



## UvA-DARE (Digital Academic Repository)

### Modeling psychopathology

*From data models to formal theories*

Haslbeck, J.M.B.

#### Publication date

2020

#### Document Version

Final published version

#### License

Other

[Link to publication](#)

#### Citation for published version (APA):

Haslbeck, J. M. B. (2020). *Modeling psychopathology: From data models to formal theories*.

#### General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

#### Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.



**MODELING  
PSYCHOPATHOLOGY**

**FROM DATA MODELS  
TO FORMAL THEORIES**

**JONAS HASLBECK**

Modeling Psychopathology:  
From Data Models to Formal Theories

Jonas Michael Benjamin Haslbeck

©2020 Jonas Haslbeck

[www.jonashaslbeck.com](http://www.jonashaslbeck.com)

Cover design: Giacomo Gallo

This thesis was typeset using  $\text{\LaTeX}$  and the thesis style design is based on the style design developed by Martijn Wieling ([www.martijnwieling.nl](http://www.martijnwieling.nl)).

**Modeling Psychopathology:  
From Data Models to Formal Theories**

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen

op donderdag 18 juni 2020, te 14:00 uur

door Jonas Michael Benjamin Haslbeck

geboren te Dachau

## Promotiecommissie

Promotor:	Dr. Lourens J. Waldorp	Universiteit van Amsterdam
Promotor:	Prof. dr. Denny Borsboom	Universiteit van Amsterdam
Overige leden:	Prof. dr. Han van der Maas	Universiteit van Amsterdam
	Prof. dr. Eric-Jan Wagenmakers	Universiteit van Amsterdam
	Prof. dr. Ellen Hamaker	Universiteit Utrecht
	Prof. dr. Casper Albers	Rijksuniversiteit Groningen
	Dr. Don Robinaugh	Harvard University
	Dr. Maarten Marsman	Universiteit van Amsterdam
	Dr. Sacha Epskamp	Universiteit van Amsterdam

Faculteit der Maatschappij- en Gedragwetenschappen

---

# Contents

---

<b>1</b>	<b>General Introduction</b>	<b>1</b>
<b>I</b>	<b>Data Models</b>	<b>5</b>
<b>2</b>	<b>Estimating Mixed Graphical Models</b>	<b>7</b>
2.1	Introduction . . . . .	8
2.1.1	Implementation and functionality . . . . .	8
2.1.2	Related implementations . . . . .	9
2.1.3	Overview of the chapter . . . . .	10
2.2	Background . . . . .	10
2.2.1	Graphical Models . . . . .	10
2.2.2	Mixed Graphical Models . . . . .	12
2.2.2.1	General Mixed Graphical Models . . . . .	12
2.2.2.2	Example: The Ising-Gaussian Model . . . . .	13
2.2.2.3	Relationship between model parameters and edges in graph . . . . .	14
2.2.3	Estimating Mixed Graphical Models . . . . .	15
2.2.4	Mixed Autoregressive Models . . . . .	18
2.2.5	Estimating time-varying models . . . . .	19
2.3	Usage and Examples . . . . .	24
2.3.1	Stationary Mixed Graphical Models . . . . .	25
2.3.1.1	Estimating Mixed Graphical Models . . . . .	25
2.3.1.2	Making Predictions from Mixed Graphical Models	27
2.3.1.3	Visualizing Mixed Graphical Models . . . . .	28
2.3.1.4	Bootstrap Sampling Distributions . . . . .	29
2.3.1.5	Sampling from Mixed Graphical Models . . . . .	31
2.3.1.6	Application: Autism and Well-being . . . . .	32
2.3.1.7	Estimating higher-order Mixed Graphical Models	34
2.3.2	Stationary mixed VAR models . . . . .	37
2.3.2.1	Estimating mixed VAR models . . . . .	37
2.3.2.2	Making Predictions from mixed VAR model . . .	39
2.3.2.3	Visualizing mixed VAR model . . . . .	39
2.3.2.4	Sampling from mixed VAR model . . . . .	40
2.3.2.5	Application: Resting state fMRI data . . . . .	42
2.3.3	Time-varying Mixed Graphical Model . . . . .	42
2.3.3.1	Estimating time-varying Mixed Graphical Model	43

2.3.3.2	Making Predictions from time-varying Mixed Graphical Model . . . . .	45
2.3.3.3	Visualizing time-varying Mixed Graphical Model . . . . .	46
2.3.3.4	Sampling from time-varying Mixed Graphical Model . . . . .	47
2.3.3.5	Bootstrap Sampling Distributions . . . . .	47
2.3.4	Time-varying mixed VAR model . . . . .	47
2.3.4.1	Estimating time-varying mixed VAR model . . . . .	47
2.3.4.2	Making Predictions from time-varying mixed VAR model . . . . .	49
2.3.4.3	Visualizing time-varying mixed VAR model . . . . .	49
2.3.4.4	Sampling from time-varying mixed VAR model . . . . .	51
2.4	Concluding Comments . . . . .	51
<b>3</b>	<b>Nodewise Predictability</b> . . . . .	<b>53</b>
3.1	Introduction . . . . .	54
3.2	Methods . . . . .	56
3.2.1	Network Models . . . . .	56
3.2.2	Making Predictions . . . . .	56
3.2.3	Quantifying Predictability . . . . .	58
3.2.3.1	Predictability in Continuous Variables . . . . .	58
3.2.3.2	Predictability in Categorical Variables . . . . .	58
3.2.4	Predictability and Model Parameters . . . . .	59
3.2.5	Application to Datasets . . . . .	59
3.3	Predictability in Cross-Sectional Networks . . . . .	60
3.4	Predictability in Temporal Networks . . . . .	62
3.5	Discussion . . . . .	64
<b>4</b>	<b>Nodewise Predictability: Reanalysis</b> . . . . .	<b>67</b>
4.1	Introduction . . . . .	68
4.2	Methods . . . . .	69
4.2.1	Literature Review & Data . . . . .	69
4.2.2	Statistical Methods . . . . .	70
4.3	Results . . . . .	71
4.3.1	Application example: node-wise predictability in data of Fried et al. (2015) . . . . .	71
4.3.2	Re-analysis of 25 datasets . . . . .	73
4.3.3	Relationship between Predictability and Edge Weights . . . . .	74
4.4	Discussion . . . . .	76
<b>5</b>	<b>Moderated Network Models</b> . . . . .	<b>79</b>
5.1	Introduction . . . . .	80
5.2	Moderated Network Models . . . . .	82
5.2.1	Moderation in Linear Regression . . . . .	82
5.2.1.1	Moderation and Interactions in Linear Regression . . . . .	82
5.2.1.2	Regression vs. Network Semantics . . . . .	84



5.2.2	Gaussian Distribution and Gaussian Graphical Model . . .	84
5.2.3	Construction of Moderated Network Model . . . . .	85
5.2.4	Visualizing Moderated Network Models . . . . .	88
5.2.5	Estimation via $\ell_1$ -regularized Nodewise Regression . . . .	89
5.2.5.1	Estimate Nodewise Regressions . . . . .	90
5.2.5.2	Combine Estimates to Joint Moderated Network Model . . . . .	91
5.3	Simulation Study . . . . .	91
5.3.1	Data generation . . . . .	92
5.3.2	Estimation . . . . .	93
5.3.2.1	Moderated Network Models . . . . .	94
5.3.2.2	Network Comparison Test (NCT) . . . . .	94
5.3.2.3	The Fused Group Lasso (FGL) . . . . .	95
5.3.3	Results . . . . .	95
5.3.4	Discussion of Simulation Results . . . . .	98
5.3.4.1	Performance of Moderated Network Models . . .	98
5.3.4.2	Performance of NCT and FGL . . . . .	99
5.3.4.3	Moderated Network Models vs. Sample-split Methods . . . . .	100
5.4	Empirical Data Examples . . . . .	100
5.4.1	R-Tutorial: Fit Moderated Network Model to Data Set of Mood Variables . . . . .	101
5.4.1.1	Fit Moderated Network Model to Data . . . . .	101
5.4.1.2	Visualize Moderated Network Model as Factor Graph . . . . .	103
5.4.2	Model Misspecification . . . . .	105
5.4.2.1	Types of Model Misspecification . . . . .	105
5.4.2.2	Different Moderation Effects across Nodewise Regressions . . . . .	106
5.5	Discussion . . . . .	108
<b>6</b>	<b>Time-varying VAR Models</b>	<b>113</b>
6.1	Introduction . . . . .	114
6.2	Estimating Time-Varying VAR Models . . . . .	116
6.2.1	Vector Autoregressive (VAR) Model . . . . .	117
6.2.2	The GAM Method . . . . .	118
6.2.3	The Kernel-smoothing Method . . . . .	120
6.2.4	Related methods . . . . .	123
6.3	Evaluating Performance via Simulation . . . . .	123
6.3.1	Simulation A: Random Graph . . . . .	124
6.3.1.1	Data generation . . . . .	124
6.3.1.2	Estimation . . . . .	126
6.3.1.3	Results . . . . .	127
6.3.1.4	Discussion . . . . .	132
6.3.2	Simulation B: Varying Sparsity . . . . .	134
6.3.2.1	Data Generation . . . . .	134

6.3.2.2	Results . . . . .	135
6.3.2.3	Discussion . . . . .	135
6.3.3	Overall Discussion of Simulation Results . . . . .	136
6.4	Estimating time-varying VAR model on Mood Time Series . . . . .	138
6.4.1	Data . . . . .	138
6.4.2	Load R-packages and Dataset . . . . .	139
6.4.3	Estimating Time-Varying VAR Model . . . . .	140
6.4.4	Assessing Reliability of Parameter Estimates . . . . .	142
6.4.5	Computing Time-Varying Prediction Error . . . . .	143
6.4.6	Visualizing Time-Varying VAR model . . . . .	144
6.4.7	Selecting between Stationary and Time-varying Models . . . . .	145
6.5	Discussion . . . . .	146
<b>7</b>	<b>Selecting between AR and VAR Models</b>	<b>149</b>
7.1	Introduction . . . . .	150
7.2	When does VAR outperform AR? . . . . .	152
7.2.1	Simulation Setup . . . . .	152
7.2.2	Simulation Results . . . . .	153
7.3	Choosing between VAR and AR based on Prediction Error . . . . .	156
7.3.1	The Relation between Prediction Error and Estimation Error . . . . .	157
7.3.2	Assessing $n_{gap}$ through simulation . . . . .	158
7.3.3	Performance of the “1 Standard Error Rule” . . . . .	160
7.4	Discussion . . . . .	162
<b>8</b>	<b>The Input Matters: Interpreting the Ising Model</b>	<b>167</b>
8.1	Introduction . . . . .	168
8.2	Different Domain, Different Interpretation . . . . .	169
8.3	Different Domain, Different Dynamics . . . . .	171
8.4	Transforming from $\{-1, 1\}$ to $\{0, 1\}$ and vice versa . . . . .	174
8.5	Conclusions . . . . .	176
<b>II</b>	<b>Formal Theories</b>	<b>177</b>
<b>9</b>	<b>Recovering Bistable Systems from Time Series Data</b>	<b>179</b>
9.1	Introduction . . . . .	180
9.2	Bistable Emotion System as Data Generating Model . . . . .	182
9.2.1	Model Specification . . . . .	182
9.2.2	Generating Time Series from Bistable System . . . . .	185
9.2.2.1	Ideal Time Series . . . . .	186
9.2.2.2	Experience Sampling (ESM) Time Series . . . . .	187
9.2.3	Qualitative Characteristics of the Model . . . . .	189
9.3	Recovering the Bistable System from Ideal Data . . . . .	190
9.3.1	Descriptive Statistics . . . . .	191
9.3.2	Hidden Markov Model . . . . .	192
9.3.3	Lag-0 Relationships / Gaussian distribution . . . . .	194
9.3.4	Lag-1 Relationships / VAR Model . . . . .	196

9.3.5	Threshold VAR Model . . . . .	199
9.3.6	Differential Equation Model Building . . . . .	203
9.3.6.1	Model Building Procedure . . . . .	203
9.3.6.2	Dynamics and Data Generated by Final Model . . . . .	205
9.3.6.3	Exact Recovery of Model Parameters . . . . .	207
9.3.7	Summary: Analysis of Ideal Time Series . . . . .	207
9.4	Recovering the Bistable Systems from ESM Data . . . . .	208
9.4.1	Descriptive Statistics, HMM and Lag-0 Relationships . . . . .	209
9.4.2	Lag-1 Relationships and VAR model . . . . .	210
9.4.3	Threshold VAR Model . . . . .	212
9.4.4	Differential Equation Model Building . . . . .	215
9.4.4.1	Model Building Procedure . . . . .	215
9.4.4.2	Dynamics and Data Generated by Final Model . . . . .	216
9.4.5	Summary: Analysis of ESM Time Series . . . . .	219
9.5	Discussion . . . . .	220
9.5.1	Implications for Studying Mental Disorders as Complex Systems . . . . .	220
9.5.2	Moving Forward: Proposing Plausible Dynamical Systems Models . . . . .	222
9.5.3	Limitations . . . . .	223
9.5.4	Summary . . . . .	225
<b>10</b>	<b>A Formal Theory of Panic Disorder . . . . .</b>	<b>227</b>
10.1	Introduction . . . . .	228
10.2	A Survey of Panic Disorder Theory and Phenomenology . . . . .	229
10.2.1	Building Blocks of a Panic Attack . . . . .	230
10.2.1.1	Physiological Arousal . . . . .	230
10.2.1.2	Perceived Threat . . . . .	230
10.2.1.3	Escape Behavior . . . . .	231
10.2.2	The Building Blocks of Panic Disorder . . . . .	231
10.2.2.1	Persistent Concern . . . . .	231
10.2.2.2	Avoidance Behavior . . . . .	232
10.2.3	Functional Relations among Building Blocks . . . . .	232
10.2.3.1	The Vicious Cycle of Panic Attacks . . . . .	232
10.2.3.2	The Vicious Cycle of Panic Disorder . . . . .	233
10.2.4	Summary . . . . .	234
10.3	A Model of Panic Disorder as a Non-Linear Dynamical System . . . . .	234
10.3.1	The Vicious Cycle of Panic Attacks . . . . .	236
10.3.1.1	Arousal (A) & the Effect of Perceived Threat on Arousal (T→A) . . . . .	236
10.3.1.2	Perceived Threat (T) & the Effect of Arousal on Perceived Threat (A→T) . . . . .	237
10.3.2	Moderating the Strength of the Vicious Cycle of Panic Attacks . . . . .	238
10.3.2.1	Arousal Schema (S) . . . . .	238
10.3.2.2	Context (C) . . . . .	238

10.3.2.3	The Moderating Effects of Arousal Schema and Context . . . . .	239
10.3.2.4	Illustrating the Moderating Effect of Arousal Schema . . . . .	239
10.3.3	Regulating Vicious Cycle of Panic Attacks . . . . .	241
10.3.3.1	The Regulating Effect of Escape Behavior on Perceived Threat ( $T \rightleftharpoons E$ ) . . . . .	241
10.3.3.2	The Regulating Effect of Homeostatic Feedback on Arousal ( $A \rightleftharpoons H$ ) . . . . .	242
10.3.3.3	Illustrating the Regulating Effects of Escape Behavior ( $E \rightarrow T$ ) and Homeostatic Feedback ( $H \rightarrow A$ ) . . . . .	242
10.3.4	Natural Variation in Arousal . . . . .	243
10.3.5	The Vicious (or Virtuous) Cycle of Panic Disorder . . . . .	244
10.3.5.1	The Effect of Panic Disorder Variables on Panic Attack Variables . . . . .	244
10.3.5.2	The Effect of Panic Attacks on Panic Disorder Variables . . . . .	245
10.4	Evaluating the Theory of Panic Disorder . . . . .	246
10.4.1	Feature 1: Individual Differences in Vulnerability to Panic Attacks . . . . .	247
10.4.1.1	Resilience . . . . .	249
10.4.2	Feature 2: The Phenomenology of Panic Attacks . . . . .	249
10.4.3	Feature 3: Coherence of the Panic Disorder Syndrome . . . . .	250
10.4.3.1	Avoidance Promotes Engineering Resilience . . . . .	251
10.4.4	Feature 4: Non-clinical Panic Attacks . . . . .	252
10.4.5	Feature 5: The Efficacy of Cognitive Behavioral Therapy for Panic Disorder . . . . .	254
10.4.5.1	Cognitive Behavioral Therapy Promotes Ecological Resilience . . . . .	256
10.5	A Theory Driven Research Agenda for Panic Disorder . . . . .	256
10.5.1	Developing Panic Disorder Theory . . . . .	257
10.5.1.1	Cumulative Theory Development with Computational Modeling . . . . .	257
10.5.1.2	Theory Evaluation . . . . .	258
10.5.2	Grounding the Model in Data . . . . .	260
10.5.3	Testing Theory Predictions . . . . .	260
10.6	Understanding and Investigating Mental Disorders as Complex Systems . . . . .	261
10.6.1	Harmful Stable States and the Definition of Disorder . . . . .	261
10.6.2	Mental Disorders as Emergent Phenomena . . . . .	262
10.6.3	Explanatory Pluralism and Equifinality . . . . .	263
10.6.4	Dysfunction from Function . . . . .	263
10.6.5	Developing Theories for Other Mental Disorders . . . . .	263
10.7	Conclusions . . . . .	264

<b>11 From Data Models to Formal Theories</b>	<b>267</b>
11.1 Introduction	268
11.2 The Nature and Importance of Formal Theories	268
11.2.1 Theories and Target Systems	269
11.2.2 The Importance of Formal Theories	270
11.2.2.1 A Formal Theory of Panic Disorder	272
11.2.3 Data and Data Models	273
11.3 Identifying Formal Theories from Data	274
11.3.1 Using Data Models as Formal Theories	275
11.3.1.1 Properties of Mental Disorder Target Systems	276
11.3.1.2 Comparing Target System Properties with Data Model Properties	276
11.3.1.3 Data Models as Formal Theories?	278
11.3.2 Using Data Models to Infer Formal Theories	279
11.3.2.1 Inferring the Panic System from Network Data Models	280
11.3.2.2 The Mapping between Data Model and Target System	282
11.3.3 Using Data Models to Develop Formal Theories	284
11.3.3.1 Obtaining Theory-Implied and Empirical Data Models	284
11.3.3.2 Theory Development: Comparing Model-Implied and Empirical Data Models	286
11.4 An Abductive Approach to Constructing Formal Theory	288
11.4.1 Generating Theory	289
11.4.1.1 Establishing the Phenomenon	289
11.4.1.2 Generate Initial Verbal Theory	290
11.4.1.3 Formalize Initial Theory	291
11.4.2 Developing Theory	293
11.4.3 Testing Theory	295
11.5 Conclusions	297
<b>III Conclusion</b>	<b>299</b>
<b>12 Discussion</b>	<b>301</b>
12.1 Data Models	301
12.2 From Data Models to Formal Theories	303
12.3 Formal Theories	304
12.4 Future Directions	305
12.4.1 Refining Theory Development	306
12.4.2 Formalizing Measurement Models	307
12.4.3 Dysfunction from Function	308
12.5 Conclusions	311

<b>IV Appendices</b>	<b>313</b>
References . . . . .	315
<b>A Nodewise Predictability: Reanalysis</b>	<b>347</b>
A.1 Details about Literature Review . . . . .	347
<b>B Moderated Network Models</b>	<b>349</b>
B.1 Mean-centering in Moderation Analysis . . . . .	349
B.2 Joint distribution for $p = 3$ . . . . .	350
B.3 Rejection Sampling . . . . .	351
B.4 Sensitivity of Moderated Network Model across Parameter Types . . . . .	353
B.5 Simulation with Isolated Interaction Types . . . . .	353
B.6 Sensitivity as Function of Number of Uncorrelated Neighbors . . . . .	355
B.7 XY is uncorrelated with X and Y if the latter are mean-centered . . . . .	355
B.8 Additional Tutorial: Estimate MNM on Iteration 2 of Simulation Study . . . . .	356
B.8.1 Fit Moderated Network Model to Data . . . . .	356
B.8.2 Moderated Network Model as Factor Graph . . . . .	358
B.9 Varying Moderation Effects across Nodewise Regressions: A closer Look . . . . .	359
B.10 Correlations between lower- and higher-order terms . . . . .	362
B.11 Simulation Results in Tables . . . . .	364
<b>C Time-varying VAR Models</b>	<b>367</b>
C.1 Sampling Variation around Aggregated Absolute Errors . . . . .	367
C.2 Sampling Variation around Absolute Errors over Time . . . . .	369
C.3 Computational Cost . . . . .	370
C.4 Code to select Appropriate Bandwidth in KS(L1) Method . . . . .	371
C.5 Estimating time-varying VAR model via GAM(st) . . . . .	372
C.5.1 Load R-packages and dataset . . . . .	372
C.5.2 Estimating time-varying VAR model . . . . .	373
C.5.3 Visualize time-varying VAR model . . . . .	373
<b>D Selecting between AR and VAR Models</b>	<b>377</b>
D.1 Sampling cells on the $R \times D$ grid . . . . .	377
<b>E The Input Matters: Interpreting the Ising Model</b>	<b>379</b>
E.1 Statistical Equivalence worked out for two variable example . . . . .	379
E.2 Increasing interaction parameters only changes the marginal probabilities domain in $\{0, 1\}$ . . . . .	380
E.3 Derivation of Transformation from $\{0, 1\}$ to $\{-1, 1\}$ and vice versa . . . . .	382
E.4 Model equivalence across domains with penalized estimation . . . . .	385

<b>F</b>	<b>Recovering Bistable Systems from Time Series Data</b>	<b>389</b>
F.1	Determine Fixed Points of Bistable System . . . . .	389
F.2	Mean-Switching Hidden Markov Model . . . . .	391
F.2.1	Model Specification . . . . .	392
F.2.2	Model Selection for Mean-Switching HMM . . . . .	392
F.3	Data Generated from Estimated Models . . . . .	393
F.3.1	Mean Switching Hidden Markov Model . . . . .	393
F.3.2	First-order Vector Autoregressive (VAR(1)) model . . . . .	394
F.3.3	Threshold VAR(1) Model . . . . .	395
F.4	Residual Partial Correlations of TVAR(1) Model . . . . .	395
F.5	Differential Equation Model Building Details . . . . .	396
F.5.1	Evaluating Model Fit . . . . .	396
F.5.2	Ideal Data . . . . .	396
F.5.3	ESM Data . . . . .	397
F.6	Additional Model Results from ESM Time Series . . . . .	399
<b>G</b>	<b>A Formal Theory of Panic Disorder</b>	<b>403</b>
G.1	Overview of Mathematical and Computational Models of Panic Disorder . . . . .	403
G.1.1	Arousal . . . . .	405
G.1.2	Perceived Threat . . . . .	406
G.1.3	Context . . . . .	407
G.1.4	Escape Behavior . . . . .	408
G.1.5	Homoestatic Feedback . . . . .	409
G.1.6	Noise . . . . .	409
G.1.7	Avoidance . . . . .	410
G.1.8	Arousal Schema . . . . .	410
G.2	Further Examining the Vicious Cycle of Panic Attacks . . . . .	412
G.3	Further Examining Cognitive Behavioral Therapy Intervention on the Model . . . . .	413
G.4	Theory Evaluation . . . . .	415
G.4.1	What can the model explain? . . . . .	415
G.4.2	Limitations to the Model's Accuracy and Consilience . . . . .	416
G.4.3	Conclusion . . . . .	418
G.5	Theory Development . . . . .	418
G.5.1	What can this modification to the model explain? . . . . .	421
G.5.2	Future Research and Theory Development . . . . .	421
<b>H</b>	<b>From Data Models to Formal Theories</b>	<b>423</b>
H.1	Simulated Data from the Panic Model . . . . .	423
H.2	Panic Model and Statistical Dependencies . . . . .	423
H.3	Details Empirical vs Simulated Ising Model . . . . .	425
H.3.1	Simulated Data and Implied Ising Model . . . . .	425
H.3.2	Empirical Symptom Data . . . . .	427

*CONTENTS*

---

<b>I</b>	<b>List of Publications</b>	<b>429</b>
I.1	Under Review . . . . .	429
I.2	Published . . . . .	429
<b>J</b>	<b>Nederlandse Samenvatting</b>	<b>431</b>
J.1	Data Models . . . . .	431
J.2	Formele Theorieën . . . . .	433
<b>K</b>	<b>Acknowledgements</b>	<b>435</b>



# GENERAL INTRODUCTION

---

In the late 19th century a considerable amount of the individuals admitted to asylums in Europe suffered from a single crippling condition (Kraepelin, 1913; Davis, 2008). Patients experienced symptoms such as fatigue, headaches, insomnia, mental deterioration, personality changes and loss of social inhibitions. In later stages, they completely lost control over their mind and body, and finally died of paralysis (Swain, 2018). This condition, now called “general paresis of the insane” (GPI), has been first characterized as a distinct disease in 1822 by the French physician Antoine Laurent Bayle, and further characterized as a disease of brain pathology with a predictable clinical history in the 1860s. However, throughout the 19th century, all efforts to identify a single cause for the condition remained futile. In the absence of a clear cause for the disease, the common view became that the causes of GPI are multifactorial and rooted in the negative influences of the urban environment (Davis, 2008).

This view changed radically in the first decade of the 20th century, when German zoologist Fritz Schaudinn and dermatologist Erich Hoffmann discovered the bacterium *Treponema pallidum* as the root cause for syphilis in 1905, and it became clear that GPI was in fact a manifestation of late-stage syphilis. Subsequently, German chemist Paul Ehrlich and Japanese bacteriologist Sahachiro Hata made a drug later to be known as Salvarsan, which cured syphilis and GPI by treating syphilis, its root cause. Salvarsan with its many side effects was replaced by Penicillin in the 1940s, and today syphilis (and GPI) is detected by a simple blood test and treated with a single dose of antibiotics (Swain, 2018).

The discovery of the single cause of GPI and its effective treatment with antibiotics is possibly the biggest success story in the history of psychiatry. As such, it had a major influence on the further course of the field in the 20th century: It set the expectation that mental disorders can be defined by a single root cause and treated by removing that root cause. This view was reinforced by the then predominant “germ theory” in medicine, which posits that diseases are caused by pathogens. This disease model triggered a golden age of medical bacteriology in the first half of the 20th century which led to the discoveries of causes and cures for many diseases such as smallpox, measles, and cholera (Blevins & Bronze, 2010; Hyland, 2011). Given this historical context, it was perhaps not unreasonable to expect the same kind of discoveries for psychopathology.

However, in the following hundred years no such discoveries have been made. No further pathogens have been found to be the cause of any major mental disorders, no single psychological mechanism has been discovered that fully explains any mental disorder, and also the emerging fields of (neuro-)biology and genetics

failed to identify any single causes (Kendler, 2019, 2005; Consortium et al., 2009; Shi et al., 2009; Wray et al., 2012; Hek et al., 2013; Ripke et al., 2013). Instead, a plethora of biological, psychological, and social risk factors with typically tiny effect sizes has been identified for each mental disorder (Kendler, 2012; Kapur, Phillips, & Insel, 2012). For example, a recent meta analysis found 44 genetic risk variants for major depression, each of which with an extremely small effect size (Wray et al., 2018). The current empirical evidence therefore suggests that there is no single cause to be discovered for any mental disorder.

But, then, where do mental disorders come from? In response to this question, several authors have suggested a different conceptualization of mental disorders: Instead of viewing them as a disease on some biological or psychological level, they should be seen as complex, mutually reinforcing networks of causal mechanisms (Borsboom, 2008; Kendler, Zachar, & Craver, 2011). In addition to being more consistent with the data, this new perspective also opens up new research agendas, and better reflects clinical practice: For example, Fried and Nesse (2015) showed that many risk factors are differentially related to individual symptoms, while previous studies trying to find single causes focused on finding risk factors for the symptom sum score; and the idea of dynamic, reinforcing mechanisms is the core principle of one of the most effective treatments, cognitive behavioral therapy (e.g., Beck, 1979). Interestingly, the field of medicine adopted a multifactorial disease model when it shifted its focus from infectious diseases to chronic diseases such as diabetes and hypertension in the second half of the 20th century. The field of psychiatry, however, kept on pursuing single causes for their major disorders (Kendler, 2019).

This alternative conceptualization has been formulated in more detail as the “network approach to psychopathology” (e.g., Borsboom & Cramer, 2013; Schmittmann et al., 2013; Borsboom, 2017). In this approach, symptoms are not caused by an underlying disorder, but rather the symptoms themselves and the causal interactions among them constitute the disorder. Such interactions are often plausible: For example, loss of appetite leads to loss of weight, insomnia brings about tiredness, and guilt predisposes suicidal ideation. However, also longer causal chains and loops seem likely, such as sleeping problems → fatigue → concentration problems → worrying → sleeping problems. The explanation for the occurrence of mental disorders in this approach is that the system of symptoms is “pushed” into a state of elevated activation, for example by an adverse life event; and then the activation is maintained by the reinforcing causal effects, even if the negative outside influence is removed (Borsboom, 2017).

This new framework struck a chord with applied researchers and led to a surge of empirical studies analyzing empirical data with statistical network models, which capture statistical dependencies between pairs of symptoms, and thereby embrace the “network approach” (for reviews see e.g., Robinaugh, Hoekstra, & Borsboom, 2019; Contreras, Nieto, Valiente, Espinosa, & Vazquez, 2019). These studies revealed fairly stable patterns of statistical relations between symptoms, and these patterns are often difficult to explain with an unknown single disease acting as the common cause of all symptoms (e.g., Fried, van Borkulo, et al., 2016), suggesting that a different explanatory framework is needed. This

---

emerging literature is heavily based on a methodological toolbox which made estimation routines and visualizations for statistical network models (i.e., data models) available to applied researchers (e.g., Epskamp et al., 2012; Van Borkulo et al., 2014; Epskamp, Borsboom, & Fried, 2016; Epskamp, Waldorp, Möttus, & Borsboom, 2018; van Borkulo et al., 2016).

The first part of this dissertation extends this toolbox of data models. While applied researchers were initially only able to estimate the Ising model (only binary variables) or the multivariate Gaussian distribution (only continuous variables), I created the R-package *mgm* which allows one to estimate network models in which variables can be continuous, count, or categorical with any number of categories (Chapter 2). Many early network studies only reported a visualization of the model parameters, which made it hard to judge how well the model fits the data and therefore how relevant the estimated relationships are. To address this issue I suggested to compute and visualize the predictability of each variable (Chapter 3) and I compared the predictability of symptom networks of different mental disorders in a reanalysis of the early network literature (Chapter 4). All network models discussed so far were pairwise models, which means that all interactions are independent of the values of all modeled variables. This means that one cannot model moderation effects. However, moderation effects are plausible for many psychological phenomena and often the very topic of study. I therefore introduced Moderated Network Models (MNMs) which allow relationships between pairs of variables to be moderated by all remaining variables (Chapter 5).

In addition to cross-sectional data, also time-series data have been analyzed from a network perspective. The most popular model in this context is the Vector Autoregressive (VAR) model, in which each variable is a linear function of all variables (including itself) at previous time points. In Chapter 7, I discuss the topics of bias-variance trade-off and model selection in the context of selecting between the VAR and the simpler AR model, and map out how well the VAR model can be estimated in typical psychological applications. A central assumption of the VAR model is that its parameters remain constant over time. However, in the network approach the key interest lies on the interactions between symptoms and how they change when transitioning from being healthy to having a mental disorder. To capture such change over time within an individual, I implemented a method to estimate time-varying VAR models (Chapter 6). Finally, in Chapter 8 I discuss the subtleties involved in interpreting and using the Ising model as a dynamic within-person model.

The wealth of empirical network studies provided a more detailed picture of the empirical phenomenon of mental disorders by uncovering the patterns of statistical relations between symptoms and related variables. Together with research on treatments and the myriad of results on biological, psychological and social risk factors, this picture makes it extremely unlikely that any major disorder can be explained by a single cause. This suggests that adopting a complex systems view on mental disorders is more promising and that, in order to explain mental disorders, we need to piece together a complex web of causal effects across biological, psychological and societal levels. This is a daunting task which

requires a strong framework of formalized theory. However, so far only few formal theories of mental disorders exist (e.g., Dujmić, Machielse, & Treur, 2018; von Kentzinsky, Wijtsma, & Treur, 2019; Burger et al., 2019); and it is generally unclear how to construct such theories from empirical data in the context of psychopathology. Developing formal theories based on empirical data is therefore a largely uncharted territory for research on mental disorders.

The second part of this dissertation explores this territory. In Chapter 9, I use a bistable dynamical system for emotion dynamics to illustrate the general difficulty of inferring a formal theory from data models, and the problem of observing a system operating on a fast time scale with low frequency measurements. This chapter shows that it is very unlikely that formal theories of mental disorders can be directly inferred from data, and that researchers have to engage in a more general theory development procedure that is typical in fields with a strong tradition in formal modeling. In Chapter 10 I present a formal theory for Panic disorder, in whose development I participated. This theory illustrates the many benefits of formalization such as making explicit what is unknown, facilitating a collaborative and cumulative science, and making specific, testable predictions. The development of this formal theory further highlights the fact that good formal theories of mental disorders are unlikely to take the form of data models that can be estimated directly from data. Finally, in Chapter 11 I attempt to connect the worlds of data models and formal theories by discussing how to use the former to obtain the latter. As a result of this discussion I put forward an abductive approach to theory construction that puts formal theories at the heart of theory development. In Chapter 12 I conclude by reflecting on how my thinking about modeling psychopathology changed in the course of my PhD and suggesting three directions for future research.

**Part I**

**Data Models**



# ESTIMATING MIXED GRAPHICAL MODELS

---

## Abstract

We present the R-package `mgm` for the estimation of  $k$ -order Mixed Graphical Models (MGMs) and mixed Vector Autoregressive (mVAR) models in high-dimensional data. These are a useful extensions of graphical models for only one variable type, since data sets consisting of mixed types of variables (continuous, count, categorical) are ubiquitous. In addition, we allow to relax the stationarity assumption of both models by introducing time-varying versions MGMs and mVAR models based on a kernel weighting approach. Time-varying models offer a rich description of temporally evolving systems and allow to identify external influences on the model structure such as the impact of interventions. We provide the background of all implemented methods and provide fully reproducible examples that illustrate how to use the package.

## 2.1 Introduction

We present *mgm*, an R-package for the estimation of (time-varying)  $k$ -order Mixed Graphical Models (MGMs) and (time-varying) mixed Vector Autoregressive (mVAR) models with a specified set of lags. The package is available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.r-project.org/>. In this chapter we introduce these models, discuss algorithms to estimate them, and present a number of fully reproducible code examples that show how to use the implementations provided by *mgm*.

Graphical models have become a popular way to abstract complex systems and gain insights into relational patterns among observed variables in a large variety of disciplines such as statistical mechanics (Albert & Barabasi, 2002), biology (N. Friedman, Linial, Nachman, & Pe'er, 2000), genetics (Ghazalpour et al., 2006), neuroscience (Huang et al., 2010) and psychology (Borsboom & Cramer, 2013). In many of these applications the dataset of interest consists of *mixed variables* such as binary, categorical, ordinal, counts, continuous and/or skewed continuous amongst others. One example is internet-scale marketing data, where it is of interest to relate variables such as clicked links (categorical), time spent on websites (possibly exponential), browsing history (categorical), social media postings (count), friends in social networks (count), and many others. In a medical context, one could be interested in interactions between person characteristics such as gender (categorical) or age (continuous), frequencies of behaviors (count), taking place of events (categorical) and the dose of a drug (continuous).

If measurements are taken repeatedly from a system, one can be either interested in relations between variables at the *same time point* or in relations between variables *across time points*. The former relations are modeled by MGMs, the latter relations are modeled by Vector Autoregressive (VAR) models, which relate variables over a specified set of time lags. For both types of models it may be appropriate in some situations to relax the assumption of stationarity, such that its parameters are allowed to vary over the measured time period. These time-varying models provide additional information for understanding and predicting organizational processes, the diffusion of information, detecting vulnerabilities and the potential impact of interventions. An example is the developmental cycle of a biological organism, in which different genes interact at different stages of development. In a medical context, the aim could be to study the impact of an intervention on the dependencies between a large number of physiological and psychological variables that capture the health of a patient. Yet another example can be found in the field of psychiatry, where one might be interested in the interaction of negative life events, social contacts and symptoms of psychological disorders such as major depression.

### 2.1.1 Implementation and functionality

The *mgm* package is written in R and uses the *glmnet* package (J. Friedman, Hastie, & Tibshirani, 2010) to fit penalized Generalized Linear Models (GLMs) to perform neighborhood selection (Meinshausen & Bühlmann, 2006). The *glm-*



*net* package is written in Fortran and is optimized for computational efficiency. In addition, *mgm* depends on the packages *matrixcalc*, *stringr*, *Hmisc*, *gtools* and *qgraph*.

The main functionality of the *mgm* package is to estimate Mixed Graphical Models (MGMs) and mixed Autoregressive (mVAR) Models, both as stationary models (`mgm()` and `mvar()`) and time-varying models (`tvmgm()` and `tvmlvar()`). In addition, we provide the S3 methods `print()` to summarize model objects and `predict()` to compute predictions and nodewise errors from all types of models, and the function `resample()` to determine the stability of estimates via resampling. Furthermore, *mgm* provides functions to sample from all four models in full flexibility in order to enable the user to investigate the performance of the estimation algorithms in a particular situation. The output of all estimation functions is designed to allow a seamless visualization with the *qgraph* package (Epskamp et al., 2012) and we therefore do not provide our own plotting functions.

## 2.1.2 Related implementations

Several packages are available to estimate Gaussian Graphical Models (GGMs): the R-packages *glasso* (J. Friedman & Tibshirani, 2014) and *huge* (Zhao et al., 2015; Zhao, Liu, Roeder, Lafferty, & Wasserman, 2012) implement the graphical lasso (Banerjee, El Ghaoui, & d’Aspremont, 2008; J. Friedman, Hastie, & Tibshirani, 2008b) which maximizes a  $\ell_1$ -penalized Gaussian log-likelihood. The *huge* package also allows to estimate GGMs via neighborhood selection (Meinshausen & Bühlmann, 2006), in which the neighborhood of each node is estimated separately and then the local estimates are combined to obtain the (global) graphical model. The R-package *IsingFit* (van Borkulo, Epskamp, & Robitzsch, 2014; van Borkulo, Borsboom, et al., 2014) implements a neighborhood selection based method to estimate the binary-valued Ising model (see e.g. Wainwright & Jordan, 2008; Ravikumar, Wainwright, & Lafferty, 2010). The *XMRF* package (Wan et al., 2015) allows to estimate Markov Random fields of the Multivariate Gaussian distribution, Ising models, log-linear Poisson based graphical model, regular Poisson graphical models, truncated Poisson graphical models and sublinear Poisson graphical models (Yang, Ravikumar, Allen, & Liu, 2015, 2013).

For VAR models, the *vars* package Pfaff (2008b) allows to fit VAR models with Gaussian noise. The *BigVAR* package (Nicholson, Matteson, & Bien, 2017) allows to fit VAR models and structured VAR models with Gaussian noise with structured  $\ell_1$ -penalties. The *mlVAR* package (Epskamp, Deserno, & Bringmann, 2017) implements multilevel VAR models with Gaussian noise. Graphical VAR models (Wild et al., 2010), in which lagged coefficients and contemporaneous effects are estimated simultaneously, can be estimated with the *graphicalVAR* package (Epskamp, 2017).

For time-varying graphical models, there is a Python implementation of the SINGLE algorithm of R. P. Monti et al. (2014) for time-varying Gaussian graphical models (R. Monti, 2014) and *GraphTime* (Immer & Gibberd, 2017), a Python implementation of time-varying (dynamic) graphical models based on the (group)

fused-lasso as presented by Gibberd and Nelson (2017). The R package *tvReg* allows to estimate linear VAR models using kernel smoothing (Casas & Fernandez-Casal, 2018).

*mgm* goes beyond the above mentioned packages in that it allows one to estimate  $k$ -order MGMs and mVAR models (with any set of lags), compute predictions from them and assess model stability via resampling, while the above packages only allow one to do this for special cases. In addition, the output of *mgm* is designed to allow a seamless visualization of estimated models using the R-package *qgraph* (Epskamp et al., 2012). Finally, *mgm* is the first package that allows to estimate time-varying MGMs and mVAR models.

### 2.1.3 Overview of the chapter

In Section 2.2, we introduce Mixed Graphical Models (MGMs) (Section 2.2.2) and mixed Vector Autoregressive (mVAR) models (Section 2.2.4), and discuss how to estimate these models in their stationary (Section 2.2.3) and time-varying (Section 2.2.5) versions. In Section 2.3, we illustrate how to use the *mgm* package to estimate parameters, compute predictions from and visualize stationary MGMs (Section 2.3.1), stationary mVAR models (Section 2.3.2), time-varying MGMs (Section 2.3.3) and time-varying mVAR models (Section 2.3.4). All presented examples are fully reproducible, with code either shown in the chapter or provided in the online supplementary material.

## 2.2 Background

In this section we provide basic concepts related to graphical models (Section 2.2.1), introduce the model classes Mixed Graphical Models (MGMs) (Section 2.2.2) and mixed Vector Autoregressive (mVAR) models (Section 2.2.4), and show how to estimate these models in their stationary (Section 2.2.3) and time-varying (Section 2.2.5) versions.

### 2.2.1 Graphical Models

Undirected graphical models are families of probability distributions that respect a set of conditional independence statements represented in an undirected graph  $G$  (Lauritzen, 1996). This connection between probability distribution and graph  $G$  is formalized by the *Global Markov Property*, which we define after introducing some notation.

An undirected graph  $G = (V, E)$  consists of a collection of nodes  $V = \{1, 2, \dots, p\}$  and a collection of edges  $E \subset V \times V$ . A subset of nodes  $U$  is a *node cutset* whenever its removal breaks the graph in two or more nonempty subsets, which is equivalent to  $U$  being the set such that all paths from disjoint node sets  $S$  and  $Q$  go through  $U$  (Lauritzen, 1996). A *clique* is a subset  $C \subseteq V$  such that  $(s, t) \in E$  for all  $s, t \in C$  where  $s \neq t$ , and is called a *maximal clique* if inclusion of any other node would make it not a clique. The neighborhood  $N(s)$  of a node

$s \in V$  is the set of nodes that are connected to  $s$  by an edge,  $N(s) := \{t \in V \mid (s, t) \in E\}$ . Throughout the chapter we use the shorthand  $X_{\setminus s}$  for  $X_{V \setminus \{s\}}$ .

To each node  $s$  in graph  $G$  we associate a random variable  $X_s$  taking values in a space  $\mathcal{X}_s$ . For any subset  $A \subseteq V$ , we use the shorthand  $X_A := \{X_s, s \in A\}$ . For three disjoint subsets of nodes,  $A$ ,  $B$ , and  $U$ , we write  $X_A \perp\!\!\!\perp X_B \mid X_U$  to indicate that the random vector  $X_A$  is independent of  $X_B$  when conditioning on  $X_U$ . We can now define graphical models in terms of the Markov property (see e.g. Loh & Wainwright, 2012):

**Definition 1** (*Global Markov property*). *If  $X_A \perp\!\!\!\perp X_B \mid X_U$  whenever  $U$  is a node cutset that breaks the graph into disjoint subsets  $A$  and  $B$ , then the random vector  $X := (X_1, \dots, X_p)$  is Markov with respect to the graph  $G$ .*

Note that the neighborhood set  $N(s)$  is always a node cutset for  $A = \{s\}$  and  $B = V \setminus \{s \cup N(s)\}$ .

In the remainder of this chapter we focus on exponential family distributions, which are strictly positive distributions. For these distributions the Global Markov property is equivalent to the *Markov factorization property* by the Hammersley-Clifford Theorem (Lauritzen, 1996). Consider for each clique  $C$  in the set of all clique sets  $\mathcal{C}$  a clique-compatibility function  $\psi_C(X_C)$  that maps configurations  $x_C = \{x_s, s \in C\}$  to  $\mathbb{R}^+$  such that  $\psi_C$  only depends on the variables  $X_C$  corresponding to the clique  $C$ .

**Definition 2** (*Markov factorization property*). *The distribution of  $X$  factorizes according to  $G$  if it can be represented as a product of clique functions*

$$P(X) \propto \prod_{C \in \mathcal{C}} \psi_C(X_C). \quad (2.1)$$

This equivalence implies that if we have distributions that are represented as a product of clique functions, then we can represent the conditional dependence statements in this distribution in a graph  $G$ . This is the case for the exponential family distributions we use in the present chapter

$$P(X) = \exp \left\{ \sum_{C \in \mathcal{C}} \theta_C \phi_C(X_C) - \Phi(\theta) \right\}, \quad (2.2)$$

where the functions  $\phi_C(X_C) = \log \psi_C(X_C)$  are sufficient statistic functions specified by the exponential family member at hand (e.g. Gaussian, Exponential, Poisson, etc.),  $\theta_C$  are parameters associated with the clique functions and  $\Phi(\theta)$  is the log-normalizing constant

$$\Phi(\theta) = \log \int_{\mathcal{X}} \sum_{C \in \mathcal{C}} \theta_C \phi_C(X_C) \nu(dx),$$

where depending on the distribution of  $X$ , the measure  $\nu$  is a counting measure or Lebesgue measure (for details see Wainwright, Jordan, et al., 2008).

The graph  $G$  represents a *family* of distributions because its edges do not indicate the strength of the dependency and the nodes can represent different conditional distributions. Hence there is a one to one mapping from the density to the graph, but a one to many mapping from graph to densities.

## 2.2.2 Mixed Graphical Models

In this section we first introduce the general class of Mixed Graphical Models (2.2.2.1), and then provide the Ising-Gaussian model as a specific example (2.2.2.2).

### 2.2.2.1 General Mixed Graphical Models

In this section, we introduce the class of Mixed Graphical Models (MGMs), which are a special case of the distribution in Equation 2.2 that allow one to combine an arbitrary set of conditional univariate members of the exponential family in a joint distribution (Yang, Baker, Ravikumar, Allen, & Liu, 2014a; S. Chen, Witten, et al., 2015).

Consider a  $p$ -dimensional random vector  $X = (X_1, \dots, X_p)$  with each variable  $X_s$  taking values in a potentially different set  $\mathcal{X}_s$ , and let  $G = (V, E)$  be an undirected graph over  $p$  nodes corresponding to the  $p$  variables. Now suppose the node-conditional distribution of node  $X_s$  conditioned on all other variables  $X_{\setminus s}$  is given by an arbitrary univariate exponential family distribution

$$P(X_s|X_{\setminus s}) = \exp\{E_s(X_{\setminus s})\phi_s(X_s) + B_s(X_s) - \Phi(X_{\setminus s})\}, \quad (2.3)$$

where the functions of the sufficient statistic  $\phi_s(\cdot)$  and the base measure  $B_s(\cdot)$  are specified by the choice of exponential family and the canonical parameter  $E_s(X_{\setminus s})$  is a function of all variables except  $X_s$ . (Wainwright et al., 2008) make these functions explicit for a number of exponential family distributions.

These node-conditional distributions are consistent with a joint distribution over the random vector  $X$  as in (2.1), that is Markov with respect to graph  $G = (V, E)$  with the set of cliques  $\mathcal{C}_k$  of size at most  $k$ , if and only if the canonical parameters  $\{E_s(\cdot)\}_{s \in V}$  are a linear combination of products of univariate sufficient statistic functions  $\{\phi_r(X_r)\}_{r \in N(s)}$  of order up to  $k$

$$\theta_s + \sum_{r \in N(s)} \theta_{s,r} \phi_r(X_r) + \dots + \sum_{r_1, \dots, r_{k-1} \in N(s)} \theta_{r_1, \dots, r_{k-1}} \prod_{j=1}^{k-1} \phi_{r_j}(X_{r_j}), \quad (2.4)$$

where  $\theta_s := \{\theta_s, \theta_{s,r}, \dots, \theta_{sr_2 \dots r_k}\}$  is a set of parameters and  $N(s)$  is the set of neighbors of node  $s$  according to graph  $G$  (Yang et al., 2014a). Factorizing  $p$  conditional distributions as in Equation 2.3 gives the joint distribution

$$\begin{aligned}
 P(X) = \exp \left\{ \sum_{s \in V} \theta_s \phi_s(X_s) + \sum_{s \in V} \sum_{r \in N(s)} \theta_{s,r} \phi_s(X_s) \phi_r(X_r) + \right. \\
 \left. \dots + \sum_{r_1, \dots, r_k \in \mathcal{C}} \theta_{r_1, \dots, r_k} \prod_{j=1}^k \phi_{r_j}(X_{r_j}) + \sum_{s \in V} B_s(X_s) - \Phi(\theta) \right\}, \quad (2.5)
 \end{aligned}$$

where  $\Phi(\theta)$  is the log-normalization constant.

The dimensionality of the parameter vector  $\theta$  depends both on the type of modeled variables and the order of interactions. If one only models continuous variables with pairwise interactions ( $k = 2$ ), the MGM simplifies to the multivariate Gaussian distribution which is parameterized by a  $1 \times p$  vector of intercepts and a  $p \times p$  matrix of  $\binom{p}{2}$  partial correlations. Including all 3-way interactions would lead to an additional  $\binom{p}{3}$  parameters, etc. At the end of Section 2.2.2, we discuss the dimensionality of the parameter vector in the presence of categorical variables.

Necessary conditions for the mixed density in Equation 2.5 to be normalizable are discussed in Yang et al. (2014a). S. Chen et al. (2015) show constraints on the parameter space to ensure normalizability for a number of MGMs with at most pairwise interactions. *mgm* does not allow one to implement the constraints, since the underlying *glmnet* package does not support the specification of these constraints. However, Trip and van Wieringen (2018) recently proposed an algorithm that allows to estimate pairwise MGMs with these constraints.

### 2.2.2.2 Example: The Ising-Gaussian Model

We take the Ising-Gaussian model as a specific example of the joint distribution in Equation 2.5. Consider a random vector  $X := (Y, Z)$ , where  $Y = \{Y_1, \dots, Y_{p_1}\}$  are univariate Gaussian random variables,  $Z = \{Z_1, \dots, Z_{p_2}\}$  are univariate Bernoulli random variables and we only consider pairwise interactions between sufficient statistics. For the univariate Gaussian distribution (with known variance  $\sigma^2$ ) the sufficient statistic function is  $\phi_Y(Y_s) = \frac{Y_s}{\sigma_s}$  and the base measure is  $B_Y(Y_s) = -\frac{Y_s^2}{2\sigma_s^2}$ . The Bernoulli distribution has the sufficient statistic function  $\phi_Z(Z_s) = Z_s$  and the base measure  $B_Z(Z_s) = 0$ . From the MGM joint distribution in Equation 2.5 follows that this mixed density is given by

$$\begin{aligned}
 P(Y, Z) \propto \exp \left\{ \sum_{s \in V_Y} \frac{\theta_s}{\sigma_s} Y_s + \sum_{r \in V_Z} \theta_r Z_r + \sum_{(s,r) \in E_Y} \frac{\theta_{s,r}}{\sigma_s \sigma_r} Y_s Y_r + \right. \\
 \left. \sum_{(s,r) \in E_Z} \theta_{s,r} Z_s Z_r + \sum_{(s,r) \in E_{YZ}} \frac{\theta_{s,r}}{\sigma_s} Y_s Z_r - \sum_{s \in V_Y} \frac{Y_s^2}{2\sigma_s^2} \right\}, \quad (2.6)
 \end{aligned}$$

where the first two terms are thresholds for Gaussian and Bernoulli variables, the third term represents pairwise interactions between Gaussians, the fourth term represents pairwise interactions between Bernoulli variables, the fifth term represents pairwise interactions between Gaussians and Bernoulli variables, and the last term sums over the base measures for the Gaussians.

When the conditional distribution is a Bernoulli random variable  $Z_r$ , it is given by

$$P(Z_r|Z_{\setminus r}, Y) \propto \exp \left\{ \theta_r Z_r + \sum_{s \in N(r)_Z} \theta_{s,r} Z_s Z_r + \sum_{s \in N(r)_Y} \frac{\theta_{s,r}}{\sigma_r} Z_r Y_s \right\}. \quad (2.7)$$

Note that the conditional distribution in Equation 2.7 has the same form as the distribution of a single variable conditioned on all remaining variables in an Ising model plus one additional term for interactions between Bernoulli and Gaussian random variables.

When the conditional distribution is a Gaussian random variable  $Y_s$ , it is given by

$$P(Y_s|Y_{\setminus s}, Z) \propto \exp \left\{ \frac{\theta_s}{\sigma_s} Y_s + \sum_{r \in N(s)_Y} \frac{\theta_{s,r}}{\sigma_s \sigma_r} Y_s Y_r + \sum_{r \in N(s)_Z} \frac{\theta_{s,r}}{\sigma_s} Y_s Z_r - \frac{Y_s^2}{2\sigma_s^2} \right\}.$$

Now, let  $\sigma = 1$ , factor out  $Y_s$  and let  $\mu_s = \theta_s + \sum_{r \in N(s)_Y} \theta_{s,r} Y_r + \sum_{r \in N(s)_Z} \theta_{s,r} Z_r$ .

Finally, when taking  $\frac{\mu_s^2}{2}$  out of the log normalization constant, we arrive with basic algebra at the well-known form of the univariate Gaussian distribution with unit variance

$$P(Y_s|Y_{\setminus s}, Z) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(Y_s - \mu_s)^2}{2} \right\}.$$

### 2.2.2.3 Relationship between model parameters and edges in graph

For pairwise MGMs (size of cliques is at most  $k = 2$ ), a pairwise interaction between two continuous variables  $s$  and  $r$  is parameterized by a single parameter  $\theta_{s,r}$ . Now, whether the edge between  $s$  and  $r$  is present depends on whether  $\theta_{s,r}$  is zero or not, that is,  $(s, r) \in E \iff \theta_{s,r} \neq 0$ . Thus, if only pairwise interactions between continuous variables are modeled, any given edge is a function of a *single* parameter. This implies that a *weighted* graph fully represents the parameterization of interactions in the underlying model (or the full parameterization minus the threshold parameters). Interactions between categorical variables with  $m > 2$ , however, are specified by more than one parameter. For instance, a pairwise interaction between two categorical variables with  $m$  and  $u$  categories is parameterized by  $R = (m-1) \times (u-1)$  parameters associated with corresponding indicator functions for all  $R$  states (e.g., Agresti, 2003). A pairwise interaction between a

categorical variable with  $m$  categories and a continuous variable has  $R = 1 \times (m-1)$  parameters associated with  $m-1$  indicator functions multiplied with the continuous variable. In this case,  $\theta_{s,r}^z$  is a parameter defining the interaction between the nodes  $s$  and  $r$  indexed by  $z \in \{1, \dots, R\}$ . In such a situation, an edge is present between  $s$  and  $r$  if all parameters do *not* have the same value, indicating that not all states have the same probability. In *mgm* we use the parameterization for multinomial regression of *glmnet*, which models the probability of each state of the predicted variable, and codes the first category of the predictor variable as the reference category that is absorbed in the intercept (for details see J. Friedman et al., 2010). This results in  $m \times (u-1)$  parameters, where  $m$  indicates the number of categories of the predicted variable. In this parameterization, an edge is present if *any* of the parameters in  $\theta_{s,r}$  are nonzero, that is,  $(s,r) \in E \iff \exists r : |\theta_{s,r}^z| > 0$ . Therefore, depending on which variables an edge connects it is defined with respect to one or several parameters.

For general  $k$ -order MGMs, an edge between nodes  $s$  and  $r$  is a function of all cliques of size up to  $k$  that include both  $s$  and  $r$ . Therefore, for instance, it is not clear from the graph  $G$  whether the edge  $(s,r)$  is due to a pairwise interaction or from higher order interactions (cliques) that include  $s$  and  $r$ , or both. The number of parameters associated to each clique discussed above for pairwise interactions extends to  $k$ -order interactions. An interaction between  $k$  continuous variables is parameterized by a single parameter  $\theta_{r_1, \dots, r_{k-1}}$  and an interaction between  $k$  categorical variables is parameterized by  $(m_1-1) \times \dots \times (m_k-1)$  parameters, where  $m_1, m_2, \dots, m_k$  are the number of categories of each categorical variable.

In this chapter we focus mainly on the estimation of pairwise MGMs, where each edge is a function of the parameter(s) of a single pairwise interaction. However, in Section 2.3.1 we estimate a higher order MGM and visualize the dependency structure in a factor graph. The factor graph representation has the advantage one can still see on which set of cliques a dependency between two nodes depends (Koller & Friedman, 2009).

### 2.2.3 Estimating Mixed Graphical Models

In this section, we discuss how to estimate the parameters of a joint distribution of the form as in Equation 2.5 from observations. The graphical model  $G$  is then obtained from the parameter estimates as discussed in the previous section.

We know that the joint distribution in Equation 2.5 can be represented as a factorization of univariate conditional distributions. Thus, if we estimate the  $p$  univariate conditional distributions with the parameterization in Equation 2.4, we obtain the joint distribution. Since all univariate conditional distributions are members of the exponential family, it is possible to estimate the joint distribution in Equation 2.5 by a series of  $p$  regressions in the Generalized Linear Model (GLM) framework (see e.g., Nelder & Baker, 1972). From a graphical models perspective this means that we estimate the neighborhood  $N(s)$  of each node  $s \in V$  and then combine all neighborhoods to obtain an estimate of the graph  $G$  (Meinshausen & Bühlmann, 2006).

In order to obtain parameter estimates that are exactly zero (and thereby im-

ply absent edges in the graph) we minimize the negative log-likelihood  $-\mathcal{L}(\theta, X)$  together with the  $\ell_1$ -norm of the parameter vector

$$\hat{\theta} = \arg \min_{\theta} \{-\mathcal{L}(\theta, X) + \lambda \|\theta\|_1\}, \quad (2.8)$$

where  $\|\theta\|_1 = \sum_{j=1}^J |\theta_j|$ ,  $J$  is the length of the parameter vector  $\theta$ , and  $\lambda$  is the regularization parameter that determines the relative weight of the negative log likelihood and the  $\ell_1$ -norm of the parameter vector. The log-likelihood  $\mathcal{L}(\theta, X)$  is defined by the exponential family distribution of the node at hand. In the Gaussian case, minimizing the negative log-likelihood is equivalent to minimizing the squared loss  $-\mathcal{L}(\theta, X) = \|X_s - X_{\setminus s} \theta\|_2^2$ . In other words, we are performing an  $\ell_1$ -penalized (LASSO) regression in the GLM framework with a link-function appropriate for the node at hand (see e.g., Nelder & Baker, 1972). The  $\ell_1$ -penalty ensures that the model is identified in the high-dimensional setting  $p > n$ , where we have more parameters than observations (Hastie, Tibshirani, & Wainwright, 2015).

The design matrix is defined with respect to the conditional distribution of node  $s$  in the  $k$ -order MGM. For example, if  $k = 2$ , the design matrix for the regression on node  $s$  contains all other variables or the corresponding indicator functions (for categorical variables). If  $k = 3$ , the design matrix for the regression on node  $s$  contains all other variables or the corresponding indicator functions, plus the products of all pairs of variables in  $V_{\setminus s}$ , or the  $(m-1) \times (u-1)$  indicator functions in the case of categorical variables with  $m$  and  $u$  categories.

To give non-asymptotic guarantees of false and true positive rates for the  $\ell_1$ -regularized regression estimator it is necessary to put a lower bound  $\tau$  on the size of the parameters in the true model. This assumption is often called the *beta min condition* (see e.g., Hastie et al., 2015). By thresholding estimates at  $\tau$ , we approximately enforce this condition (see e.g., Loh & Wainwright, 2012). For estimating the joint distribution in Equation 2.5 we show in (Haslbeck & Waldorp, 2015) that

$$\tau \asymp s_0 \sqrt{\log \frac{p}{n}} \leq s_0 \lambda, \quad (2.9)$$

where  $s_0$  is the true number of neighbors. If all variables are continuous, the number of neighbors is equal to the number of nonzero parameter estimates  $s_0 = \|\theta^*\|_0$ , where  $\theta^*$  is the true parameter vector. In the case of categorical variables, interactions are parameterized by several parameters. In this case the categorical neighbor is present if at least one of the parameters defining the interaction is nonzero. Since the true parameter vector  $\theta^*$  is unknown, we plug in the estimated parameter vector  $\hat{\theta}$  to obtain the *estimated* number of neighbors  $\hat{s}_0 = \|\hat{\theta}\|_0$ . For interactions involving more than one parameter, we plug in the aggregated parameter (see Algorithm 1). Note that *mgm* allows to switch off this thresholding (see Section 2.3.1). Of course, switching off the thresholding gives a solution that does not have the guarantees of false and true positive rates.

We determine whether an edge is present or not as described in Section 2.2.2. In addition, we compute a *weight* from the set of parameters of each interaction.



If the interaction only involves continuous variables there is only one parameter and we take its value. If the interaction involves categorical variables, we take the mean of the absolute value of all parameters as the weight of the edge. From the nodewise regressions we obtain  $k$  edge-weights for each  $k$ -order interaction. For example, for a pairwise interaction ( $k = 2$ ) between nodes  $s$  and  $r$ , we obtain one parameter  $\theta_{s,r}$  from the regression on  $s$  and  $\theta_{r,s}$  from the regression on  $r$ . To obtain a final conditional dependence graph  $G$  we need to combine these into a final weight. This can be done either by using the OR-rule (take arithmetic mean of  $k$  parameter estimates) or the AND-rule (take arithmetic mean of  $k$  parameter estimates if all parameter estimates are nonzero, otherwise set the parameter to zero). Algorithm 1 summarizes this procedure:

**Algorithm 1** (*Estimating Mixed Graphical Models via Neighborhood Regression*)

1. For each  $s \in V$ 
  - (a) Construct design matrix defined by  $k$ , the order of the MGM
  - (b) Solve the lasso problem in Equation 2.8 with regularization parameter  $\lambda$
  - (c) Threshold the estimates at  $\tau$
  - (d) Aggregate interactions with several parameters into a single edge-weight
2. Combine the edge-weights with the AND- or OR-rule
3. Define  $G$  based on the zero/nonzero pattern in the combined parameter vector

The regularization parameter  $\lambda$  can be selected using cross-validation or a model-selection criterion such as the Extended Bayesian Information Criterion (EBIC):

$$EBIC_{\gamma}(\hat{\theta}) = -2\mathcal{L}(\hat{\theta}) + \hat{s}_0 \log n + 2\gamma \hat{s}_0 \log p, \quad (2.10)$$

where  $\mathcal{L}$  is the log likelihood of the conditional density specified by the estimated parameter vector  $\hat{\theta}$ ,  $\hat{s}_0$  is the number of nonzero neighbors in the candidate model, and  $\gamma$  is a tuning parameter. Note that if  $\gamma = 0$  the EBIC is equal to the BIC (Schwarz et al., 1978). The EBIC has been shown to perform well asymptotically in selecting sparse graphs (Foygel & Drton, 2010, 2014) for any value of  $\gamma$ . In practice, the choice of  $\gamma$  will control the trade-off between sensitivity and precision. (Foygel & Drton, 2010) used values  $\gamma \in \{0, .25, .5, .75, 1\}$  and showed that increasing  $\gamma$  from 0 to 0.25 led to a considerable decrease in false positives, without increasing false negatives too much. We therefore adopted  $\gamma = 0.25$  as a default value. However, to make an *optimal* choice for  $\gamma$ , it is necessary to take into account the true model, the number of available observations and the cost of false positives and false negatives. While the true model is unknown in real data, a reasonable  $\gamma$  can be selected by running a simulation study roughly reflecting the scenario at hand and choosing the  $\gamma$  with the most desirable performance. To this end we provide flexible sampling functions (see Section 2.3).

The computational complexity of Algorithm 1 is  $\mathcal{O}(p \log(p2^{k-1}))$ . Thus the algorithm does not scale well for large  $k$ , the order of interactions in the MGM.

However, in most situations  $k$  will be small, because interactions with a high order are increasingly difficult to interpret and therefore often not of interest.

Note that using a single regularization parameter  $\lambda$  for a model including different edge types may lead to a different penalization for different edge types. This is because edge-parameters are scaled with the sufficient statistic they are associated with and this scaling can differ across exponential family members. While we can bring Gaussian variables on the same scale by subtracting their mean and dividing by their standard deviation, this is not possible for categorical or Poisson random variables. A potential solution would be to introduce a different penalization parameter for each edge type. But this would make the selection of regularization parameters  $\lambda$  considerably more complicated, because now a  $u$ -dimensional space of  $\lambda$  values has to be searched, where  $u$  is the number of different edge types. This is why we currently do not have a procedure in *mgm* that allows different penalties for different edge types.

The performance of Algorithm 1 depends on the number of variables, the order of interactions, type of variables, the size of parameters relative to the variance of associated variables, the sparsity of the parameter vector and the structure of the dependency graph. The best way to determine the performance for a given situation is therefore to obtain it with a simulation study. To this end *mgm* provides a flexible function to sample from MGMs such that the performance of Algorithm 1 in a given situation can be evaluated via simulations.

### 2.2.4 Mixed Autoregressive Models

In Vector Autoregressive (VAR) models, each node  $s$  at time point  $t$  is modeled as a linear combination of all variables (including  $s$ ) at a set of earlier time points. The standard VAR model is defined with a Gaussian noise process, such that the model can be split up into  $p$  conditional Gaussian distributions (see e.g. Hamilton, 1994; Pfaff, 2008a). Instead of a univariate conditional Gaussian distribution, one can also associate other univariate exponential family members with a given node. This leaves us with an almost identical model and estimation problem as discussed in the previous section (Section 2.2.3). The only difference is that the canonical parameter of the node-conditional at hand is not a function of parameters associated with interactions of variables at the *same time point*, but a function of parameters associated with variables at *previous time points*. To distinguish this VAR model over mixed variables from the VAR model that is typically defined with only Gaussian variables, we call this model *mixed Autoregressive (mVAR) model*.

The mVAR model can be estimated by estimating the parameters of the conditional probability of each variable  $s$  as a function of all variables (including itself) at a set of specified previous time points, denoted by  $L$ . For example,  $L = \{1, 2, 3\}$  specifies a VAR model with lags 1, 2 and 3. We introduce a time index as a superscript  $t$  for all variables since we are now dealing with time-ordered observations. We then define the canonical parameter  $E_s^t(X)$  of the conditional distribution  $P(X_s^t | X^{t-1}, X^{t-2}, X^{t-3})$  at time  $t$

$$E_s^t(X) = \theta_s + \sum_{j \in L} \sum_{r \in N(s)} \theta_{s,r}^{t-j} \phi_r(X_r^{t-j}). \quad (2.11)$$

We only included pairwise interactions because *mgm* does not implement higher order interactions for mVAR models. The canonical parameter function in Equation 2.11 defines the log-likelihood  $\mathcal{L}(\theta, X)$  in Equation 2.8 and we can therefore use Algorithm 1 with two modifications: first, we define the design matrix as a function of the included lags  $L$  instead of the maximal order of the interactions, which we here fix to  $k = 2$  (only pairwise interactions). Second, we do not apply an AND/OR rule, because the cross-lagged effect of  $X_s^{t-1}$  on  $X_r^t$  is a different effect than the cross-lagged effect of  $X_r^{t-1}$  on  $X_s^t$  and thus no parameter is estimated twice. Here we state the modified algorithm explicitly:

**Algorithm 2** (*Estimating mixed VAR models via nodewise regression*)

1. For each  $s \in V$ 
  - (a) Construct design matrix defined by  $L$ , the set of included lags
  - (b) Solve the lasso problem in Equation 2.8 with regularization parameter  $\lambda$
  - (c) Threshold the estimates at  $\tau$
2. Define the directed graphs  $D_j$  based on the zero/nonzero pattern in the combined parameter vector for each lag  $j \in L$ .

The computational complexity of Algorithm 2 is  $\mathcal{O}(p \log(p|L|))$ . Similarly to Algorithm 1, the regularization parameter  $\lambda$  can be selected using cross validation or an information criterion such as the EBIC.

Note that the directed graphs in the  $p \times p \times |L|$  array  $D$  are not encoding conditional independence statements as the graph  $G$  for MGMs. But they are a useful summary of the parameters of the mixed VAR model, especially because it allows a visualization as a series of directed networks (see Section 2.3.1 for illustrations).

Note that the performance of Algorithm 2 depends on the number of variables and the number of lags, the type of variables, the size of parameters relative to the variance of associated variables, the sparsity of the parameter vector and the structure of the dependency graph. *mgm* offers a flexible function to sample from mixed VAR models such that the performance of Algorithm 2 in a given situation can be evaluated via simulation studies. (Haslbeck, Bringmann, & Waldorp, 2020) report the performance of Algorithm 2 in recovering VAR models with Gaussian noise process in a variety of situations.

## 2.2.5 Estimating time-varying models

For both MGMs (Section 2.2.2) and mixed VAR models (Section 2.2.4) for time series data, we assumed so far that the models are stationary. This means that the observations at each time  $t$  point are generated from the same distribution parameterized by  $\theta$ . In time-varying models we relax this assumption, such that the parameters  $\theta^t$  can be different at each time point  $t \in \mathcal{T}_n = \{\frac{1}{n}, \frac{2}{n}, \dots, 1\}$ , where

$n$  is the number time points in the time series. Note that we use  $n$  to denote the number of observations both for cross-sectional data (observations are measurements of different systems from some population) and time series data (repeated measurements of the same system).

Since one cannot estimate a model from a single time point, we have to make assumptions about how the parameters of the true model vary as a function of time. These assumptions are usually assumptions about *local stationarity* (e.g., Zhou, Lafferty, & Wasserman, 2010a) and come in one of two flavors: we either assume that there exists a partition  $\mathcal{B}$  of  $\mathcal{T}_n$  in which time points are consecutive and in each of the subsets  $B \in \mathcal{B}$  the model is stationary, that is,  $\forall i, j \in B : \theta^i = \theta^j$ . These piecewise constant time-varying models can be estimated with a fused lasso penalty, which puts an additional penalty on parameter changes from one time point to the subsequent time point (see e.g., R. P. Monti et al., 2014; Kolar & Xing, 2012; Gibberd & Nelson, 2015, 2017).

The other type of local stationarity, which we focus on in this chapter, requires that the model  $\theta^t$  is a smooth function of time. In this case we can combine observations close in time for estimation, because we know that their generating models are similar. This idea is implemented by fitting local models  $\hat{\theta}^{t^e}$  across the time series, which only give high weight to data points close to the given estimation point  $t^e$ . The weight function is usually non-negative and symmetric over  $t^e$  (see e.g., Song, Kolar, & Xing, 2009; Zhou, Lafferty, & Wasserman, 2010b; Kolar, Song, Ahmed, & Xing, 2010; Kolar & Xing, 2009; Tao, Huang, Wang, Xi, & Li, 2016; X. Chen & He, 2015). The full time-varying model is then the set of all local estimates  $\{\theta^{e_1}, \theta^{e_2}, \dots, \theta^{|\mathcal{E}|}\}$  at estimation points  $\mathcal{E} = \{t_1^e, t_2^e, \dots, t^{|\mathcal{E}|}\}$ , where the entries in  $\mathcal{E}$  are usually equally spaced across the time series and the number of estimation points  $|\mathcal{E}|$  is chosen depending on how fine-grained one would like to describe  $\theta^t$  as a function of time  $t$ .

Stating the above formally, we estimate the model  $\theta^{t^e}$  at time point  $t^e$  by minimizing a weighted version of the loss function in Equation 2.8 in Section 2.3.1.1

$$\hat{\theta}^{t^e} = \arg \min_{\theta} \left\{ -\frac{1}{\sum_{t=1}^n w_t^{t^e}} \sum_{t=1}^n w_t^{t^e} \mathcal{L}(\theta, X^t) + \lambda \|\theta\|_1 \right\}, \quad (2.12)$$

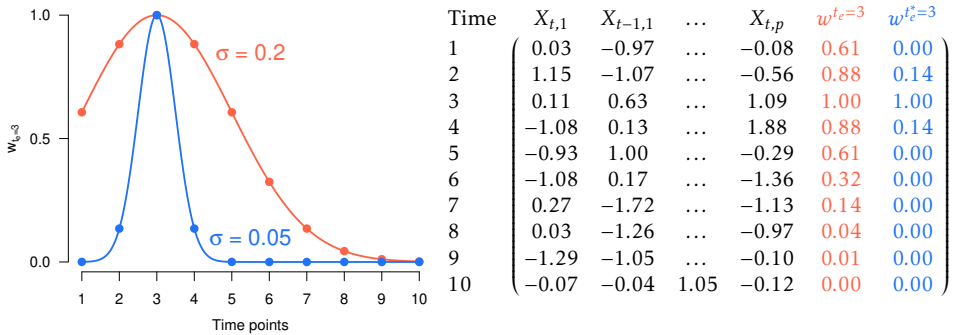
where  $w_t^{t^e}$  is a function of  $t$  defined by a kernel centered over  $t^e$ . Specifically, we define the weight function  $w_t^{t^e}$  to be a Gaussian kernel, normalized such that the largest weight is equal to one (Zhou et al., 2010a)

$$w_t^{t^e} = \frac{Z_t}{\max_{(t \in \mathcal{T}_n)} \{\cup_t Z_t\}}, \quad \text{where } Z_t = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(t-t^e)^2}{2\sigma^2}\right\}. \quad (2.13)$$

This particular scaling of the weight function has the convenient property that the sum of all weights  $n_{\sigma, t^e} = \sum_{t=1}^n w_t^{t^e}$  (or the area under the curve) used at a given estimation point  $t^e$  indicates amount of data used for estimation at  $t^e$  relative to the full time series (the full rectangle). Note that we indexed  $n_{\sigma, t^e}$  also with the estimation point  $t^e$ , because less data is used at the beginning and the

end of the time series, where the weighting function is truncated (see left panel Figure 2.1).

The example in Figure 2.1 illustrates this estimation procedure. Here we have a time series of  $n = 10$  measurements of  $p$  continuous variables, and we would like to estimate the model at time point  $t^e = 3$ . To this end we first define a kernel function  $w_t^{t^e}$  as in Equation 2.13. The bandwidth  $\sigma$  of the kernel, which is here equal to the standard deviation of the Gaussian distribution, indicates how many observations close in time we combine to estimate the node at estimation point  $t^e$ .



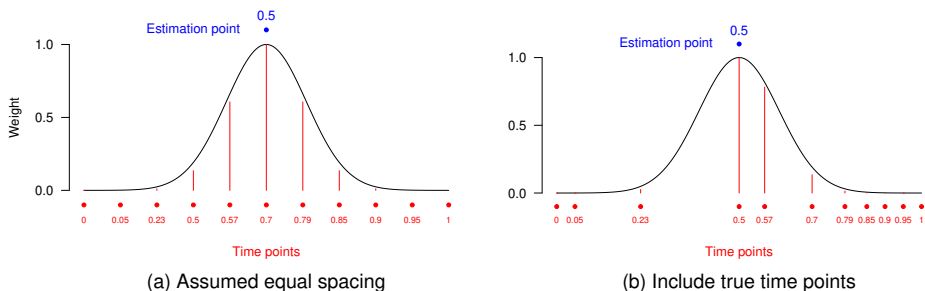
**Figure 2.1:** Illustration of two kernel weighting functions with different bandwidth parameter defined for the estimation point  $t^e = 3$ ; left panel: weights as a function of time; right panel: equivalent representation of the weights across time, combined with the time series data.

Figure 2.1 displays the kernel function  $w_t^{t^e}$  for two different choices of bandwidth,  $\sigma = 0.05$  and  $\sigma = 0.2$ . The kernel function with  $\sigma = 0.05$  gives only time points very close to  $t^e = 3$  a nonzero weight, while other time points get a weight close to zero and have therefore almost no influence on the parameter estimated at  $t^e = 3$ . In contrast, the kernel function with  $\sigma = 0.2$  distributes weights more evenly, which implies that also time-points quite distant from  $t^e = 3$  influence the parameter estimates at  $t^e = 3$ . The values of both weighting functions at the measured time points are also illustrated together with the data matrix in the right panel of Figure 2.1.

The choice of bandwidth involves a trade-off between the sensitivity to time-varying parameters and the stability of the estimates: if we combine only a few observations close in time (small bandwidth  $\sigma$ ) the algorithm can detect parameter-variation at small time scales, however, because we use little data, the estimates will be unstable. If we combine many observations around the estimation point (large  $\sigma$ ), parameter-variation at small time scales will be lost due to aggregation, however, the estimates will be relatively stable. Note that if we keep increasing the bandwidth  $\sigma$ , the weights on  $[0, 1]$  will converge to a uniform distribution and give the same estimates as the stationary version of the model, thereby becoming relatively stable, but losing all sensitivity to detect changes in parameters over time.

The ideal bandwidth  $\sigma^*$  results in the estimated parameter vector  $\hat{\theta}^t$  which minimizes the distance to the true time-varying model  $\hat{\theta}^{t^*}$  as a function of  $\sigma$ . We can estimate the ideal bandwidth  $\sigma^*$  using a time-stratified cross-validation scheme, where one searches a specified  $\sigma$ -sequence and selects the  $\sigma$  which minimizes the mean (across folds and variables) out of sample prediction error (see Section 2.3.3 for a description of the time-stratified cross-validation scheme).

So far we assumed that the measurements in the time-series are taken at equal time intervals. But this need not be the case, because measurements can be missing randomly or by design. Simply treating the time points as equally distributed leads to an incorrect estimate of the time-varying model. Figure 2.2 illustrates this issue:



**Figure 2.2:** Left panel: the weighting function is computed by assuming that the true time points are equally spaced. Since the true time points are not equally spaced, this creates a mismatch between the time scale of estimation points and the true time scale; right panel: the true time interval is used to compute the weights and hence the two scalings match.

Here we have a time series with  $n = 10$  time points, measured at irregular time intervals. In Figure 2.2(a) we distribute these time points evenly across the time interval, which results in that the assigned time points in the normalized time interval  $[0, 1]$  do not correspond to the true time points (values in red). Now if we estimate the time-varying model at time point  $t^e = 0.5$ , we see that the true time point 0.7 gets the highest weight. Thus, the model at  $t^e = 0.5$  is more strongly influenced by the observations at the true time point 0.7 than by the observations at time point 0.5. Clearly, this is undesirable.

In Figure 2.2(b) we avoid this problem by using the true time points in order to define the weighting function  $w_t^{t^e}$ . We again estimate the model at  $t^e = 0.5$  and see that the time scale of the estimation point is now aligned with the true time scale. This results in a different amount of data used for estimation  $n_{\sigma, t^e}$ , depending on how many measurements are available around a given time point. If there is less data available, the algorithm becomes more conservative, since we plug in  $n_{\sigma, t^e}$  for  $n$  in the  $\tau$  threshold in Equation 2.9. In the extreme case where there is no data close to an estimation point,  $n_{\sigma, t^e}$  will be extremely small, which implies that the algorithm sets all estimates to zero. This makes sense, because if there is no data available close to a given estimation point  $t^e$ , we cannot expect to obtain reliable estimates at  $t^e$ .

Note that the only difference between the stationary models and the time-varying models is that we introduce a weight for each time point in the cost function in Equation 2.12 and repeatedly estimate the model at different estimation points. Therefore we can easily adapt the estimation algorithms for the stationary MGM (Algorithm 1) and mixed VAR model (Algorithm 2) to their time-varying versions. We first state the algorithm for time-varying MGMs.

**Algorithm 3** (*Estimating time-varying MGMs via kernel-smoothed neighborhood regression*)

1. For each estimation point  $t^e \in \mathcal{E}$ 
  - (a) For each variable  $s \in V$ 
    - i. Construct design matrix defined by  $k$ , the order of the MGM
    - ii. Solve the weighted lasso problem in Equation 2.12 with regularization parameter  $\lambda$  and the weighting function  $w^{t^e}$  defined by  $t^e$  and bandwidth  $\sigma$
    - iii. Threshold the estimates at  $\tau_{n_\sigma, t^e}$
  - (b) Combine the parameter estimates with the AND- or OR-rule
  - (c) Define  $G^e$  based on the zero/nonzero pattern in the combined parameter vector  $\theta^e$

Thus we obtain a parameter vector  $\theta^{t^e}$  of the MGM in Equation 2.5 and a graph  $G^{t^e}$  defined by  $\theta^{t^e}$ , for each estimation point  $t^e \in \mathcal{E}$ . From Algorithm 1 follows that Algorithm 3 has a computational complexity of  $\mathcal{O}(|\mathcal{E}|p \log(p2^{k-1}))$ .

Similarly, we can adapt Algorithm 2 for the estimation of time-varying mixed VAR models:

**Algorithm 4** (*Estimating time-varying mixed VAR models via kernel-smoothed neighborhood regression*)

1. For each estimation point  $t^e \in \mathcal{E}$ 
  - (a) For each variable  $s \in V$ 
    - i. Construct design matrix defined by the  $L$ , the set of included lags
    - ii. Solve the weighted lasso problem in Equation 2.12 with regularization parameter  $\lambda$  and the weighting function  $w^{t^e}$  defined by  $t^e$  and bandwidth  $\sigma$
    - iii. Threshold the estimates at  $\tau_{n_\sigma, t^e}$
  - (b) Define the directed graphs  $D_j^e$  based on the zero/nonzero pattern in the parameter vector  $\theta^e$  for each lag  $j \in L$ .

Here we obtain a parameter vector  $\theta^{t^e}$  of the mVAR model and a directed graph  $D_j^{t^e}$  for each lag, defined by  $\theta^{t^e}$ , for each estimation point  $t^e \in \mathcal{E}$ . From Algorithm 2 follows that Algorithm 4 has a computational complexity of  $\mathcal{O}(|\mathcal{E}|p \log(p|L|))$ . Haslbeck et al. (2020) report the performance of Algorithm 4 in

recovering time-varying VAR models with Gaussian noise process for a variety of situations.

Fitting a time-varying model with the above method requires to specify an appropriate bandwidth parameter  $\sigma$ . In Section 2.3.3, we describe a time-stratified cross-validation scheme to select  $\sigma$  in a data-driven way. The EBIC is not suitable to select  $\sigma$ . The reason is that threshold (intercept) parameters are neither included in the  $\ell_1$ -penalty, nor in the EBIC. This results in the EBIC selecting always the model with the smallest specified bandwidth, which includes no interaction parameters, but achieves an extremely good fit through highly local (time-varying) thresholds (intercepts). This problem is avoided when using a cross-validation scheme, where fitting local means leads to high out-of-fold prediction error.

Note that the performance of Algorithm 3 and 4 depends on the number of variables, the type of variables, the size of parameters relative to their variance, the sparsity of the parameter vectors, the structure of the dependency graph and the how non-linear the parameters vary as a function of time. The best way to obtain the performance of Algorithm 3 and 4 is to set up a suitable simulation study. To this end *mgm* offers flexible functions to sample from time-varying MGMs and time-varying mVAR models.

### 2.3 Usage and Examples

The *mgm* package can be installed from the Comprehensive R Archive Network (CRAN) (<http://CRAN.r-project.org/>):

```
install.packages("mgm")
library(mgm)
```

In the following sections, we show for each of the four model types how to

1. sample observations from a specified model
2. estimate the model from data
3. make predictions from an estimated model
4. visualize the estimated model
5. and assess the stability of estimates.

The sampling functions are included to enable the user to determine the performance of the estimation algorithm in a specific situation via simulations. All used datasets are loaded automatically with the *mgm*-package. All analyses in the chapter are fully reproducible, and the necessary code is either shown in the chapter or can be found in the online supplementary material or the Github repository <https://github.com/jmbh/mgmDocumentation>. For all code examples we use the *mgm* version 1.2-9 and R-version 3.6.



### 2.3.1 Stationary Mixed Graphical Models

In this section we first use a simulated data set to show how to estimate a pairwise MGM, compute predictions from it, visualize it and assess the stability of its parameters. Then we fit a pairwise MGM to a larger empirical data set related to Autism Spectrum Disorder (ASD). Finally, we give an example of a higher-order MGM by showing how to estimate a  $k = 3$  MGM to a data set consisting of symptoms of Post-traumatic Stress Disorder (PTSD).

#### 2.3.1.1 Estimating Mixed Graphical Models

In this section we show how to use the function `mgm()` to estimate a pairwise MGM to a data set with  $n = 500$  observations of two continuous, and two categorical with  $m = 2$  and  $u = 4$  categories, respectively. The true model includes the pairwise interactions 1-4, 2-3 and 1-2. For the exact parameterization of the true model and for a description of how to sample from this MGM using the `mgmsampler()` function see the section on sampling below.

Next to the data, we specify the type of each variable ("g" for Gaussian, "p" for Poisson, "c" for categorical) and the number of levels of each variable (1 for continuous variables by convention). Here we use the example data `mgm_data` which is automatically loaded with `mgm`. Next, we indicate the order of the graphical model: we choose  $k = 2$ , which corresponds to a pairwise MGM (containing at most 2-way interactions). If we specified  $k = 3$ , we would fit an MGM including all 2-way and all 3-way interactions,  $k = 4$  would include all 2-way, 3-way and 4-way interactions, etc. After that, we specify how we select the penalization parameter  $\lambda$  in Algorithm 1. The two available options are the EBIC or cross-validation. Here we choose cross-validation with 10 folds. If not otherwise specified via the argument `lambdaSeq`, the considered  $\lambda$ -sequence is determined as in the `glmnet` package: a sequence is defined from  $\lambda_{max}$ , the smallest (data derived) value for which all coefficients are zero, and  $\lambda_{min}$ , a fraction of  $\lambda_{max}$ , which is 0.01 in the high-dimensional setting ( $n < p$ ) and 0.0001 if  $n > p$ . Finally, indicate that estimates across neighborhood regressions should be combined with the AND-rule. Since we use cross-validation, we set a random seed outside the function to ensure that the analysis is reproducible.

```
set.seed(1)
fit_mgm <- mgm(data = mgm_data$data,
              type = c("g", "c", "c", "g"),
              levels = c(1, 2, 4, 1),
              k = 2,
              lambdaSel = "CV",
              lambdaFolds = 10,
              ruleReg = "AND")
```

The function `mgm()` returns a list with the following entries: `fit_mgm$call` returns the call of the function; `fit_mgm$pairwise` contains the weighted adjacency matrix and the signs (if defined) of the parameters in the weighted adjacency matrix; `fit_mgm$interactions`— contains a list that shows all recovered interactions (cliques) and a list that returns the parameters associated with

all cliques; `fit_mgm$intercepts` stores all estimated thresholds/intercepts and `fit_mgm$nodemodels` is a list with the  $p$  *glmnet* objects from which all above results are computed. We inspect the weighed adjacency matrix stored in `fit_mgm$pairwise$wadj`

```
round(fit_mgm$pairwise$wadj, 2)
> [,1] [,2] [,3] [,4]
> [1,] 0.00 0.53 0.00 0.46
> [2,] 0.53 0.00 0.09 0.00
> [3,] 0.00 0.09 0.00 0.00
> [4,] 0.46 0.00 0.00 0.00
```

and see that we correctly recovered the pairwise dependencies 1-4, 2-3 and 1-2. The list entry `fit_mgm$pairwise$signs` indicates the sign for each interaction, if a sign is defined. By default, a sign is only defined for interactions between non-categorical variables (Gaussian, Poisson). Interactions involving categorical variables with  $m > 2$  categories are defined by more than one parameter and hence no sign can be defined. The function `showInteraction()` provides an alternative way to inspect a given interaction. For instance, one can obtain the details about the interaction 1-4 like this:

```
showInteraction(object = fit_mgm,
                int = c(1,4))
> Interaction: 1-4
> Weight: 0.4586544
> Sign: 1 (Positive)
```

We use the *glmnet* package to fit the regularized nodewise regressions, which directly models the probabilities of categorical variables instead of the ratio relative to a reference category. This is possible, because the regularization ensures that this model is identified (for details see J. Friedman et al., 2010). This means that an interaction between two categorical variables  $X_1 \in \{1, \dots, m\}$  and  $X_2 \in \{1, \dots, u\}$  has  $m \times (u - 1)$  parameters in the regression on  $X_1$  and  $u \times (m - 1)$  parameters in the regression on  $X_2$ . In addition, all estimation functions in *mgm* also allow an overparameterization (specified via the argument `overparameterize = TRUE`), where an indicator function is defined for *each* state of the categorical predictor variable. In the previous example of a pairwise interaction, this leads to  $m \times u$  parameters specifying the interaction between  $X_1$  and  $X_2$ . The overparameterization is useful when one is interested in parameters associated with indicator functions that are otherwise absorbed by the threshold (intercept) parameters (also called reference category). We give an example for estimating a  $k = 3$  order MGM at the end of this section.

If the argument `binarySign` is set to `TRUE`, all binary variables have to be coded as  $\{0, 1\}$  and a sign is defined in the following way: for an interaction between two binary variables  $X_1, X_2 \in \{0, 1\}$ , if the parameter associated with the indicator function  $\mathbb{I}_{X_2=1}$  in the equation modeling  $P(X_1 = 1)$  has a positive sign (which implies that the parameter associated with  $\mathbb{I}_{X_2=1}$  in the equation modeling  $P(X_1 = 0)$  has a negative sign, see (J. Friedman et al., 2010)), then we assign

a positive sign to the binary-binary interaction. For an interaction between a binary variable  $X_1$  and a *continuous* variable  $X_2$  we take the sign of the parameter associated with  $X_2$  in the equation modeling  $P(X_1 = 1)$ . In addition, it is possible to specify a weight for each observation via the argument `weights` to perform weighted regression.

In the example above we used an  $\ell_1$ -penalized GLM to estimate the MGM, which implies that we assume that the true MGM is sparse. However, a different penalty may be appropriate in some situations. Via the argument `alphaSeq` one can specify any convex combination of the  $\ell_1$ - and  $\ell_2$ -penalty (the elastic net penalty, see (Zou & Hastie, 2005)). `alphaSeq = 1` corresponds to the  $\ell_1$ -penalty (default) and `alphaSeq = 0` to the  $\ell_2$ -penalty. If a sequence of values is provided to `alphaSeq`, the function will select the best  $\alpha$  value based on the EBIC or cross validation, specified via the argument `alphaSel`.

### 2.3.1.2 Making Predictions from Mixed Graphical Models

We now use the `predict()` function to compute predictions and nodewise errors from the model estimated in the previous section. This function takes the model object and data of the same format as the data used for estimation as input. It also allows to specify which error functions should be used to compute nodewise prediction errors. The error functions  $F(\hat{y}, y)$  for continuous and categorical variables are specified via the `errorCon` and `errorCat` arguments, respectively. Here we specified the Root Mean Squared Error ("RMSE") and the proportion of explained variance ("R2") as error functions for the continuous variables, and the proportion of correct classification (or accuracy, "CC") and the normalized proportion of correct classification ("nCC") for categorical variables. "nCC" is the increase in accuracy beyond the intercept model, divided by the maximal possible increase, and thereby captures how well a node is predicted by other nodes beyond the intercept model. Specifically, let  $\mathcal{A} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i = \hat{y}_i)$  be the proportion of correct classifications, and let  $p_0, p_1, \dots, p_m$  be the marginal probabilities of the categories, where  $\mathbb{I}$  is the indicator function for the event  $R_i = \hat{R}_i$ . In the binary case these are  $p_0$  and  $p_1 = 1 - p_0$ . We then define normalized accuracy as

$$\mathcal{A}_{\text{norm}} = \frac{\mathcal{A} - \max\{p_0, p_1, \dots, p_m\}}{1 - \max\{p_0, p_1, \dots, p_m\}}.$$

For details see (Haslbeck & Waldorp, 2018). If one is not interested in computing nodewise prediction errors, the arguments `errorCon` and `errorCat` can be simply ignored.

We provide the `mgm` fit object and the data as input arguments and a choice of prediction error measures to the `predict()` function:

```
pred_mgm <- predict(object = fit_mgm,
                    data = mgm_data$data,
                    errorCon = c("RMSE", "R2"),
                    errorCat = c("CC", "nCC"))
```

The output object `pred_mgm` is a list that contains the function call, the predicted values, the predicted probabilities of each category in case the model includes categorical variables, and a table with nodewise prediction errors. Here we print the nodewise error table in the console:

```
pred_mgm$errors
> Variable Error.RMSE Error.R2 Error.CC Error.nCC
> [1,]      1      0.781    0.389      NA      NA
> [2,]      2      NA      NA    0.842    0.225
> [3,]      3      NA      NA    0.342    0.000
> [4,]      4      0.855    0.268      NA      NA
```

The RMSE and  $R^2$  are shown for the two continuous variables, the accuracy and normalized accuracy are shown for the two categorical variables. It is possible to provide an arbitrary number of customary error functions for both continuous and categorical variables to `predict()`, for details see `?predict.mgm`.

In this example we used the same data for estimation and prediction, which means that we computed *within sample* prediction errors. In order to evaluate how well the model generalizes out of sample, the predictions have to be made on a fresh test data set. This can be done by providing new data of the same format to the `predict()` function.

### 2.3.1.3 Visualizing Mixed Graphical Models

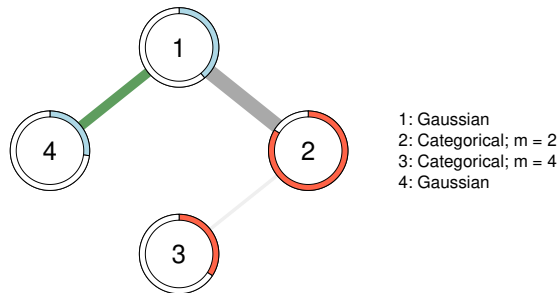
We visualize interaction parameters of the pairwise model together with the nodewise errors using the *qgraph* package (Epskamp et al., 2012). To this end we first install and load the *qgraph* package and compute a vector containing the nodewise errors we would like to display:

```
install.packages("qgraph")
library(qgraph)
errors <- c(pred_mgm$errors[1, 3],
            pred_mgm$errors[2:3, 4], pred_mgm$errors[4, 3])
```

Here we decided to display the proportion of explained variance for the continuous variables and the accuracy for the categorical variables. In order to plot the model, one provides the weighted adjacency matrix and the errors to the function `qgraph()`. We also provide a matrix of edge colors that specify the sign of each interaction (green = positive, red = negative, grey = undefined) that is stored in the *mgm* fit object. Finally we provide colors for the different error measures and variable names for the legend.

```
qgraph(fit_mgm$pairwise$wadj,
       edge.color = fit_mgm$pairwise$edgecolor,
       pie = errors,
       pieColor = c("lightblue", "tomato", "tomato", "lightblue"),
       nodeNames = c("Gaussian", "Categorical; m = 2",
                    "Categorical; m = 4", "Gaussian"),
       legend = TRUE)
```

Figure 2.3 shows the resulting visualization:



**Figure 2.3:** Visualization of the edge-parameters and nodewise errors of the estimated MGM. Green edges indicate positive relationships. Grey edges indicate pairwise interactions for which no sign is defined (interactions involving categorical variables). The width of the edges is proportional to the absolute value of the associated edge-parameter.

The green edge between variable 1 and variable 2 indicates a positive linear relationship between the two Gaussian variables and the two grey edges indicate relationships between categorical variables, for which no sign is defined. The exact nature of these interactions can be found by inspecting them using the output object of the `showInteraction()` function. The width of the edges is proportional to the size of the corresponding edge-parameter. The blue rings indicate the proportion of variance explained by neighboring nodes for the Gaussian variables, and the red rings indicate the accuracy of the categorical nodes.

#### 2.3.1.4 Bootstrap Sampling Distributions

Obtaining the sampling distributions for parameter estimates can be useful if one is interested in the stability of estimates (Hastie et al., 2015). The function `resample()` obtains empirical sampling distributions with the nonparametric bootstrap (Efron, 1992; Efron & Tibshirani, 1994). Its input is the *mgm* model object `fit_mgm` (the output of the function `mgm()`, see above), the data, and the desired number of bootstrap samples  $B$  via the argument `nB`. The argument `quantiles` specifies lower/upper quantiles of the sampling distributions, which are added to the output. Here we choose `quantiles = c(.05, .95)`. Finally, we set a random seed to make the analysis reproducible.

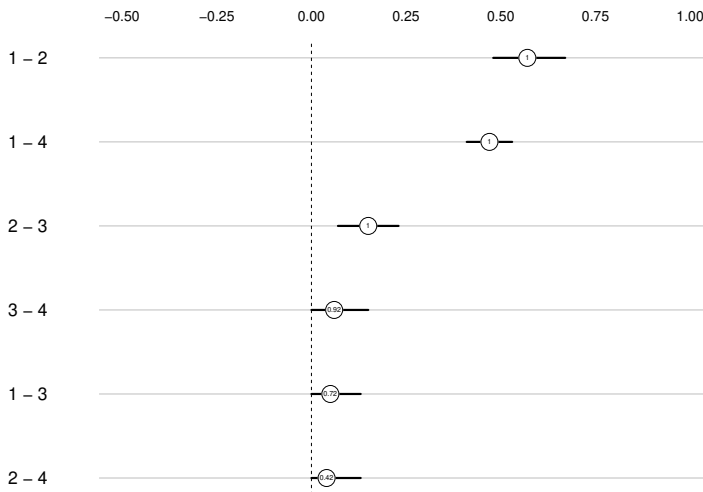
```
set.seed(1)
res_obj <- resample(object = fit_mgm,
                   data = mgm_data$data,
                   nB = 50,
                   quantiles = c(.05, .95))
```

The bootstrapped sampling distributions of the edge weights can be found in a  $B \times p \times p$  array stored in the list entry `res_obj$bootParameters`. For example,

the vector of length  $B$  with the bootstrapped sampling distribution of the weight of the edge 3-4 is stored in the entry `res_obj$bootParameters[, 3, 4]`. The output object `res_obj` also contains the specified lower/upper quantiles of each sampling distribution, the function call, the  $B$  estimated models and the running time for each bootstrap sample. The function `plotRes()` provides a plot that summarizes the bootstrapped sampling distributions. For each edge weight, it displays the proportion of nonzero estimates across all  $B$  models, printed at the arithmetic mean of the sampling distribution. In addition, it displays specified lower/upper quantiles. Here we choose the 5% and 95% quantiles by setting `quantiles = c(.05, .95)`.

```
plotRes(object = res_obj,
        quantiles = c(.05, .95))
```

The resulting plot is displayed in Figure 2.4. It shows that the sampling distributions for the edges 1-2 and 1-4 are located far from zero, have a small standard deviation and 100% of the  $B$  estimates were nonzero. For the edges 3-4, 2-3 and 1-3 that are absent in the true graph, the sampling distribution is close to zero and the proportion of estimated nonzero effects is much smaller.



**Figure 2.4:** Summaries of bootstrapped sampling distributions separately for the weight of each edge. The value indicates the proportion of nonzero estimates across the  $B$  bootstrap samples and is plotted on the arithmetic mean of the sampling distribution. The black horizontal lines indicate the 0.05 and 0.95 quantiles of the bootstrapped sampling distribution.

While bootstrapped sampling distributions are useful to determine the stability of estimates, they are not suited for performing hypothesis tests, for instance, against the null hypothesis that the population parameter is equal to zero. The reason is that the sampling distributions of parameters obtained with  $\ell_1$ -regularized regression have a zero mass around zero (Bühlmann, Kalisch, & Meier, 2014). The R-package *bootnet* (Epskamp, Borsboom, & Fried, 2018) also

implements a bootstrap scheme for *mgm* objects and provides similar plotting options.

### 2.3.1.5 Sampling from Mixed Graphical Models

Here we illustrate how to use the function `mgmsampler()` to create the dataset `mgm_data` that was used for estimation above. These data were created by specifying an MGM consisting of two continuous-Gaussian nodes ("g"), and two categorical nodes ("c") with  $m = 2$  and  $u = 4$  categories, and three pairwise interactions between these four variables. A third option in `mgmsampler()` are Poisson nodes ("p"). Note that we use the overparameterized representation of interactions between categorical variables to specify the model, which means that the pairwise interaction between the categorical variables has  $m \times u$  parameters. We begin by specifying the type and number of categories for each node. By convention, for continuous variables we set the number of categories to 1.

```
type <- c("g", "c", "c", "g")
level <- c(1, 2, 4, 1)
```

Next, we specify a list containing the thresholds for each variable:

```
thresholds <- list()
thresholds[[1]] <- 0
thresholds[[2]] <- rep(0, level[2])
thresholds[[3]] <- rep(0, level[3])
thresholds[[4]] <- 0
```

We specify a zero threshold (intercept) for the two Gaussian nodes, and for each of the categories of both categorical variables. Thresholds correspond to the first summation in the joint MGM density in Equation 2.5. Next, we specify a vector containing the standard deviations for the Gaussian variables:

```
sds <- rep(1, 4)
```

The entries in `sds` corresponding to non-Gaussian nodes (here 2 and 3) are ignored. Finally, we specify three pairwise interactions between the variables 1-2, 2-3 and 1-4 in two steps: first, we create a matrix, in which each row indicates one pairwise interaction:

```
factors <- list()
factors[[1]] <- matrix(c(1,4,
                        2,3,
                        1,2), ncol = 2, byrow = T)
```

We assign the matrix to the first list entry `factors[[1]]`, which contains pairwise interactions. The second list entry `factors[[2]]` contains a  $q \times 3$  matrix of  $q$  3-way interactions, the third entry contains a  $w \times 4$  matrix of  $w$  3-way interactions, etc. Since we only specify pairwise interactions in this example, we only use the first entry. A description and examples of how to specify higher order interactions are given in the help file `?mgmsampler`. In a second step, we specify the parameters of the three interactions:

```
interactions <- list()
interactions[[1]] <- vector("list", length = 3)

# 2-way interaction: 1-4
interactions[[1]][[1]] <- array(.5, dim = c(level[1], level[4]))

# 2-way interaction: 2-3
int_2 <- matrix(0, nrow = level[2], ncol = level[3])
int_2[1, 1:2] <- 1
interactions[[1]][[2]] <- int_2

# 2-way interaction: 1-2
int_1 <- matrix(0, nrow = level[1], ncol = level[2])
int_1[1, 1] <- 1
interactions[[1]][[3]] <- int_1
```

The interaction between the continuous variables 1-2 is parameterized by one parameter with value 0.5. The interaction between the two categorical variables is specified by a  $2 \times 4$  parameter matrix. We give the entries (1, 1) and (1, 2) a value of 1, which means that these two states have a higher probability than the remaining states, which are associated with a value of 0. Finally, we specify the interaction between the continuous-Gaussian node 1 and the binary node 2, which has two parameters associated with the two indicator functions for the binary variable multiplied with the continuous variable. Now we provide these arguments to the `mgmsampler()` function, together with  $n = 500$ , which samples 500 observations from the model:

```
set.seed(1)
mgm_data <- mgmsampler(factors = factors,
                      interactions = interactions,
                      thresholds = thresholds,
                      sds = sds,
                      type = type,
                      level = level,
                      N = 500)
```

The function returns a list containing the function call in `mgm_data$call` and the data in `mgm_data$data`. For more details on how to specify  $k$ -order MGMs we refer the reader to the help file `?mgmsampler`.

### 2.3.1.6 Application: Autism and Well-being

Here we show how to estimate an MGM on a real data set consisting of responses of 3521 individuals from the Netherlands, who were diagnosed with Autism Spectrum Disorder (ASD), to 28 questions on demographics, psychological aspects, conditions of the social environment and medical measurements (for details see Begeer, Wierda, & Venderbosch, 2013; Deserno, Borsboom, Begeer, & Geurts, 2017). The dataset is included in the `mgm` package and loaded automatically. It includes continuous variables, count variables and categorical variables



(see `autism_data_large$type`), and the latter have between 2 and 5 categories (see `autism_data_large$level`)

We choose a pairwise model ( $k = 2$ ) and select the regularization parameters  $\lambda$  using the EBIC with a hyperparameter  $\gamma = 0.25$ :

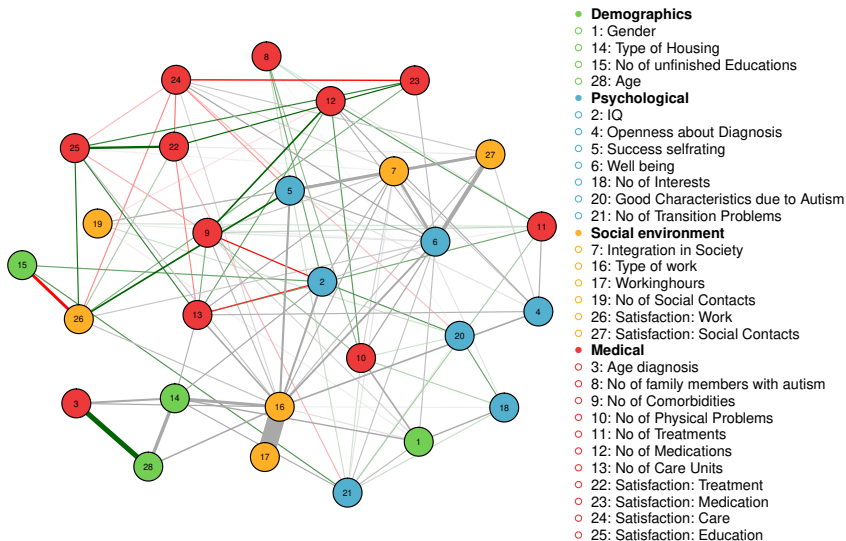
```
fit_ADS <- mgm(data = autism_data_large$data,
               type = autism_data_large$type,
               level = autism_data_large$level,
               k = 2,
               lambdaSel = "EBIC",
               lambdaGam = 0.25)
```

The  $28 \times 28$  weighted adjacency matrix is too large to be displayed here. Instead, we directly visualize it using the `qgraph` package. In addition to the weighted adjacency matrix and the matrix containing edge colors that indicate the signs of edge parameters, we also provide a grouping of the variables into the categories *Demographics*, *Psychological*, *Social environment* and *Medical* measurements as well as colors for the grouping, both of which are contained in the data list `autism_data_large`. The remaining arguments are chosen to improve the visualization, for details we refer the reader to the help file `?qgraph`.

```
qgraph(fit_ADS$pairwise$wadj,
       layout = "spring", repulsion = 1.3,
       edge.color = fit_ADS$pairwise$edgecolor,
       nodeNames = autism_data_large$colnames,
       color = autism_data_large$groups_color,
       groups = autism_data_large$groups_list,
       legend.mode="style2", legend.cex=.4,
       vsize = 3.5, esize = 15)
```

The resulting visualization is shown in Figure 2.5. The layout of node positions was computed with the Fruchterman Reingold algorithm, which places nodes such that all the edges are of more or less equal length and there are as few crossing edges as possible (Fruchterman & Reingold, 1991). Green edges indicate positive relationships, red edges indicate negative relationships and grey edges indicate relationships involving categorical variables for which no sign is defined. The width of the edges is proportional to the absolute value of the edge-weight. The node color indicates to the different categories of variables.

We observe a strong positive relationship between *age* and *age of diagnosis*, which makes sense because the two variables are logically connected. The negative relationship between *number of unfinished educations* and *satisfaction at work* seems plausible, too. Well-being is strongly connected in the graph, with the strongest connections to *satisfaction with social contacts* and *integration in society*. These three variables are categorical variables with 5, 3 and 3 categories, respectively. In order to investigate the exact nature of these interactions, one can look up all parameters using the function `showInteraction()`.



**Figure 2.5:** Visualization of the MGM estimated on the autism dataset. Green edes indicate positive relationships, red edges indicate negative relationships and grey edges indicate relationships involving categorical variables for which no sign is defined. The width of the edges is proportional to the absolute value of the edge-parameter. The colors of the nodes map to the different domains *Demographics, Psychological, Social Environment* and *Medical*.

### 2.3.1.7 Estimating higher-order Mixed Graphical Models

In the previous section, we focused on the estimation of pairwise ( $k = 2$ ) MGMs. Here, we show how to estimate an MGM of order  $k = 3$  to a dataset consisting of Post-traumatic Stress Disorder (PTSD) symptoms reported from 344 survivors of the Wenchuan earthquake in China reported in McNally et al. (2015). The data is loaded automatically with *mgm* and includes the following symptoms:

```
PTSD_data$names
> [1] "intrusion" "dreams" "flash" "upset" "physior" "avoidth"
```

We first specify the data, type, levels and the desired method to select the regularization parameter  $\lambda$ , similarly to the pairwise MGM. But here we specify with  $k = 3$  to estimate *all* pairwise and *all* 3-way interactions.

In addition, we choose to use the overparameterized version of the representation of categorical variables by setting `overparameterize = TRUE`. This results in that all states of categorical variables up to degree  $k$  are modeled explicitly. This overparameterization is possible due to the  $\ell_1$ -penalization (for details see J. Friedman et al., 2010). The standard and the overparameterized parameterization are statistically equivalent and therefore one has to choose one over the other based on which parameterization lends itself to the most useful interpretation in a given application: if it is more sensible to compare all categories to a reference category the standard parameterization is preferable. If one is interested in all

categories equally, the overparameterization might be better. We call `mgm()` with the above discussed specifications:

```
fit_mgmk <- mgm(data = PTSD_data$data,
               type = PTSD_data$type,
               level = PTSD_data$level,
               lambdaSel = "EBIC",
               lambdaGam = 0.25,
               k = 3,
               overparameterize = TRUE)
```

The output object `fit_mgmk` has the same structure as the pairwise MGM discussed above. We still find the pairwise interactions in `fit_mgmk$pairwise` but these do not represent the full parameterization anymore, since we also estimated 3-way interactions. All interactions that have been estimated to be nonzero can be found in the list `fit_mgmk$interactions`: the entry `fit_mgmk$interactions$indicator` contains a list showing all nonzero estimated interactions, separately for each order (here 2 and 3):

```
fit_mgmk$interactions$indicator
[[1]]
[,1] [,2]
[1,]  1  2
[3,]  4  5

[[2]]
[,1] [,2] [,3]
[1,]  1  3  4
[2,]  1  3  5
[3,]  2  3  4
[4,]  3  5  6
[5,]  4  5  6
```

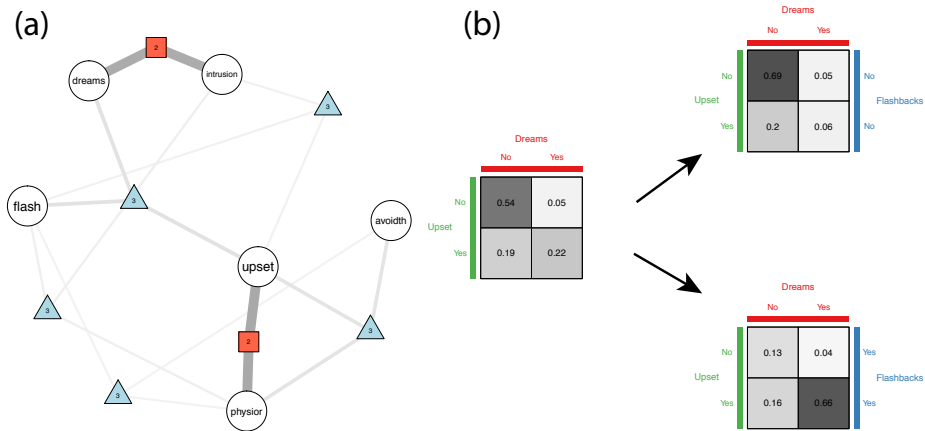
The output indicates that we estimated two nonzero pairwise interactions, and five nonzero 3-way interactions. For example, the third row in the second list entry indicates that there is a 3-way interaction between variables 2-3-4 (*Dreams*, *Flashbacks* and *Upset*). The list `fit_mgmk$interactions` also contains additional entries for the strength of each interaction, and all parameters specifying the interaction (more than one parameter in case of categorical variables, see Section 2.2.1).

If we were to visualize the dependency structure of this  $k = 3$ -order MGM in a common undirected graph, we would lose the information about on which interaction(s) a dependency (edge) is based on. For instance, an edge between the nodes 1 and 2 could either be due to a pairwise interaction between 1 and 2, or due to any 3-way interaction including the nodes 1 and 2, or both. A visualization that allows to represent different orders of interactions is the factor graph (e.g. Koller & Friedman, 2009). A factor graph is a bipartite graph that includes nodes for variables on the one hand, and nodes for interactions on the other hand. We use the function `FactorGraph()` to plot such a factor graph

## 2. Estimating Mixed Graphical Models

```
FactorGraph(object = fit_mgmk,
            labels = PTSD_data$names,
            PairwiseAsEdge = FALSE)
```

which results in Figure 2.6 (a). The six circle nodes represent the six variables in the dataset. The red square factor nodes indicate pairwise interactions and the blue square factor nodes indicate 3-way interactions. Each factor node connects to two (pairwise) or three (3-way) variables, indicating an interaction between the respective variables. The width of the edges are proportional to the absolute value of the weight of the corresponding interaction.



**Figure 2.6:** (a) Factor graph visualization of the estimated  $k = 3$  MGM. The circle nodes refer to variables, the quadratic nodes refer to factors over two variables, and the triangle nodes refer to factors over three variables. The width of the edges is proportional to the strength of the factor; (b) the marginal sample probability cross-table of *Dreams* and *Upset*, and the same table conditioned on the two states of *Flashbacks*. We see that the relationship between *Dreams* and *Upset* depends on *Flashbacks*

We have a closer look at the 3-way interaction 2-3-4 (*Dreams*, *Flashbacks*, and *Feeling Upset*) in Figure 2.6 (b): First look at the marginal probability cross-table of the variables *Dreams* and *Upset*, which shows unequal cell probabilities and hence an interaction between those two variables. Now we condition on the two states of a third variable *Flashbacks* and see that the interaction between *Dreams* and *Upset* considerably depends on whether an individual has *Flashbacks* or not.

Interpreting a  $k$ -way interaction by interpreting the  $k - 1$  way interaction for several levels of one of the variables  $X_j$  in the interaction can be seen as a moderation by  $X_j$ . In Haslbeck, Borsboom, and Waldorp (2019) we explain this approach of interpreting  $k = 3$  interactions as moderation in more detail, and provide further examples for estimating and interpreting higher-order MGMs for the special case of continuous variables.

## 2.3.2 Stationary mixed VAR models

In this section, we first show how to estimate a mixed VAR model, compute predictions from it and visualize it based on simulated data. Then we show how to specify and sample from a mixed VAR model in order to generate the data used earlier for estimation. We then fit a mixed VAR model of order 3 to resting state fMRI data.

### 2.3.2.1 Estimating mixed VAR models

Here we show how to use the function `mvar()` to fit a mixed VAR model to a time series of six variables, consisting of four categorical variables (with 2, 2, 4 and 4 categories) and two Gaussian variables. In the true mVAR model from which the time series was sampled, there are effects of lag order 1 from variable 6 on 5, from 5 on 1 and from 3 on 1. The exact parameterization of these interactions is shown in below in this section, where we create this data set with the function `mvarsampler()`.

We provide the data (which is an example dataset automatically loaded with `mgm`), and specify the type of each variable in `type`, where "g" stands for Gaussian, "p" for Poisson, and "c" for categorical. Next, we provide the number of levels for each variable via `levels`, where we choose 1 for continuous variables by convention. We specify a lag of order 1 and select the regularization parameters  $\lambda$  using the EBIC with tuning parameter  $\gamma = .25$ :

```
fit_mvar <- mvar(data = mvar_data$data,
                type = c("c", "c", "c", "c", "g", "g"),
                level = c(2, 2, 4, 4, 1, 1),
                lambdaSel = "EBIC",
                lambdaGam = .25,
                lags = 1)
```

The function `mvar()` returns a list with several entries: `fit_mvar$wadj` is a  $p \times p \times |L|$  array of edge weights, where  $|L|$  is the number of specified lags. For example, `fit_mvar$wadj[3, 5, 1]` corresponds to the parameter for the crossed lagged effect of 5 on 3 over the first lag specified in `lags` (in this example we only specified one lag). `fit_mvar$signs` has the same dimension as `fit_mvar$wadj` and contains the signs of all parameters, if defined. `fit_mvar$rawlags` contains the full parameterization of the cross-lagged effects. If the mixed VAR model contains only continuous variables, the information in `fit_mvar$wadj` and `fit_mvar$rawlags` is equivalent. Similarly to `mgm()`, the entry `fit_mvar$intercepts` contains all thresholds (intercepts) and `fit_mvar$nodemodels` contains the  $p$  *glmnet* models of the  $p$  neighborhood regressions. Here we show the interaction parameters of the fitted VAR model for the single specified lag of order 1:

```
round(fit_mvar$wadj[, , 1], 2)
[,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0 0.33 0.06 0.41 0.00
```

```
[2,]  0  0 0.00 0.00 0.00 0.00
[3,]  0  0 0.00 0.00 0.00 0.00
[4,]  0  0 0.00 0.00 0.00 0.00
[5,]  0  0 0.00 0.00 0.00 0.31
[6,]  0  0 0.00 0.00 0.00 0.00
```

The autoregressive effects are on the diagonal and the cross-lagged effects are on the off-diagonal. We use a representation in which columns predict rows, which means that the entry `fit_mvar$wadj[5, 3, 1]` corresponds to the cross-lagged effect of 3 on 5 at lag 1. Comparing the estimates with the true cross-lagged effects indicated above, we see that all three true cross-lagged effects have been recovered and all other effects are correctly set to zero.

The additional arguments that can be provided to `mvar()` are similar to the ones in `mgm()`: the regularization parameter  $\lambda$  can be selected using the EBIC with a specified hyperparameter  $\gamma$  or with cross-validation with a specified number of folds. The candidate  $\lambda$  sequence is computed as in `mgm()` (see Section 2.3.1). The  $\alpha$  in the elastic-net penalty can be selected with the EBIC or cross validation, similarly to how  $\lambda$  is selected. Again similarly to `mgm()`, the `weights` argument allows to weight observations, `binarySign` allows signs for interactions involving binary variables, `threshold` defines the type of thresholding (see Section 2.2.3) and `overparameterize` allows to choose the preferred type of parameterization of interactions involving categorical variables. For additional input arguments see `?mvar`.

In many situations, one fits a VAR model to data that do *not* consist of a sequence of measurements that are equally spaced in time. The reason for this can be (randomly) missing measurements and gaps implied by the measurement process: for instance, in an experience sampling study, individuals may be asked to respond to questions about symptoms 6 times a day at equal time intervals of three hours. A mixed VAR model would then show how the presence of a symptom at a given time point is related to the presence of that and other symptoms at earlier time points (3h ago, 6h ago, etc.). However, because the individual sleeps at night, there are gaps in the time series. If one did not take this information into account, every seventh data point in the time series would represent a lag with the length of the night-gap, whereas the other six are representing a lag of three hours. This problem can be avoided by providing an integer sequence via the argument `consec`, which indicates whether measurements are consecutive. For instance if one has a time series with 12 time points (2 days of measurements in the above example), one would provide the vector `c(1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6)`. If one specifies a lag of order 1, `mvar()` then excludes the time step (over night) from measurement 6 to 7. Alternatively one can specify the notification number and the day number via the arguments `beepvar` and `dayvar`, respectively. Then the `consec` variable is computed internally. If a larger number of lags is included, more measurements are excluded accordingly. Information about which cases were excluded as well as the final data matrix used for estimation can be found in `mvar$call`. For more details see `?mvar` and the application example for the time-varying mVAR model below.

### 2.3.2.2 Making Predictions from mixed VAR model

Here we show how to use the `predict()` function to compute predictions and nodewise errors from the model estimated in the previous section. We provide the fit object `mvar_fit` and the data as arguments:

```
pred_mgm <- predict(object = fit_mvar,
                    data = mvar_data$data,
                    errorCon = c("RMSE", "R2"),
                    errorCat = c("CC", "nCC"))
```

`pred_mgm$call` contains the function call, `pred_mgm$predicted` the predicted values for each row in the provided data matrix, and `pred_mgm$probabilities` contains the predicted probabilities for categorical variables. `pred_mgm$errors` contains a table of nodewise errors. Similarly to Section 2.3.1 we specified the Root Mean Squared Error (RMSE) for continuous variables and the (normalized) accuracy for categorical variables:

```
pred_mgm$errors
> Variable  RMSE    R2    CC    nCC
> 1         1    NA    NA 0.754 0.495
> 2         2    NA    NA 0.523 0.000
> 3         3    NA    NA 0.302 0.000
> 4         4    NA    NA 0.266 0.000
> 5         5 0.916 0.157    NA    NA
> 6         6 0.998 0.000    NA    NA
```

Node 1 has the highest normalized accuracy, which makes sense because it is predicted by three other nodes at the previous time point. Nodes 2, 3 and 4 have a normalized accuracy of 0, because they are not predicted by any other node. Node 6 has a proportion of explained variance of 0, because it is not predicted by any other node, and node 5 has a nonzero proportion of explained variance because it is predicted by node 6.

One can also provide customary error functions via the `errorCon` and `errorCat` arguments. For details, see `?predict.mgm`.

### 2.3.2.3 Visualizing mixed VAR model

We visualize the lagged interaction parameters of the mixed VAR model estimated above together with the nodewise errors computed in the previous section. Specifically, we visualize the proportion of explained variance for the two continuous variables, and the normalized accuracy for the four categorical variables:

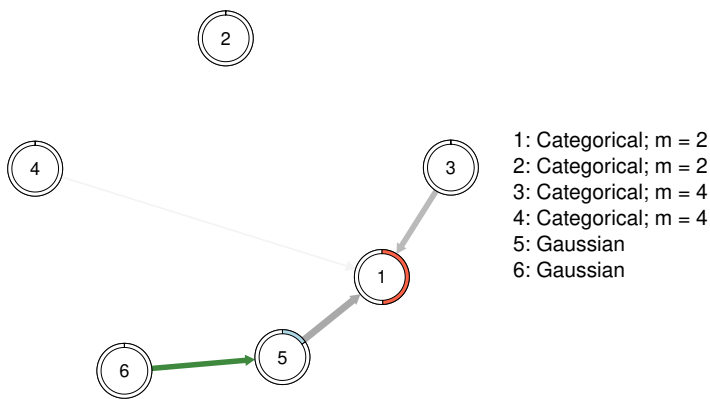
```
errors <- c(pred_mgm$errors[1:4, 5], pred_mgm$errors[5:6, 3])
```

```

qgraph(t(fit_mvar$wadj[, , 1]),
       edge.color = t(fit_mvar$edgecolor[, , 1]),
       pie = errors,
       pieColor = c(rep("tomato", 4), rep("lightblue", 2)),
       nodeName = c(paste0("Categorical; m=", c(2,2,4,4)),
                    rep("Gaussian", 2)))

```

We transposed the parameter matrix `fit_mvar$wadj[, , 1]` because `qgraph()` draws arrows from rows to columns instead of columns to rows, the latter of which is the data structure used in `mvar()`. The resulting plot is shown in Figure 2.7.



**Figure 2.7:** We visualize the lagged interaction parameters of the mixed VAR model estimated above together with the nodewise errors computed in the previous section. Green edges indicate positive relationships. Grey edges indicate that no sign is defined for the pairwise interaction (in the case the interaction involves categorical variables). The width of the edges is proportional to the absolute value of the edge-parameter.

The green edge indicates a positive linear relationship for the cross-lagged effect from node 6 on node 5. The remaining edges are grey, indicating that no sign is defined. This is because these interactions are defined by several parameters, so no sign can be defined. The width of the edges is proportional to the absolute value of the estimated edge-weights in (the values in `fit_mvar$wadj[, , 1]`).

### 2.3.2.4 Sampling from mixed VAR model

We now use the function `mvarsampler()` to sample the data set `mvar_data` used in the previous section. We specify a model with only one lag of order one and  $p = 6$  variables, four categorical (with 2, 2, 4 and 4 categories) and two Gaussians:



```

p <- 6
type <- c("c", "c", "c", "c", "g", "g")
level <- c(2, 2, 4, 4, 1, 1)
max_level <- max(level)
lags <- 1
n_lags <- length(lags)

```

Next, we specify the thresholds for each variable. We assign one threshold (intercept) to the Gaussians, and a separate threshold for each of the categories of each of the categorical variables. These thresholds correspond to the first summation in the joint MGM density in Equation 2.5. In addition, we define a vector indicating the standard deviations of the Gaussian nodes. Note that entries of that vector that do not correspond to Gaussian variables in `type` are ignored.

```

thresholds <- list()
for(i in 1:p) thresholds[[i]] <- rep(0, level[i])
sds <- rep(1, p)

```

Finally, we specify the lagged effects in a  $5 \times p \times \max\{\text{levels}\} \times \max\{\text{levels}\} \times |L|$  array, where  $|L|$  is the number of lags `n_lags`. We first specify the lagged effect from the continuous variable 6 on the continuous variable 5, which consists of a single parameter:

```

# Create coefficient array
coefarray <- array(0, dim=c(p, p, max_level, max_level, n_lags))
# Lagged effect: 5 <- 6
coefarray[5, 6, 1, 1, 1] <- .4

```

We specify two additional lagged effects: one from the categorical variable 3 on the categorical variable 1, which is parameterized by  $2 \times 4$  parameters; and one from the continuous variable 5 to the binary variable 1, which is parameterized by  $2 \times 1$  parameters.

```

# Lagged effect 1 <- 5
coefarray[1, 5, 1:level[1], 1:level[5], 1] <- c(0, 1)
# Lagged effect 1 <- 3
m1 <- matrix(0, nrow=level[2], ncol=level[4])
m1[1,1:2] <- 1
m1[2,3:4] <- 1
coefarray[1, 3, 1:level[2], 1:level[4], 1] <- m1

```

Finally, all arguments are provided to `mvarsampler()`:

```
mvar_data <- mvarsampler(coefarray = coefarray,
                        lags = lags,
                        thresholds = thresholds,
                        sds = sds,
                        type = type,
                        level = level,
                        N = 200,
                        pbar = TRUE)
```

These sampled data correspond to the example dataset in `mvar_data` we used above to illustrate how to estimate a mVAR model.

### 2.3.2.5 Application: Resting state fMRI data

We fit an mVAR model with lags 1, 2 and 3 to resting state fMRI data of a single participant. The dataset consists of BOLD measurements of 68 voxels for 240 time points, where the average sampling frequency is 2 seconds (for details see Schmittmann, Jahfari, Borsboom, Savi, & Waldorp, 2015). The data is loaded automatically with the *mgm* package. All BOLD measurements are modeled as conditional Gaussians, and accordingly we specify the number of levels to be equal to 1 for all variables. We select the regularization parameters  $\lambda$  with the EBIC with tuning parameter  $\gamma = 0.25$ , and we include lags of order 1, 2 and 3.

```
rs_mvar <- mvar(data = restingstate_data$data,
               type = rep("g", 68),
               level = rep(1, 68),
               lambdaSel = "EBIC",
               lambdaGam = 0.25,
               lags = c(1, 2, 3))
```

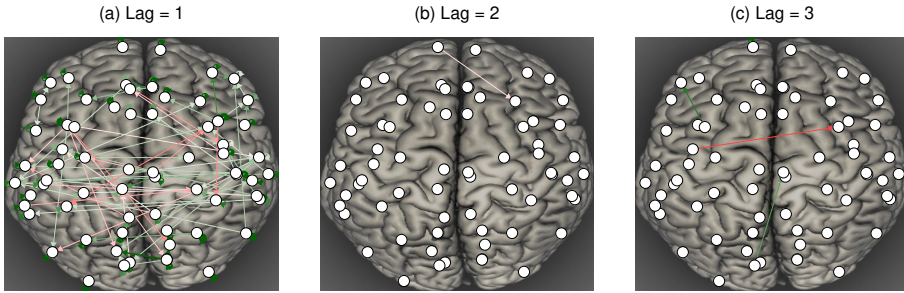
We visualize the  $68 \times 68 \times 3$  interaction parameters of this VAR model in `rs_mvar$wadj` in three separate network plots in Figure 2.8, one for each lag. We provide code to reproduce Figure 2.8 from the package example data set `restingstate_data` in the online supplementary materials and on the Github repository <https://github.com/jmbh/mgmDocumentation>.

For the lag of size one, many coefficients are nonzero. In contrast, for the lags of size two and three only few coefficients are nonzero. For a typical fMRI data analysis, this could mean that it is sufficient to fit a VAR model of lag 1 in order to reduce the variance for further analyses.

Similarly to MGMs, the function `resample()` can be used to obtain bootstrapped sampling distributions for the parameters of the mVAR model.

### 2.3.3 Time-varying Mixed Graphical Model

In this section we show how to estimate a time-varying MGM, how to compute predictions from it and how to visualize it.



**Figure 2.8:** Visualization of the fitted mVAR model, where we depict the parameters separately for each lag. Red edges indicate positive relationships, green edges indicate negative relationships. The width of the edges is proportional to the absolute value of the edge-parameter.

### 2.3.3.1 Estimating time-varying Mixed Graphical Model

We fit a time-varying MGM to gene expression data used by Gibberd and Nelson (2017), who took a subset of the data presented by Arbeitman et al. (2002). Specifically, we model  $p = 150$  gene expressions related to the immune system of *D. melanogaster* (the fruit fly) measured at  $n = 67$  time points across its whole life span. Since  $p > n$ , this is an example of a high-dimensional estimation problem. Figure 2.9 (top panel) shows that the 67 measurements are distributed unequally across the time interval.

Estimating the type of time-varying models introduced in Section 2.2.5 requires the specification of a bandwidth parameter  $\sigma$  that reflects how many time points are combined locally for estimation. The bandwidth parameter  $\sigma$  is the standard deviation of the Gaussian distribution that defines the weighting function (see Section 2.2.5). The empirical time points are normalized to the interval  $[0, 1]$  and the Gaussian weighting function is defined on this interval. This allows some intuition about which  $\sigma$  is appropriate. For example,  $\sigma = 2$  implies weights that are close to uniform on the interval  $[0, 1]$  and therefore gives similar results as the stationary model. This intuition allows to specify a *candidate*  $\sigma$ -sequence. We select  $\sigma$  with the function `bwSelect()`, which computes prediction errors on leave-out sets for all candidate  $\sigma$ -values and selects the  $\sigma^*$  that has the lowest mean error. In the next paragraph we describe this approach in detail.

The function `bwSelect()` fits time varying models on an equally spaced sequence between the time points  $j$  and  $n - F + j - 1$  of length  $J$  with  $j \in \{1, 2, \dots, F\}$ , where  $F$  is the number of folds (times the procedure is repeated) while leaving out (weighting to zero) the time point at which the model is estimated. In a second step, the data at this time point are predicted with the time-varying model and an error measure is computed (RMSE for non-categorical, 0/1-loss for categorical). This procedure is repeated  $F$  times. Then we take the arithmetic mean over  $J$  estimation points,  $p$  variables and  $F$  folds. If  $J = \frac{n}{F}$ , this procedure is equal to a time-stratified  $F$ -fold cross-validation scheme. We allow to specify  $J < \frac{n}{F}$  to save computational cost.  $J$  is specified by the argument `bwFoldsSize` and  $F$  is specified by the argument `bwFolds`. Selecting the ratio between `bwFoldsSize` and

$n$  corresponds to the problem of selecting the number of folds in cross-validation (see e.g., J. Friedman, Hastie, & Tibshirani, 2001).

For the present illustration we select `bwFolds = 5` and `bwFoldsize = 5` to keep the computation time short. We provide the candidate  $\sigma$ -sequence `{0.1,0.2,0.3,0.4}`. And we provide all arguments for the time-varying MGM. This is because we repeatedly fit the type of model we want as our final model (then with fixed  $\sigma$ ). We provide the time points of measurements `fruitfly_data$timevector` via the argument `timepoints` (see Figure 2.2 in Section 2.2.5 for an explanation of why one has to provide the time points if they are not equally spaced). Finally, we specify the class of time-varying model `modeltype = "mgm"` and pass the arguments `k`, `threshold` and `ruleReg`, to `tvmgm()` (see Section 2.3.1 on `mgm()` for a description of these arguments).

```
set.seed(1)
p <- ncol(fruitfly_data$data)
bw_tvmgm <- bwSelect(data = fruitfly_data$data, # Takes around 3h
  type = rep("g", p),
  level = rep(1, p),
  bwSeq = c(0.1, 0.2, 0.3, 0.4),
  bwFolds = 5,
  bwFoldsize = 5,
  timepoints = fruitfly_data$timevector,
  modeltype = "mgm", k = 2,
  threshold = "none", ruleReg = "OR")
```

We would like to know which candidate bandwidth minimized the average prediction error. This information is stored in `bw_tvmgm$meanError`:

```
round(bw_tvmgm$meanError, 3)
> 0.1 0.2 0.3 0.4
> 0.826 0.707 0.630 0.640
```

We see that  $\sigma = 0.3$  minimizes the error in this dataset. If the smallest/largest candidate  $\sigma$  minimized the prediction error, it is advisable to extend the candidate  $\sigma$  sequence to smaller/larger values.

After obtaining a reasonable bandwidth for this data set, we can estimate the final time-varying MGM. The estimation points are specified on the unit interval  $[0,1]$  to which the provided time scale is normalized internally. We choose 20 equally spaced time points across the time series by setting `estpoints = seq(0, 1, length = 20)`. Finally, we specify the above obtained bandwidth with `bandwidth = 0.3` and set a random seed to ensure reproducibility.

```
set.seed(1)
fit_tvmgm <- tvmgm(data = fruitfly_data$data,
  type = rep("g", p),
  level = rep(1, p),
  timepoints = fruitfly_data$timevector,
  estpoints = seq(0, 1, length = 20),
```

```

k = 2,
bandwidth = 0.3,
threshold = "none",
ruleReg = "OR")

```

The output list in the fit object `fit_tvMGM` is similar to the list returned by `mgm()`. The difference is that all parameter matrices are now 3 dimensional arrays, with an additional dimension for the estimated time points  $t^e \in \mathcal{E}$ . For instance, the edge parameters of the pairwise MGM estimated at the third estimation point  $t^e = 3$  are stored in the matrix `fit_tvMGM$pairwise$wadj[, , 3]`. For a detailed description of all output provided in `fit_tvMGM`, see the help file `?tvMGM`.

### 2.3.3.2 Making Predictions from time-varying Mixed Graphical Model

When making predictions with time-varying MGMs, in principle we would need to estimate the time-varying model at the maximum resolution, that is, at every time point. However, this would be computationally expensive: for example, for a time-series of  $n = 1000$  time points, we would need to fit 1000 models in order to compute predictions. The `predict` method in `mgm` provides two different options in order to compute predictions and nodewise errors across time, without requiring to estimate  $n$  models.

The first option, `tvMethod = "weighted"`, computes predictions for each of the  $n$  time points from *all* models  $t^e \in \mathcal{E}$ . It then computes a weighted average over the predictions of all models at each time point. The weight is equal to the weight of the kernel function at  $t$  for the respective model estimated at  $t^e$ . The second option is `tvMethod = "closestModel"`, which for each time point determines the closest estimation point  $t^e$ , and then uses this model for prediction. Accordingly, local nodewise errors are calculated only from the closest model. Note that if one estimates  $n$  models at equally spaced time points, this method corresponds to the above described situation of estimating a time-varying model for each time point.

In order to compute predictions we call the `predict()` function and provide the data, the fit object and the desired method to compute predictions. Here we pick `tvMethod = "weighted"`:

```

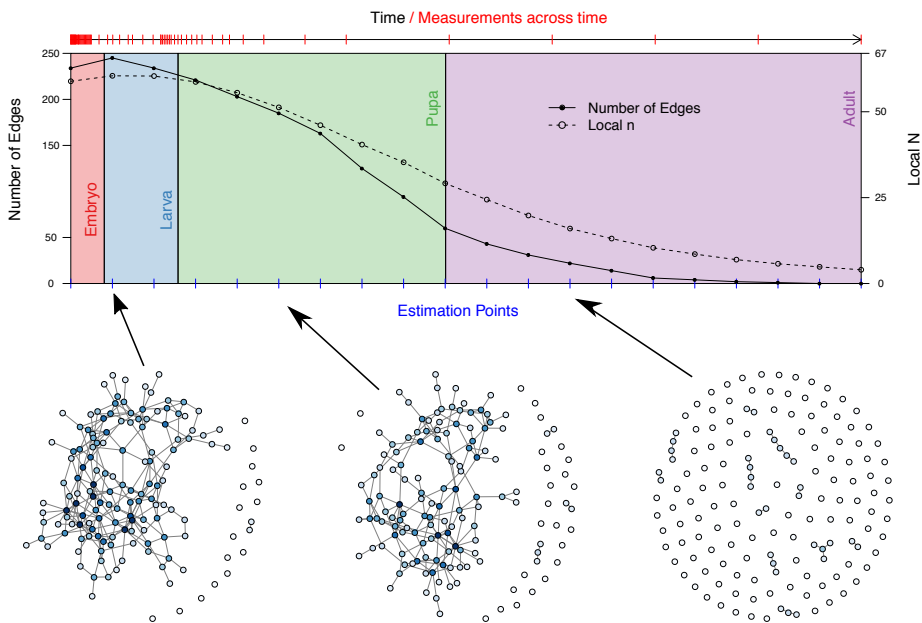
pred_tvMGM <- predict(object = fit_tvMGM,
                     data = fruitfly_data$data,
                     tvMethod = "weighted")

```

The output object `pred_tvMGM` is a list containing the function call `pred_tvMGM$call`, the predicted values `pred_tvMGM$predicted` and `pred_tvMGM$probabilities` (in the case of categorical variables) computed by the method `tvMethod = "weighted"`. `pred_tvMGM$true` contains the true data matrix and `pred_tvMGM$errors` contains an array of local nodewise error, where the third dimension indicates the estimation points.

### 2.3.3.3 Visualizing time-varying Mixed Graphical Model

Figure 2.9 displays several aspects of the time-varying MGM estimated on the fruit-fly data above. The top panel shows the number of edges (solid line) estimated across the time series of 67 measurements, which decreases across the time series. This can be explained by the small number of measurements available at the end of the time series (see red dashes on the time arrow). To make this explicit, we plot  $n_{\sigma=0.3,t^e}$ , the used sample size at a given estimation point (see Section 2.2.5). We see that extremely few data points are available in the end of the time series, resulting in a very low sensitivity to detect edges. The lower panel shows the undirected network at the 2nd, 6th and 13th estimation point out of 20 equally spaced estimation points across the whole time series (blue dashes).



**Figure 2.9:** Top panel: the number of estimated edges (solid line) and the local sample size  $n_{\sigma=0.3,t^e}$  (dashed line) at each estimation point. The red dashes indicate the available measurement on the true time scale. The four colored areas indicate the four stages of the life cycle of the fruit fly. Bottom panel: the undirected network plotted at three different estimation points 2, 6, 13 (with 20 estimation points equally distributed across the 67 time points).

While we can interpret the MGM at each estimation in context of the local  $n_{\sigma,t^e}$ , it is difficult to interpret changes over time, because the sensitivity of the algorithm decreases towards the end of the time series (because less data is available) and hence it is unclear whether edges in the end of the time series are absent in the true model or whether the sensitivity of the algorithm was too low to detect them. This highlights the importance of collecting data with a roughly constant sampling frequency. We provide code to exactly reproduce Figure 2.9 from the

package example data set `fruitfly_data` in the online supplementary materials and on the Github repository <https://github.com/jmbh/mgmDocumentation>.

#### 2.3.3.4 Sampling from time-varying Mixed Graphical Model

The function `tvmgmsampler()` allows to sample from a time-varying MGM. The function input is identical to the input to `mgmsampler()`, the sampling function of the stationary MGM described in Section 2.3.1.5, except that the arguments `thresholds`, `sds` and `interactions` have an additional dimension for time. The number of indices in this additional time dimension defines the length of the time series. Thus, a separate model is specified for each time point in the time series. For details see `?tvmgmsampler`.

#### 2.3.3.5 Bootstrap Sampling Distributions

Similarly to stationary MGMs, the function `resample()` can be used to obtain bootstrapped sampling distributions for the parameters of the time-varying MGM. The only difference is that we use a block-bootstrapping scheme to ensure that data points remain reasonably distributed across time. The number of blocks can be specified with the argument `blocks` in the `resample()` function. The larger the number of blocks, the more evenly distributed the bootstrap samples are across the time interval and the higher the similarity between bootstrap samples. Since even distribution across time and low similarity across bootstrap samples is desirable, the number of blocks controls this trade-off. For more details see the help file `?resample`.

### 2.3.4 Time-varying mixed VAR model

We illustrate how to fit a time-varying mixed VAR model on a symptom time series with 51 variables measured on 1478 time points during 238 consecutive days from an individual diagnosed with major depression (Wichers, Groot, Psychosystems, Group, et al., 2016). The measured variables include questions regarding mood, self-esteem, social interaction, physical activity, events and symptoms of depression (see also legend in Figure 2.10). During the measured time interval, a double-blind medication dose reduction was carried out, consisting of a baseline period, the dose reduction, and two post assessment periods (See Figure 2.10, the points on the time line correspond to the two dose reductions). For a detailed description of this data set see Kossakowski, Groot, Haslbeck, Borsboom, and Wichers (2017).

#### 2.3.4.1 Estimating time-varying mixed VAR model

We provide the data, the type (continuous and categorical), and the levels for each variable, all of which are contained in the data list `symptom_data` (automatically loaded with `mgm`), similarly to specifying `mvar()`. Next, we provide the day number with `dayvar` and the number of notification on each day with `beepvar`.

Alternatively, one could manually compute a single vector that indicates the consecutiveness of measurements and provide it via the argument `consec`. We provide this information because the measurements in this data set are not consecutive, both because of the day-night break in which no measurements are taken and because of randomly missing measurement points. If we did not provide this information, the resulting parameters represent a mixture of effects across different lags and are therefore not interpretable anymore. We explained this in detail in Section 2.2.5. The function `tvmvar()` uses this information to fit the model only on rows of the time series for which sufficient previous measurements are available (1 for lag 1, 2 for lag 2, etc.).

In order to fit a time-varying MGM we need to choose an appropriate bandwidth parameter  $\sigma$ , which determines how many observations close in time we combine in order to estimate a local model (see Section 2.2.5). In Section 2.3.3, we provided an explanation of how to use `bwSelect()` to select an appropriate  $\sigma$  using a time-stratified cross validation scheme. Here we choose  $\sigma = 0.2$ .

We specify a lag of order 1 and via the argument `estpoints` we specify that we would like to estimate the model at 20 equally spaced time intervals throughout the time series. We specify the sequence of estimation points on the unit interval  $[0, 1]$ , to which the provided time scale is normalized internally. Finally, we set thresholding `threshold = "none"` and set a random seed to ensure reproducibility.

```
set.seed(1)
fit_tvmvar <- tvnvar(data = symptom_data$data, # Takes around 15min
                   type = symptom_data$type,
                   level = symptom_data$level,
                   beepvar = symptom_data$data_time$beepno,
                   dayvar = symptom_data$data_time$dayno,
                   lags = 1,
                   estpoints = seq(0, 1, length = 20),
                   bandwidth = 0.2,
                   threshold = "none",
                   saveData = TRUE)
```

The output of `tvmvar()` is similar to the output of `?mvar` described in Section 2.3.2. The difference is that all entries have now an additional dimension for estimation points. For example, the entry of the parameter array `fit_tvmvar$wadj[4, 9, 2, 15]` indicates the cross lagged effect of 9 on 4 over the second specified lag in `lags` at the 15th estimation point. The array `fit_tvmvar$signs` has the same dimension and specifies the signs of the parameters in `fit_tvmvar$wadj`, if defined. For a discussion of when a sign is defined for an edge-parameter see Section 2.3.1. The object `fit_tvmvar$intercepts` contains a list with time-varying thresholds/intercepts and `fit_tvmvar$tvmodels` contains the models at each of the  $|\mathcal{E}|$  estimation points.

We provided the day and notification number of each measurement and `tvmvar()` used this information to only include measurements in the model for which sufficient previous measurements are available. By executing the model



object in the console, we get the number of measurements that were actually used for estimation:

```
fit_tvmvar
> mgm fit-object
> Model class: Time-varying mixed Vector Autoregressive (tv-mVAR) model
> Lags: 1
> Rows included in VAR design matrix: 876 / 1476 ( 59.35 %)
> Nodes: 48
> Estimation points: 20
```

We see that for 876 of 1476 measurement points the previous measurement (requirement of lag 1) is available and were therefore used for estimation. If we included lags with higher order the number of usable measurements would become smaller.

### 2.3.4.2 Making Predictions from time-varying mixed VAR model

In order to compute predictions from the mixed VAR model we have to choose between the two options `tvMethod = "weighted"` and `tvMethod = "closestModel"`. For a discussion of these two methods see Section 2.3.3 or the help file `?predict.mgm`. Next to the fit object we provide the data and information about the consecutiveness of measurements to `predict()`:

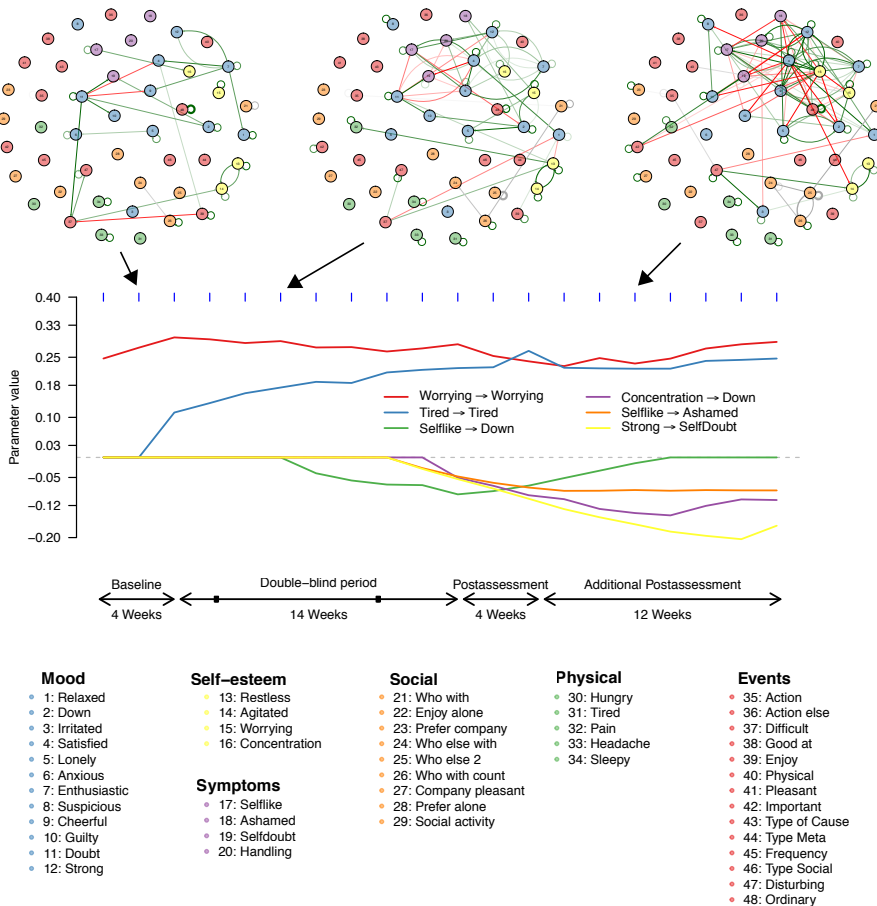
```
# Compute Predictions
pred_tvmvar <- predict(object = fit_tvmvar,
                      data = symptom_data$data,
                      tvMethod = "weighted",
                      beepvar = symptom_data$data_time$beepno,
                      dayvar = symptom_data$data_time$dayno)
```

The output object `pred_tvmvar` is a list containing the function call `pred_tvmgm$call`, the predicted values `pred_tvmgm$predicted` and `pred_tvmgm$probabilities` (in the case of categorical variables) computed by the method `tvMethod = "weighted"`. `pred_tvmgm$true` contains the true data matrix, which is useful in the case of VAR models, when not all rows in the original data matrix are necessarily used to fit the VAR model (see previous section). Finally, `pred_tvmgm$errors` is an array of local nodewise errors, where the third dimension indexes estimation points. For instance, `pred_tvmgm$errors[, , 9]` contains the nodewise errors for estimation point 9.

### 2.3.4.3 Visualizing time-varying mixed VAR model

Figure 2.10 displays some aspects of the time varying mixed VAR model estimated in the previous section.

## 2. Estimating Mixed Graphical Models



**Figure 2.10:** Top row: network visualization of VAR(1) parameters at the estimation points 2, 6, and 16. Green edges indicate positive relationships, red edges indicate negative relationships and grey edges indicate that no sign is defined. The color of the nodes corresponds to the group the variable belongs to (see legend); second row: six autoregressive (e.g.,  $Worrying^{t-1} \rightarrow Worrying^t$ ) or cross-lagged effects (e.g.,  $Selflike^{t-1} \rightarrow Down^t$ ) depicted as a function of time.

In the top row of Figure 2.10 we depict a network plot of the VAR(1) parameters at the estimation points 2, 6, and 16. Green edges indicate positive relationships and red edges indicate negative relationships. Grey edges indicate that no sign is defined, because the edge-weight is a function of several parameters, which is the case for interactions including categorical variables (see Section 2.3.1). The width of edges is proportional to the absolute value of the edge-weight. It is evident from the three network plots that the model changes considerably over time which suggests that a stationary model is not appropriate for these data. The second row depicts six autoregressive or cross-lagged effects across the measured time interval. We see that parameters change considerably

over time, for instance the autoregressive effect of *Tired* is strong at the beginning of the time series and decreases almost monotonously until the end of the measured time interval.

We provide code to exactly reproduce Figure 2.10 from the example data set `symptom_data` in the online supplementary materials and on the Github repository <https://github.com/jmbh/mgmDocumentation>.

#### 2.3.4.4 Sampling from time-varying mixed VAR model

The function `tvmvarsampler()` allows to sample from a time-varying mVAR model. The function input is identical to the input to `mgmsampler()`, the sampling function of the stationary mVAR described in Section 2.3.2, except that the arguments `thresholds`, `sds` and `coefarray` have an additional dimension for time. The number of indices in this additional time dimension defines the length of the time series. Thus, a separate model is specified for each time point in the time series. For details see `?tvmvarsampler`.

Similarly to time-varying MGMs, the function `resample()` allows to obtain bootstrapped sampling distributions for the parameters of time-varying mixed VAR models.

## 2.4 Concluding Comments

We presented the R-package `mgm` which allows to fit stationary and time-varying Mixed Graphical Models and stationary and time-varying mixed Vector Autoregressive Models. In addition to the estimation functions, we provide methods to compute predictions and nodewise errors and assess the stability of estimates via resampling. Furthermore, flexible sampling functions for all model classes allow the user to evaluate the performance of the estimation algorithms in a given situation via simulations. Finally, we provided fully reproducible code examples that illustrate how to use the software package.

The `mgm` package is under continuous development. We aim to add functions that allow one to inspect higher order interactions in an accessible way. We plan to implement different ways to select tuning parameters ( $\lambda$  penalization parameter,  $\alpha$  elastic net parameter,  $\sigma$  bandwidth parameter), for instance with stability-selection (Meinshausen & Bühlmann, 2010; Liu, Roeder, & Wasserman, 2010). And we will implement other estimators than  $\ell_\alpha$ -penalized regression, which might be more appropriate in some situations. Finally, since all estimation algorithms are based on sequential regressions, considerable performance gains can be made by parallelizing the estimation algorithms.

## Acknowledgements

We would like to thank three anonymous reviewers for their detailed and constructive feedback, and Fabian Dablander and Oisín Ryan for their comments on

an earlier version of this manuscript.

# NODEWISE PREDICTABILITY

---

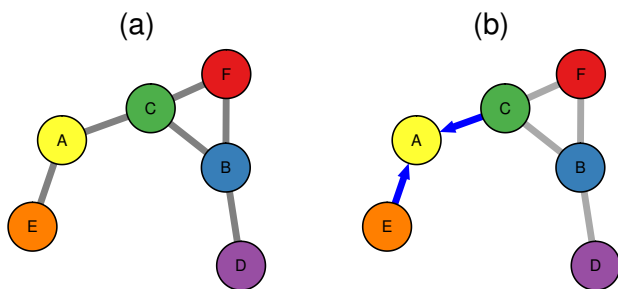
## Abstract

Network models are an increasingly popular way to abstract complex psychological phenomena. While the study of the *structure* of network models has led to many important insights, little attention is paid to how well they *predict* observations. This is despite the fact that predictability is crucial for judging the *practical relevance* of edges: for instance in clinical practice, predictability of a symptom indicates whether a an intervention on that symptom through the symptom network is promising. We close this methodological gap by introducing nodewise predictability, which quantifies how well a given node can be predicted by all other nodes it is connected to in the network. In addition, we provide fully reproducible code examples of how to compute and visualize nodewise predictability both for cross-sectional and time-series data.

### 3.1 Introduction

Network models graphically describe interactions between a potentially large number variables: each variable is represented as a dot (node) and interactions are represented by lines (edges) connecting the nodes (for an illustration see panel (a) of Figure 3.1). These models have been a popular way to abstract complex systems in a large variety of disciplines such as statistical mechanics (Albert & Barabasi, 2002), biology (N. Friedman et al., 2000), neuroscience (Huang et al., 2010) and are recently also applied in psychology (Costantini et al., 2015) and psychiatry (Borsboom, 2017).

Particularly in psychology, network models are attractive because many psychological phenomena are considered to depend on a large number of variables and interactions between them. In such a situation, the graphical representation ensures that the model can be understood intuitively even if the number of variables is large. In addition, network models open up the possibility to study the network structure: for instance, one can use network summary measures like density or centrality to describe the global structure of the network (M. Newman, 2010). These could allow inferences about the behavior of the whole network that would not be possible from the edge parameters alone. One could also run generative models on the network, e.g. diffusion models of diseases to explain how symptoms of psychological disorders activate each other (Shulgin, Stone, & Agur, 1998).



**Figure 3.1:** (a) Example network with six nodes. An edge between two nodes indicates a pairwise interaction between those two nodes; (b) Illustration of predicting node A by all its neighboring nodes (E and C).

Currently, most applications are in the field of clinical psychology (e.g., Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016; Fried et al., 2015; Beard et al., 2016; McNally et al., 2015; Boschloo et al., 2015) but network models are also applied in other subfields such as health psychology (Kossakowski et al., 2016) and personality psychology (Cramer et al., 2012; Costantini et al., 2015). While initially they were used to model cross-sectional data, there is increasing interest in analyzing data obtained using the Experience Sampling Method (ESM),

which consists of repeated measurements of the same person (e.g., Bringmann et al., 2013; Pe et al., 2015). The focus in these papers is the global network structure and how specific nodes are connected in the network, which provides a new perspective on many psychological phenomena. For instance, Cramer and colleagues (Cramer, Waldorp, Van Der Maas, & Borsboom, 2010) suggested an alternative view on the concept of comorbidity by analyzing how symptoms of different psychological disorders are connected to each other.

The key idea of this chapter is to analyze the *predictability* of nodes in the network *in addition* to the network structure. By predictability of node A we mean how well node A can be predicted by all its neighboring nodes in the network (see Figure 1b). The predictability of nodes is important for several reasons:

1. The edges connected to node A should be interpreted taking into account how much of the variance of A is explained by the nodes connected to A. For instance, edges will be interpreted differently, depending on whether 0.5% or 50% of the variance of A is explained. This issue is particularly important for networks estimated on a large number of observations, where small edge weights can be detected that are likely to be meaningless in practice.
2. In many areas of psychology the goal is to design effective interventions. Using the predictability measure of node A, one can estimate how much we can influence this node by intervening on nodes that are connected to it.
3. Predictability across nodes tells us whether a (part of a) network is largely determined by itself through strong mutual interactions between nodes (high predictability) or whether it is mostly determined by other factors that are not included in the network (low predictability).

The problem addressed here is similar to the problem of modeling only the covariance matrix in Structural Equation Modeling (SEM) (Byrne, 2013): one might find a model that perfectly fits the covariance matrix, but if the variance of variables is much larger than their covariance, the model might be of limited use in practice.

Predictability in general cannot be inferred by the network structure but has to be computed from the network model and the data. Unfortunately, currently there is no easy-to-use tool available for researchers to compute predictability and include it in network visualizations. In the present chapter, we address this methodological gap by making the following contributions:

1. We present a method to compute easy to interpret nodewise predictability measures for state-of-the-art network models (Section 3.2).
2. We provide a step-by-step description of how to use the R-packages *mgm* and *qgraph* to compute and visualize nodewise predictability, both for cross-sectional (Section 3.3) and time-series networks (Section 3.4). The provided code is fully reproducible, which means that the reader can run the code and reproduce all figures while reading. The data in our applications are from two published studies and will be downloaded automatically with the provided code.

## 3.2 Methods

In order to determine the predictability of a given node  $A$ , we need to know which nodes are connected to  $A$  in the network model. Therefore the first step is to estimate a network model, which we describe in Section 3.2.1. In a second step, we use the network model to predict the given node  $A$  by the nodes that are connected to it (its neighbors). In Section 3.2.2, we describe in detail how to compute these predictions. Finally, we quantify how close these predictions are to the actual values of  $A$ . The closer the predictions are to the actual values, the higher the predictability of  $A$ . A description of predictability measures for both continuous and categorical variables is given in Section 3.2.3. In Section 3.2.4 we discuss the relationship between the predictability and the parameters of the network model. Finally we describe the data (3.2.5) that is used in the application examples in Sections 3.3 and 3.4.

### 3.2.1 Network Models

We model cross-sectional data using pairwise Mixed Graphical Models (MGMs) (Yang, Baker, Ravikumar, Allen, & Liu, 2014b; Haslbeck & Waldorp, 2015), which generalize well-known exponential family distributions such as the multivariate Gaussian distribution or the Ising model (Wainwright & Jordan, 2008). This is the model used in all papers mentioned in the introduction. MGMs are estimated via  $\ell_1$ -regularized (LASSO) neighborhood regression as implemented in the R-package *mgm* by the authors (Haslbeck & Waldorp, 2020). In this approach, one estimates the neighborhood of each node and combines all neighborhoods to obtain the complete graph (network) (Meinshausen & Bühlmann, 2006). The *neighborhood* of a node is the set of nodes that is connected to that node. For example, in Figure 1(a), the neighborhood of node  $A$  consists of the nodes  $E$  and  $C$ . The  $\ell_1$ -regularization ensures that spurious edge-parameters are put to exactly zero, which makes the network model easier to interpret. The parameter that controls the strength of the regularization is selected via 10-fold cross validation.

To model time-series data we use the Vector Autoregressive (VAR) model, which is a popular model for multivariate time series in many disciplines (see e.g. Hamilton, 1994; Pfaff, 2008a). The VAR model is different from the MGM in that associations are now defined between time-lagged variables. Specifically, in its simplest form with a time-lag of order one, in this model all variables  $X^{t-1}$  at time  $t-1$  are regressed on *each* of the variables  $X_i^t$  at time  $t$ , where  $i$  indexes different variables. Note that this also includes the variable  $X_s$  itself at an earlier time point: that is, one predicts  $X_s^t$  at time  $t$  by itself and all other variables at time  $t-1$ . For the analyses in this chapter we use the implementation of mixed VAR models in the R-package *mgm* (Haslbeck & Waldorp, 2020).

### 3.2.2 Making Predictions

We are interested in how well a node can be predicted by all adjacent nodes in the network. This means that we would like to compute the mean of the conditional



distribution of the node at hand given all its neighbors. To provide understanding of what this means exactly, we show how to compute predictability for the node  $A$  in Figure 3.1 (b), for (1) the case of  $A$  being a continuous-Gaussian variable and (2) the case of  $A$  being binary.

We begin with (1): the conditional mean of  $A$  given its neighbors  $C$  and  $E$ , which is given by

$$P(A = x|C, E) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}, \quad (3.1)$$

where the mean  $\mu = \beta_0 + \beta_C C + \beta_E E$  is a linear combination of the two neighbors  $C$  and  $E$ . This conditional distribution is obtained from the multivariate exponential family distribution of the MGM, for details see Yang et al. (2014b); Haslbeck and Waldorp (2015). This prediction problem corresponds to the familiar linear regression problem with Gaussian noise. Now, how to make predictions? Let's say the intercept is  $\beta_0 = 0.25$  and  $\beta_C = 0.1, \beta_E = -0.5$ . Then, if the  $i^{\text{th}}$  case in the sample is  $C_i = 2, E_i = 1$ , then for the  $i^{\text{th}}$  sample of  $A$  we predict  $A_i = 0.25 + 0.1 \times 2 - 0.5 \times 1 = -0.05$ . A measure of predictability should evaluate how close this is the actual observation for node  $A_i$ .

In example (2), where  $A$  is categorical, we compute a predicted probability for each category using a multinomial distribution

$$P(A = k|C, E) = \frac{\exp\{\mu_k\}}{\sum_{l=1}^K \exp\{\mu_l\}}, \quad (3.2)$$

where  $k$  indicates the category,  $K$  is the number of categories and  $\mu_k = \beta_{0k} + \beta_{Ck} C + \beta_{Ek} E$ . Now let's assume  $A$  is binary ( $K = 2$ ) and we have  $\beta_{01} = 0, \beta_{C1} = 0.5, \beta_{E1} = 1$  and  $\beta_{02} = 0, \beta_{C2} = -0.5, \beta_{E2} = -1$  and if for the  $i^{\text{th}}$  cases we have  $C_i = 1$  and  $E_i = 1$ . When filling in the numbers in equation (3.2) we get  $P(A = 1|C, E) \approx 0.95$  and  $P(A = 2|C, E) \approx 0.05$ , and predict category  $k = 1$  for the  $i^{\text{th}}$  sample of  $A$ , because  $0.95 > \frac{1}{2}$ . Of course, all probabilities have to add up to 1, so we have  $1 - P(A = 1|C, E) = P(A = 2|C, E)$ . This direct approach of modeling the probabilities of categories is possible due to the regularization used in estimation (see e.g. Hastie et al., 2015), otherwise this model would not be identified. Note that predicting  $A$  by all its neighbors is the same as predicting  $A$  by all nodes in the network. This is because all nodes that are *not* in the neighborhood of  $A$  have a zero weight associated to them in the regression equation on  $A$  (3.1 or 3.2) and can hence be dropped.

In the case of other exponential family distributions, such as Poisson or Exponential, one similarly uses the univariate conditional distribution to make predictions (Yang et al., 2014b). Importantly, the joint distribution of the MGM can be represented as a factorization of  $p$  conditional distributions and hence our method to compute predictions is based on a proper representation of the joint distribution. Indeed, this factorization is used when estimating the MGM in the neighborhood regression approach (see Section 3.2.1).

### 3.2.3 Quantifying Predictability

After computing predictions, we would like to know how close these are to the observed values in the data. Because we are interested in how well a given node can be predicted *by all other nodes in the network*, we need to remove any effects of the intercept (continuous variables) and the marginal (categorical variables). The marginal indicates the probabilities of categories, when ignoring all other variables. For example, the marginal of a binary variable is described by relative frequency of observing category 1, e.g.  $P(X = 1) = 0.7$ .

#### 3.2.3.1 Predictability in Continuous Variables

For continuous data, we choose the proportion of explained variance as predictability measure as it is well-known in the literature and easy to interpret:

$$R_A^2 = 1 - \frac{\text{var}(\hat{A} - A)}{\text{var}(A)},$$

where  $\text{var}(X)$  is the variance of  $X$ ,  $\hat{A}$  is a vector of predictions for  $A$  as described in Section 3.2.2, and  $A$  is the vector of observed values in the data. In order to remove any influences of the intercepts, all variables are centered to mean zero. Hence, all intercepts will be zero and cannot affect to the predictability measure. Thus, we can interpret  $R^2$  as follows: a value of 0 means that a node cannot be predicted at all by its neighboring nodes in the network, whereas a value of 1 means that a node can be perfectly predicted by its neighboring nodes.

#### 3.2.3.2 Predictability in Categorical Variables

For categorical variables it is slightly more difficult to get a measure with the same interpretation as the  $R^2$  for continuous variables, because there is no way to center categorical variables. The following example shows that it is, however, important to somehow take the marginal into account: let's say we have 100 observations of a binary variable  $A$  and observe ten 1s and 90 0s. This means that the marginal probabilities of  $A$  are  $p_0 = 0.9$  and  $p_1 = 0.1$ . Now, if all other nodes contribute nothing to predicting whether there is a 0 or 1 present in case  $A_i$ , one can just predict a 0 for all cases and get a proportion of correct classification (or accuracy, see below) of 90%. For our purpose of determining how well a node can be predicted by all other nodes, this is clearly misleading, because actually *nothing* is predicted by all other nodes. We therefore compute a *normalized accuracy* that removes the accuracy that is achieved by the trivial prediction using marginal of the variable ( $p_1 = 0.1$ ) alone:

Let  $\mathcal{A} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i = \hat{y}_i)$  be the proportion of correct predictions, the accuracy, and let  $p_0, p_1, \dots, p_m$  be the marginal probabilities of the categories, where  $\mathbb{I}$  is the indicator function for the event  $F_i = \hat{F}_i$ . In the binary case these are  $p_0$  and  $p_1 = 1 - p_0$ . We then define normalized accuracy as

$$\mathcal{A}_{norm} = \frac{\mathcal{A} - \max\{p_0, p_1, \dots, p_m\}}{1 - \max\{p_0, p_1, \dots, p_m\}}.$$

Hence,  $\mathcal{A}_{norm}$  indicates how much the node at hand can be predicted by all other nodes in the network, *beyond* what is trivially predicted by the marginal distribution.  $\mathcal{A}_{norm} = 0$  means that none of the other nodes adds anything to the marginal in predicting the node at hand, while  $\mathcal{A}_{norm} = 1$  means that all other nodes perfectly predict the node at hand (together with the marginal).

Let's return to the above example: in contrast to the high accuracy of  $\mathcal{A} = 0.9$ , the normalized accuracy  $\mathcal{A}_{norm}$  is zero, indicating that the node at hand cannot be predicted by other nodes in the network. However, notice that both  $\mathcal{A}$  and  $\mathcal{A}_{norm}$  are important for interpretation. For instance if we have a marginal of  $p_1 = .9$  in a binary variable, then it is less impressive if all other predictors account for 80% of the remaining accuracy that can be achieved (.98 instead of .9) than in a situation where  $p_1 = .5$ , where accounting 80% of the remaining accuracy would mean an improvement from .5 to .9. We therefore visualize both  $\mathcal{A}$  and  $\mathcal{A}_{norm}$  for the binary variable in Figure 3.2.

### 3.2.4 Predictability and Model Parameters

Given the above definition of measures of predictability, it is evident that there is a close relationship between the parameters of the network model and predictability: if a node is not connected to any other node then the explained variance/normalized accuracy of this node *has* to be 0. Also, the more edges are connected to a node, the higher predictability tends to be. There is a strong linear relationship between predictability and edge parameters for Gaussian Graphical Models (GGM), where the edge parameters (partial correlation) are restricted to  $[-1, 1]$ . This linear relationship is much weaker for models including categorical variables, where the model parameters are only constrained to be finite.

This implies that also centrality measures (like degree centrality), which are a function of edge parameters, are strongly correlated with predictability for GGMs, but much less for MGMs (e.g., Haslbeck & Fried, 2017). However, note that even if a given centrality measure would correlate perfectly with predictability, it would not be a substitute, because it would only allow us to order nodes by predictability but would *not* tell us the predictability of any node. Hence, while centrality measures are related to predictability, they are not a good proxy for predictability.

### 3.2.5 Application to Datasets

We illustrate how to compute and visualize nodewise predictability for network models for both cross-sectional and time-series data. We use a cross-sectional dataset from (Fried et al., 2015) ( $N = 515$ ) with 11 variables on the relationship on bereavement and depressive symptoms. In order to illustrate predictability for the VAR model, we use a dataset consisting of up to 10 daily measurements of nine variables related to mood over a long period of time ( $N = 1478$ ) of a single individual (Wichers et al., 2016). A detailed description of the time-series data can be found in (Kossakowski et al., 2017).

## 3.3 Predictability in Cross-Sectional Networks

Here we show how to obtain the proposed predictability measures using the *mgm* package. We will provide the code below so all steps can be reproduced exactly by the reader. First, we download the preprocessed data. The raw data and the preprocessing file can be found in the same Github repository.

```
library(httr)
url <- "https://github.com/jmbh/NetworkPrediction/raw/master/Fried2015_nD.RDS"
GET(url, write_disk("Fried2015.RDS", overwrite=TRUE))
datalist <- readRDS("Fried2015.RDS")
```

Next, we fit a MGM using the *mgm*-package:

```
library(mgm)
fit_obj <- mgm(data = datalist$data,
               type = c(rep("g", 11), "c"),
               lev = c(rep(1, 11), 2),
               ruleReg = "OR",
               k = 2, binarySign = TRUE)
```

In addition to the data, one has to specify the type and the number of categories for each variable. The remaining arguments are tuning parameters and are selected such that the original results in Fried et al. (2015) are reproduced. For the general usage of the *mgm* package see Haslbeck and Waldorp (2020). After estimating the model, which is saved in `fit_obj`, we use the `predict()` function to compute the predictability for each node in the network. For categorical variables, we specify the predictability measures accuracy / correct classification ("CC") and normalized accuracy ("nCC"). In addition, we request the accuracy of the intercept (marginal) model ("CCmarg"), which we will use to visualize the accuracy decomposition in intercept model and the contribution of other variables. For continuous variables, we specify explained variance ("R2") as predictability measure.

```
p_obj <- predict(fit_obj, datalist$data,
                errorCat = c("CC", "nCC", "CCmarg"),
                errorCon = c("R2"))
```

To display both the accuracy of the intercept model and the normalized accuracy (contribution by other variables), we require a list for the ring-segments and a list for the corresponding colors:

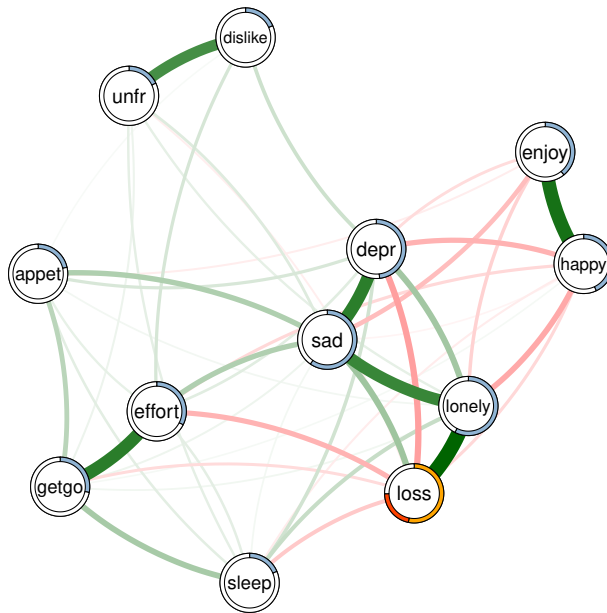
```
error_list <- list() # List for ring-segments
for(i in 1:11) error_list[[i]] <- pred_obj$errors[i, 2]
error_list[[12]] <- c(p_obj$errors[12,5],
                    p_obj$errors[12,3]-p_obj$errors[12,5])

color_list <- list() # List for Colors
for(i in 1:11) color_list[[i]] <- "#90B4D4"
color_list[[12]] <- c("#ffa500", "#ff4300")
```

We now provide the weighted adjacency matrix and the list containing the nodewise predictability measures to `qgraph`, resulting in Figure 3.2:

```
pieColor <- c(rep("#90B4D4", 11), rep("#EB9446", 1)) # pick nice color

library(qgraph)
qgraph(fit_obj$pairwise$wadj, pie = error_list,
       layout="spring", labels = datalist$names,
       pieColor = color_list, label.cex = .9,
       edge.color = fit_obj$pairwise$edgecolor,
       curveAll = TRUE, curveDefault = .6,
       cut = 0, labels = datalist$names)
```



**Figure 3.2:** Mixed Graphical Model estimated on the data from Fried et al. (2015). Green edges indicate positive relationships, red edges indicate negative relationships. The blue ring shows proportion of explained variance (for continuous nodes). For the binary variable "loss", the orange part of the ring indicates the accuracy of the intercept model. The red part of the ring is the *additional* accuracy achieved by all remaining variables. The sum of both is the accuracy of the full model  $\mathcal{A}$ . The normalized accuracy  $\mathcal{A}_{norm}$  is the ratio between the additional accuracy due to the remaining variables (red) and one minus the accuracy of the intercept model (white + red).

The color of the pie chart behind the node can be controlled using the `pieColor` argument. The remaining arguments are not necessary but improve the visualization. The argument `layout="spring"` specifies that the placement of

the nodes in the visualization is determined by the force-directed Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), which places nodes such that all edges have more or less equal length and that there are as few edge crossings as possible. Note that there is no analytic relation between the distance of nodes and model parameters, however, the algorithm tends to group strongly connected nodes together in order to avoid edge crossings. Green and red edges indicate positive and negative relationships, respectively, and the width of the edges is proportional to the absolute value of the edge-weight. For a detailed description of the `qgraph`-package see Epskamp et al. (2012).

This code returns a network that is very similar to the one in the original paper of Fried et al. (2015). Note that the network is not identical as we did not dichotomize ordinal variables but treat them as continuous instead. For the 11 continuous variables, the percentage of explained variance is indicated by the blue part in the pie chart. For the single binary variable, the normalized accuracy is indicated by the orange part in the pie chart.

As expected, nodes with more/stronger edges can be predicted better (e.g. *lonely*) than nodes with fewer/weaker edges (e.g. *unfriendly unfri*). While this trivially follows from the construction of the predictability measure (see Section 3.2.4), this does not mean that one can use the network structure to infer the predictability of a node: by looking at the network visualization in Figure 3.2, we are quite certain that predictability of *lonely* is higher than of *unfri*. However, we do not know *how* high predictability is in either of the two nodes (0.55 and 0.13, respectively), which is highly relevant for interpretation and practical applications.

Because we used the same data for estimating the network and calculating the predictability (or error) measures, we estimated the *within sample prediction error*. In order to see how well the model *generalizes*, one has to calculate the *out of sample prediction error*. This can be done by splitting the data in two parts (or using a cross validation scheme) and providing one part to the estimation function, and the other part to the prediction function.

## 3.4 Predictability in Temporal Networks

Note that the interpretation of predictability is slightly different for VAR networks because we predict each node by *all* nodes at the previous time point, which also includes the predicted node itself. We begin again by downloading the example dataset:

```
url<- "https://github.com/jmbh/NetworkPrediction/raw/master/Wicherts2016_Mood.RDS"
GET(url, write_disk("Wicherts2016_Mood.RDS", overwrite=TRUE))
datalist_ts <- readRDS("Wicherts2016_Mood.RDS")
```

Next, we provide the data and the type and number of categories of variables as input. In addition, we specify that we would like to estimate a VAR model with lag 1

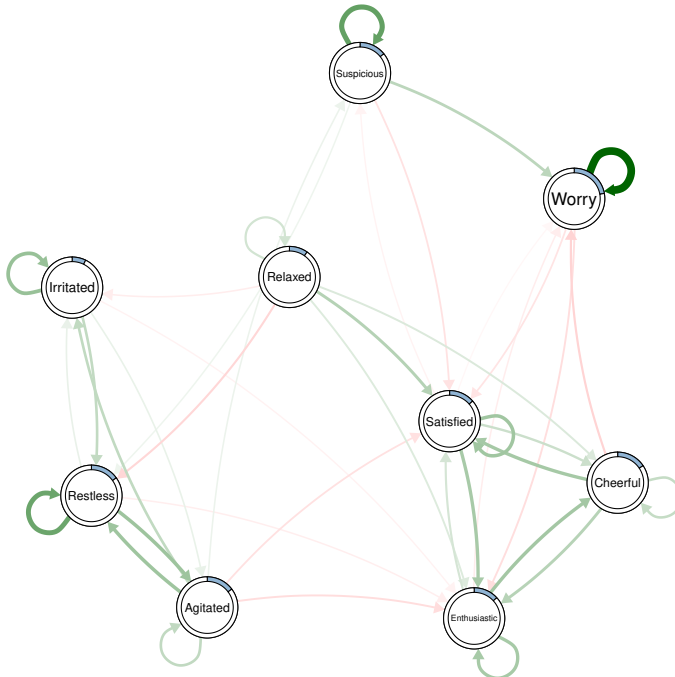
```
var_obj <- mvar(data = datalist_ts$data_mood,
               type = rep("g", 9), lev = rep(1, 9), lags = 1,
               consec = datalist_ts$data_time$beepno)
```

and compute the predictability of each node similarly to above:

```
p_obj2 <- predict(var_obj, datalist_ts$data_mood,
                 errorCon = c("R2"))
```

Finally, we visualize the network structure together with the nodewise predictability measures, which results in Figure 3.3. Because we have only one predictability measure for each node, we can provide them in a vector via the `pie` argument:

```
qgraph(var_obj$wadj[, , 1],
       edge.color = var_obj$edgecolor[, , 1],
       labels = datalist_ts$labels,
       pie = p_obj2$errors[, 2],
       pieColor = rep('#90B4D4', 9),
       curveAll = TRUE, curveDefault = .6, cut = 0)
```



**Figure 3.3:** Visualization of VAR network of the mood variables in Wicherts et al. (2016). Green edges indicate positive relationships, red edges indicate negative relationships. The self-loops refer to the effect of the variable on itself over one time lag. The blue rings around the nodes indicate the proportion of explained variance in that node by all other nodes.

We see two groups of self-engaging mood variables in Figure 3.3: (a) the positive mood variables *Cheerful*, *Enthusiastic* and *Satisfied* and (b) the negative mood variables *Irritated*, *Agitated*, *Restless* and *Suspicious*. *Worrying* seems to be influenced by both groups and *Relaxed* is rather disconnected and has a weak negative influence on group (b). These insights can be used to judge the effectiveness of possible interventions on these mood variables: for instance, if the goal is to change variables in group (a), one can do this by intervening on other variables in (a). In addition, we would expect an effect on *Worrying* when intervening on groups (a) and (b), however, the reverse is not true. *Relaxed* has a small influence on group (b), however, is itself not influenced by any of the variables in the network. Hence, in order to intervene on *Relaxed*, one has to search for additional variables influencing *Relaxed* that were not yet taken into account in the present network.

## 3.5 Discussion

In this chapter we introduced a method and easy-to-use software to compute nodewise predictability in network models and to visualize it in a typical network visualization. Predictability is an important concept that *complements* the network structure when interpreting network models: it gives a measure of how well a node can be predicted by all its neighboring nodes and is hence crucial information whenever one needs to judge the practical significance of a set of edges. An example is clinical practice, where it is important to make predictions of the outcome of interventions on an interpretable scale to optimally select treatments.

The analyses shown in the present chapter can be extended to networks that are changing over time, which allows to investigate how edge-parameters and nodewise predictability change over time. The time-varying parameters can then be modeled by a second model, which could include variables from inside and outside the time-varying network. With this modeling approach, it would be possible to gather evidence for the event of one (or several) variables causing the system to transition into another state, which is possibly reflected by a different network structure and nodewise predictability. For details about how to fit time-varying network models and time-varying predictability measures see Haslbeck and Waldorp (2020); Haslbeck et al. (2020).

Several limitations for the interpretation of nodewise predictability require discussion. First, we can only interpret the predictability of a node as the influence of its neighboring nodes if the network model is an appropriate model. A network model can be inappropriate for a number of reasons:

1. Two or more variables in the network models are caused by a variable that is not included in the network. This results in estimated edges between these variables in the network, even though they are only related via an unobserved common cause. In this situation we cannot interpret predictability as influence by neighboring nodes, because we know that the nodes are not influencing each other but are caused by a third variable outside the network.



2. In some situations variables are logically dependent, for instance *age* and *age of diagnosis* are always related, because one cannot be diagnosed before being born. Clearly, in this situation the relation between the variables must be interpreted differently.
3. If two or more variables measure the same underlying construct (e.g., five questions about sad mood). In this situation the edge-parameters indicate how similar the variables are and do not reflect mutual causal influence. Consequently, we would not interpret the predictability of these variables as the degree of determination by neighboring nodes. See Fried and Cramer (2016) for a discussion of this problem. Solutions could be to determine the topological overlap (Zhang, Horvath, et al., 2005) and choose only one variable in case of large overlap or to incorporate measurement models into the network model (Epskamp, Rhemtulla, & Borsboom, 2016).

Second, if the data was generated by a Directed Acyclic Graph (DAG) (e.g., Peters, Janzing, & Schölkopf, 2017) and if a collider structure is present in this DAG (i.e., A causes B and C), then the network models discussed in this chapter imply a small spurious edge between variables B and C. This spurious edge will lead to a small over-estimation of the predictability measures of nodes B and C.

Third, if we interpret the predictability of node A as a measure of how much it is determined by its neighbors, we assumed that the causal influence of the edges goes from the neighbors to node A. However, the direction of edges is generally unknown when the model is estimated from cross-sectional data. Estimates about the direction of edges can be made using causal search algorithms like the PC algorithm (Spirtes et al., 2000) or by using substantive theory. This means that the predictability of a node is an upper bound and in practice often lower, because the causal effect points away from the node at hand or is bi-directional. While this is a major limitation, note that this is true for any model estimated on cross-sectional data. In models with lagged predictors like the VAR model, this problem does not exist, because we use the direction of time to determine the causal direction.

Finally, it is important to stress that a topic we did *not* cover here is to investigate how well A can be predicted *by node B*. This is different from the problem studied in this chapter, where the interest was in how well node A can be predicted *by all other nodes*. Unfortunately, there are no straightforward solutions for the former problem in the situation of correlated predictors, which is always the case in practice. For linear regression, there is work on decomposing explained variance (Grömping, 2012) and in the machine learning literature there are methods to determine variable importance by replacing predictor variables by noise and investigate the drop in predictability (e.g., Breiman et al., 2001). It would certainly be interesting to try to extend these ideas to the general class of network models.

To sum up, if the network model is an appropriate model for the phenomena at hand, predictability is an easy to interpret measure of how strongly a given node is influenced by its neighbors in the network. This allows researchers to judge the practical relevance of edges connected to a node A on an absolute scale

(0 = no influence on A at all, 1 = A fully determined) and thereby helps to predict intervention outcomes. In addition, the predictability of (parts of) the network is interesting on a theoretical level, as it indicates how self-determined the network is.

## Acknowledgements

We would like to thank Pia Tio, Joris Broere, Max Haslbeck, Benjamin Rosche, Adela Isvoranu, Matthias Huber and Fabian Dablander for helpful comments and Sacha Epskamp for incorporating nodewise error visualizations in the *qgraph* package. In addition, we would like to thank two anonymous reviewers for helpful comments.

# NODEWISE PREDICTABILITY: REANALYSIS

---

## Abstract

Network analyses on psychopathological data focus on the network structure and its derivatives such as node centrality. One conclusion one can draw from centrality measures is that the node with the highest centrality is likely to be the node that is determined most by its neighboring nodes. However, centrality is a relative measure: knowing that a node is highly central gives no information about the extent to which it is determined by its neighbors. In this chapter, we provide an absolute measure of determination (or controllability) of a node its predictability. We introduce predictability, estimate the predictability of all nodes in 18 prior empirical network papers on psychopathology, and statistically relate it to centrality. We investigate predictability in 25 datasets from 18 published papers that studied psychopathology (several mood and anxiety disorders, substance abuse, psychosis, autism, and transdiagnostic data) from a network perspective. Predictability was unrelated to sample size, moderately high in most symptom networks, and differed considerable both within and between datasets. Predictability was higher in community than clinical samples, highest for mood and anxiety disorders, and lowest for psychosis. We argue that predictability is an important additional characterization of symptom networks because it gives an absolute measure of the controllability of each node. It also allows to judge how self-determined a symptom network is, and may help to inform intervention strategies.

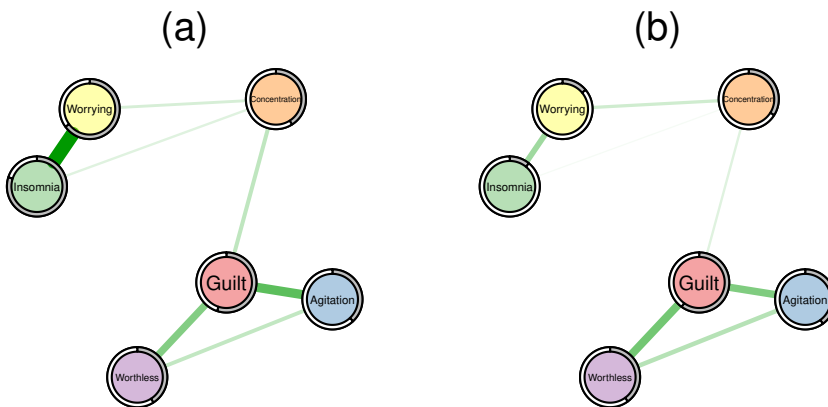
---

This chapter has been adapted from: Haslbeck, J. M. B., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? A reanalysis of 18 published datasets. *Psychological Medicine*, 47(16), 2767-2776.

## 4.1 Introduction

In the network approach to psychopathology, mental disorders are understood as networks of interacting symptoms, and by studying the structure of these networks one hopes to find explanatory models for the etiology of disorders and effective interventions (Cramer et al., 2010; Borsboom & Cramer, 2013). This perspective has provided new and intuitively appealing explanations of psychopathological phenomena, and has been applied to many different mental disorders, has been described in detail elsewhere (for a review, see Fried et al., 2017).

While the analysis of the *structure* of symptom networks has led to important insights, in this chapter we focus on another important characteristic that has not been considered so far in the literature: *predictability*, i.e. the degree to which a given node can be predicted by all other nodes in the network. Predictability is an important measure because it tells us on an interpretable absolute scale (e.g. 40% variance explained) how much a node is determined by other nodes in the network. Thereby, predictability gives us an idea of how clinically relevant connections (also called ‘edges’) are: if node A is connected to many other nodes, but these together explain only 1% of its variance, then these edges are not interesting in many situations. As an example, take the problem of selecting an intervention on insomnia in two hypothetical symptom networks in Figure 4.1:



**Figure 4.1:** Two example symptom networks with different predictability measures. Left: insomnia is strongly determined by the nodes connected to it (80% variance explained as indicated by the grey pie chart). Right: insomnia is weakly determined by the nodes connected to it (11% variance explained).

In the network of the first patient (a), 80% of the variance of insomnia is explained by the two nodes that are connected to it, worrying and concentration problems, as indicated by the grey area in the ring around the node; it is plausible that an intervention on worrying may have a considerable impact on the sleep problems. In contrast, in the network of the second patient (b), insomnia is only weakly determined by its neighbors (11% variance explained), and an ef-

efficient intervention on insomnia via worrying seems questionable. Instead, we should search for relevant variables outside the current network that have an effect on insomnia, or may want to consider intervening directly on insomnia, e.g. by administering sleeping pills. Predictability thus helps us to estimate the potential success of clinical interventions on a symptom via the symptom network and could thereby guide treatment selection.

Clearly, predictability depends on the number and the strength of the edges a node is connected to: a node with many strong edges tends to have a higher predictability than a node with few weak edges. For instance, we can expect from the edge weights in Figure 4.1 that insomnia is *better* predicted in network (a) than (b). However, we do not know how well we can predict insomnia on an absolute scale in either case. In contrast, predictability does provide such an absolute scale and thereby goes beyond the network structure and centrality indices reported in prior literature.

In summary, this work makes the following contributions:

1. We introduce predictability as an additional measure to characterize network models, and discuss the relationship between predictability and derivatives of the network structure, such as centrality measures.
2. We provide a reproducible example (including R syntax) of how to estimate and interpret predictability in psychopathological networks using the data on bereavement and depression from Fried et al. (2015), serving as a tutorial for researchers.
3. We provide a first glimpse into predictability in the field of psychopathology by re-analyzing 25 datasets from 18 published papers that used network analyses. We discuss theoretical implications of the variability of predictability within and between networks and the relation between predictability and the network structure. In addition, we make our syntax, all datasets we are allowed to share (4/25), as well as the correlation matrices and adjacency matrices (i.e. the network structures) of *all* 25 datasets available.

## 4.2 Methods

### 4.2.1 Literature Review & Data

We aimed to identify all papers in the field of psychopathology that applied network analysis techniques to cross-sectional data. To this end, we combined all papers known to the authors with the results of a literature review: we searched the databases PsycNET, ISI Web of Science and GoogleScholar using the names of the most prevalent mental disorders in combination with “Network” as keywords. This literature review yielded 23 papers published between 2010 and 2016. We excluded one paper as the used data was identical to the data used in another paper. We contacted the authors of the remaining 22 papers and were

able to obtain the data of the 12 papers described in Table 4.1. For further details about the literature review see Appendix A.1. The authors in the respective papers estimated Gaussian Graphical Models (GGMs) using the `qgraph` package (Epskamp et al., 2012), Ising models using the `IsingFit` package (van Borkulo, Borsboom, et al., 2014), and the parameters for relative importance networks using the `relaimpo` package (Grömping, 2012; Grömping et al., 2006) (see column “original analysis” in Table 4.1). Datasets predominantly feature symptoms or clinical problems as nodes, although some contain contextual variables (e.g., *age of diagnosis* in Deserno et al. (2017)).

### 4.2.2 Statistical Methods

We fitted GGMs to the continuous datasets and Ising models to the binary datasets. These models are considered the state-of-the-art and were also used in most of the papers included in our re-analysis (see Table 1). For an accessible tutorial on how to estimate GGMs, see Epskamp and Fried (2018). We computed predictability measures using the R-package `mgm` (Haslbeck & Waldorp, 2020). Note that in the case of GGMs, our estimation procedure was slightly different than the one in the original analyses as we did not estimate polychoric correlations before using the correlation matrix to estimate the graph structure using the graphical lasso (e.g., Epskamp et al., 2012). We instead used the neighborhood regression approach implemented in the `mgm` package, which is necessary to compute predictability. In the case of the Ising Model, there are no differences since the node-wise estimation of `mgm` is identical to the node-wise estimation in `IsingFit` (van Borkulo, Borsboom, et al., 2014). Note that the reported sample size in Table 4.1 in some cases differs from the one reported in the original study. In these cases, the authors deleted missing values pairwise to compute the sample covariance matrix and reported the full sample size. With the neighborhood regression approach, however, we have to delete missing values casewise, resulting in a smaller number of observations.

As predictability measures we selected the proportion of explained variance for (centered) continuous variables and a normalized accuracy measure for binary variables. The normalized accuracy measure quantifies how a node is determined by its neighboring nodes beyond the intercept model. This is important, because for instance if a binary variable with 100 cases has 5 zeros and 95 ones, then the intercept model (which predicts a one for each case) alone would already lead to an accuracy of 95% without considering any other nodes. The normalized predictability measure takes this into account and is zero when other variables do not predict the node at hand beyond the intercept model; a more detailed explanation of both proportion of explained variance and normalized accuracy can be found in Haslbeck and Waldorp (2018). Both measures range from 0 to 1: 0 means that we cannot at all predict a node by other nodes in the network, whereas 1 implies perfect prediction. In addition to predictability, we computed the the following node centrality measures: weighted degree centrality, betweenness centrality, closeness centrality and eigenvector centrality (M. E. Newman, Barabási, & Watts, 2006).

Paper	Subfield	Datatype	p	n	Original Analysis
Anderson et al. (2015)	Autism	Continuous	14	477	Correlation
Armour et al. (2017)	PTSD	Continuous	27	221	GGM
Beard et al. (2016)	Anxiety, Depression	Continuous	17	1029	GGM
Borsboom and Cramer (2013)	Anxiety, Depression	Binary	18	9282	Ising model
Boschloo, van Borkulo, et al. (2016)	General	Binary	12	501	Ising model
Deserno et al. (2017)	Autism	Continuous	17	301	GGM
Fried et al. (2015)	Bereavement	Binary	12	515	Ising model
Fried, Epskamp, et al. (2016)	Depression	Continuous	28	3463	GGM
Goekoop and Goekoop (2014)	General	Continuous	63	192	Correlation
Hoorelbeke et al. (2016)	Depression	Continuous	6	69	GGM
Koenders et al. (2015)	Bipolar	Continuous	16	126	Correlation
McNally et al. (2015)	PTSD	Binary	17	362	Ising model
Rhemtulla et al. (2016)	Substance Abuse	Binary	11	2405	Ising model
Robinaugh et al. (2014)	Bereavement	Continuous	19	1532	GGM
Robinaugh et al. (2016)	Complicated Grief	Continuous	13	195	GGM
Ruzzano et al. (2015)	Autism	Binary	17	213	Correlation
Santos Jr et al. (2017)	Depression	Continuous	20	503	GGM
Wigman et al. (2016)	Psychosis	Binary	56	283	GGM

**Table 4.1:** Characteristics of Papers included in the Data Reanalysis. GGM stands for Gaussian Graphical Model. Datatype refers to the variables after preprocessing as performed in the original papers; this means that some datasets were actually on an ordinal scale with more than two categories, but were binarized for the analysis by the original papers (we did the same)

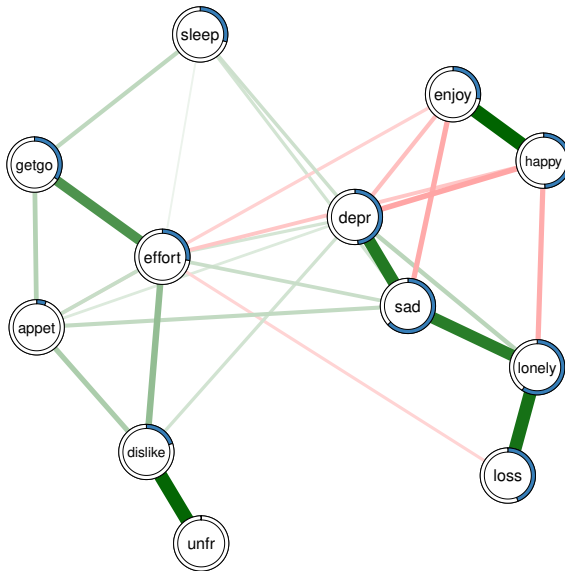
Rhemtulla et al. (2016) split their data in six subgroups (abuse of cannabis, sedatives, stimulants, cocaine, opioids or hallucinogens) and Koenders et al. (2015) used three subgroups (mildly depressed, predominantly depressed, cycling). We followed the analyses in their papers and estimated six and three separate networks (see also Figure 4.3), respectively. Overall, this led to 25 datasets/networks from 18 papers.

## 4.3 Results

### 4.3.1 Application example: node-wise predictability in data of Fried et al. (2015)

Before discussing the results of the re-analysis of all papers, we provide an example how to estimate and interpret predictability using the depression and bereavement dataset analyzed in Fried et al. (2015). In their paper, they re-analyzed the Changing Lives of Older Couples study (Carr, Nesse, & Wortman, 2005). The network in Figure 4.2 represents the cross-sectional network structure of 10 dichotomous depression symptoms (measured via the 10-item CES-D) and 1 condition node (*loss*), which codes whether participants belong to the bereaved group who had lost their spouse prior to this follow-up time point, or the still-married control group. Several results of the predictability analysis are noteworthy.

First, the average predictability across all nodes is 0.34, indicating that 34% of the variance of a node that is not predicted by the intercept model is explained by its neighbors. Compared to the predictability results of all other datasets (see below), this is an average level of predictability.



**Figure 4.2:** Ising model estimated on the data of Fried et al. 2015. Green edges indicate positive relationships, red edges indicate negative relationships. The blue ring around each node represents its predictability. *loss*, spousal loss; *depr*, depressed; *effort*, everything is an effort; *sleep*, restless sleep; *unfr*, people are unfriendly; *enjoy*, enjoy life; *appet*, poor appetite; *dislike*, people dislike me; *getgo*, cannot get going.

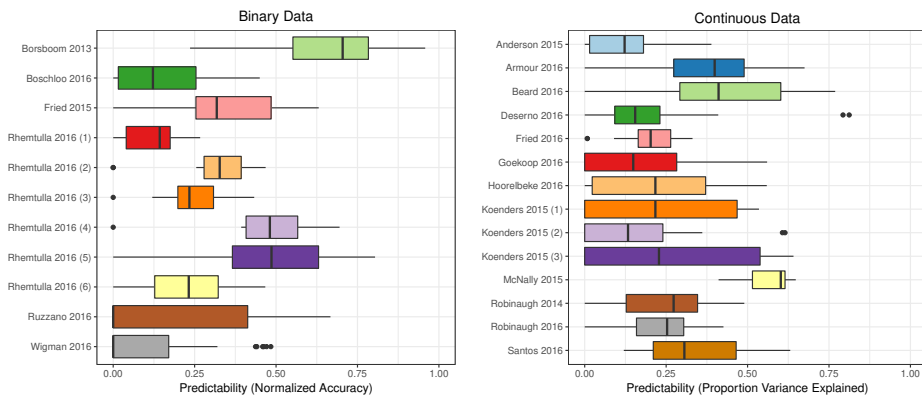
Second, *appet* (poor appetite) and *unfr* (people are unfriendly) stand out with the lowest predictability estimates in the network (.06 and 0), implying that all other nodes together share nearly no variance with these items. The low predictability of poor appetite is consistent with psychometric studies of depression scales, showing that weight and appetite items often form a distinct cluster of nodes (or factor) and show only weak partial correlations with other depression symptoms (e.g., Fried et al., 2016). In contrast, the low predictability *unfr* is likely explained by the low variance in this variable: 94% of the cases report the symptom to be absent. This leads to a situation in which the model including the neighbors gives the same predictions as the intercept model. Because the normalized predictability measure used here captures the predictability beyond the intercept model, we get a measure of zero (for details see Haslbeck & Waldorp, 2018).

Finally, negative emotions such as *depr*, *sad*, and *lonely* have comparably high predictability values (.48, .63 and .59). This could either be due to the fact that these items measure different concepts that strongly influence each other, or because they capture similar constructs (e.g., *depr* and *sad* may tap into the same emotion).



### 4.3.2 Re-analysis of 25 datasets

We now turn to the re-analysis of 25 datasets from 18 published papers in the psychopathology network literature. Figure 4.3 shows box plots describing the distributions of predictability measures for all included datasets. In general, we see that symptoms in networks can often be predicted reasonably well by all other symptoms in the network.



**Figure 4.3:** Summaries of the distribution of predictability measures for datasets with continuous (left) and binary (right) data. The x axis shows the predictability measure (ranging from 0 to 1): ‘normalized accuracy’ for binary variables and ‘proportion of explained variance’ for continuous variables. The box plot whiskers show 1.5 times the Interquartile Range (IQR)

A few things stand out. First, node-predictability varies considerably within datasets, as can be seen by the spread in the distributions of predictability measures that is summarized in the boxplots: the bold vertical bar corresponds to the median, the box indicates the 25% and 75% quantiles, and the whiskers show 1.5 times the interquartile range (IQR).

Second, there is a considerable amount of node-predictability variation across datasets. This difference is not trivially explained by differences in sample size between datasets: the Spearman correlation between mean predictability and sample size is only 0.07. In addition, we explored whether predictability differences across samples were related to severity of psychopathology. To that end, we classified all datasets into an ordinal variable indicating severity (0 = all healthy, 1 = mixed, 2 = clinical populations). The weighted (by number of observations) Spearman correlation between this severity variable and predictability was -0.82, providing evidence that networks of clinical samples may have a lower mean predictability than networks of healthy samples. This is consistent with findings of lower dimensionality of symptom networks of healthier patients (Fried, van Borkulo, et al., 2016). However, these results are at best preliminary because a clear classification of datasets into predominantly healthy, sick, or mixed was difficult, and because we analyzed a small number of often highly heterogeneous datasets. Like many other results presented in this chapter, these analyses serve as example of what research questions can be explored with predictability results,

rather than strong evidence.

Third, it stands out that the six substance abuse subsamples of Rhemtulla et al. (2016) differ considerably in their mean predictability. A possible explanation for these differences is that the symptoms are consequences of a common cause—the consumed substance—and that the influence of this common cause is differentially strong for different substances (e.g., stronger for opioids than cannabis). A similar argument could be made for the datasets on PTSD: symptoms may co-vary (and hence predict each other well) because they are all caused by the traumatic experience. This contrasts with the network approach to psychopathology, and we will turn to this issue of (unobserved) common causes in the discussion.

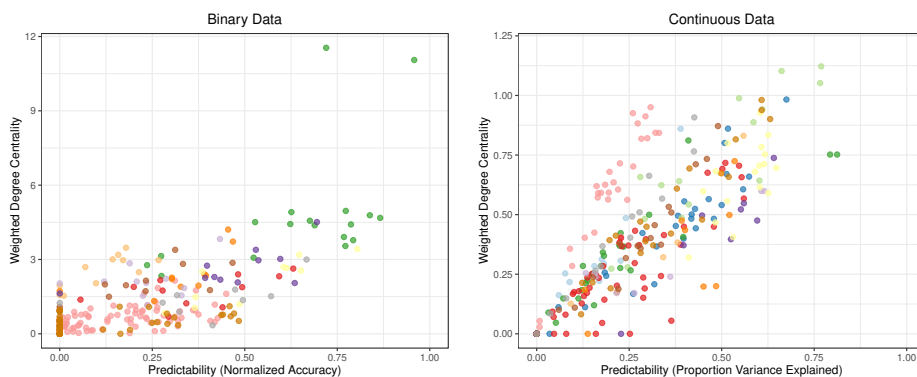
Fourth, we observed a very high mean predictability for the depression network of Borsboom and Cramer (2013). This finding is at least in part explained by the fact that the authors replaced skipped questions with zeros (Borsboom et al., 2017). This procedure leads to spurious relationships, because variables become related via their shared missing value pattern that is determined by the structure of the skip questions (0s are imputed for multiple variables at the same time, inducing correlations among these items). We also observed a very high predictability of 2 items in the paper on autism by Deserno et al. (2017) (see the outliers in Figure 4.3)—age and age of diagnosis. These have to be strongly correlated, because the former is an upper bound for the latter, i.e. a person cannot get a diagnosis at the age of 15 if the person is 9 years old.

Finally, the question arises whether predictability differs consistently across different types of datasets, for instance, across mental disorders. Differences in predictability across mental disorders can be interpreted as evidence for how self-determined a symptom network is: if predictability is high, the symptoms are largely determined by each other, if predictability is low, symptoms are largely influenced by additional variables (e.g. biological, environmental or additional symptoms) that are not included in the network. Figure 4.3 suggests that symptoms of depression and anxiety disorders might be more self-determined (average predictability = 0.38), while the symptoms of psychosis might be determined to a larger degree by other influences such as genes or environmental variables (0.10). Other explanations for the pattern of findings could be that the measurement error is larger for symptoms of psychosis, or that depression and anxiety assess very similar problems multiple times, which increases their respective predictability. Apart from comparing predictability across types of mental disorders, we could also investigate whether the predictability is higher for female vs male, or younger vs. older patients. While we do not have sufficient data to answer these questions, Figure 4.3 provides numerous possibilities that should be investigated in more focused future studies.

### 4.3.3 Relationship between Predictability and Edge Weights

It is clear that there has to be a close relationship between the predictability of a node and the edge weights connected to that node: if a node is unconnected, its predictability by other nodes has to be zero. And the more edges are connected

to a node, the higher its predictability tends to be. We illustrate this relationship using weighted degree centrality (the sum of absolute edge-weights connected to a node), which had the highest mean correlation with predictability (0.79, 0.70, 0.36 and 0.27 for weighted degree centrality, eigenvalue centrality, closeness centrality and betweenness centrality, respectively): in Figure 4.4, we plot weighted degree against predictability, for all datasets shown in Figure 4.3.



**Figure 4.4:** The relationship between weighted degree centrality (x axis) and predictability (y axis) of each node in the datasets with continuous (left) and binary (right) data. The colors of the points correspond to the colors used for different papers in Figure 4.3.

Each point corresponds to one node and its color indicates to which dataset it belongs (see Figure 4.3). As expected, we observe a positive relationship between the centrality of a node and its predictability. This relationship is stronger for continuous-Gaussian variables, because here the edge weights (which are partial correlations in this special case) are always between  $-1$  and  $1$ , whereas edge weights in the Ising model for binary data are only constrained to be finite. However, the relationship is far from perfect: for example, for continuous-Gaussian variables, a centrality measure of  $0.25$  can coincide with a predictability measure between  $0.1$  and  $0.5$  and for binary variables, a centrality measure of  $3$  can coincide with a predictability measure between  $0.1$  and  $0.8$ .

It is crucial to note, however, that centrality gives us only relative information about predictability: even if both measures would be correlated perfectly, we could only order all nodes by predictability, but we would not know the absolute value of the predictability of any node. This is similar to the correlation of the actual height of children in a classroom with their position in an ordering by height: these two metrics may be considerably related, but we can never know how tall Alice is from knowing she is the 5th tallest girl in class.

It would be possible to fit a regression model to predict predictability from degree centrality. However, both the mean predictability (see Figure 4.3) and the strength of the linear relationship between both measures (see Figure 4.4) differ greatly between data sets, which implies that a predictability inferred from centrality would be highly inaccurate. Given that predictability can be easily

computed with freely available software (Haslbeck & Waldorp, 2020), we see no reason to accept these inaccuracies.

## 4.4 Discussion

We showed that predictability is an important characteristic of network models in addition to their structure. Furthermore, we provided a first overview of how high predictability typically is in the field of psychopathology and suggest that analyzing predictability across disorders and groups of individuals may generate new theoretical insights.

Predictability was moderately high in most datasets, indicating that a considerable amount of the variation of nodes can be explained by other nodes in the network. We found that the average predictability was higher for certain disorders (e.g. depression, anxiety, PTSD) than for others (e.g. psychosis). This suggests that the symptom network of the former disorders is more self-determined, while nodes for the latter disorders are more strongly influenced by other factors that are not included in the network, such as additional symptoms or biological and environmental variables. We thus see predictability as a first attempt towards characterizing the *controllability* of the symptom network: if predictability is high, we can control symptoms via their neighboring symptom in the network – if it is low, we have to search for additional variables or intervene on the symptom directly. If our findings of low predictability for specific disorders or groups of patients can be replicated in future studies, this calls for research on important variables beyond common symptoms.

In clinical practice, predictability enables us to judge the efficacy of a planned treatment: if the neighbors of symptom A explain 90% of its variance, an intervention on symptom A via its neighbors seems viable. In contrast, if they explain only 5% of the variance, one would rather search for additional variables outside the network or try to intervene on the node directly (instead of trying to control the node via neighboring nodes).

It is important to note several limitations of the present chapter. First, we only analyzed a small and heterogeneous sample of datasets (all available datasets we could obtain for this project), and a much larger database of studies is required to draw any strong conclusions when comparing, for instance, the predictability of different types of mental disorders. Due to the increasing popularity of network models in psychopathology, we look forward to having more data available in the next few years to tackle these and related questions.

Second, the present chapter explored how well node A is predicted by all its neighbors. Another interesting question is how well node A is predicted by one particular neighboring node B. Unfortunately, there is no straightforward solution to this problem in the case of correlated predictors, which is nearly always the case in psychopathology data. For continuous-Gaussian data, solutions to this problem have been proposed that are based on variance decomposition (Grömping, 2012), and there are more general methods in the machine learning literature based on replacing a predictor by noise and investigating the drop

in predictability (e.g., Breiman et al., 2001). While the performance of these methods is not always clear and requires further work, it would be important to extend and apply these approaches to the network models used in psychopathology research. From this limitation follows that we cannot quantify the ‘predictive power’ or ‘degree of determination’ of a given node on its neighboring nodes on an absolute scale (the causal opposite of predictability or controllability). However, if predictability is low for all nodes in the network, we do know that no node exerts a strong influence on any other node in the network.

Third, when calculating predictability of node A, we assume that all edges are directed towards that node A, i.e. that all adjacent nodes are causing A, but not vice versa. However, we do not know whether this is true because the direction of edges is generally unknown in cross-sectional data (Pearl, 1991). It follows that the predictability of a node is an upper bound for how much it is determined by the nodes it is connected to. While it is important to keep this limitation in mind, it may not matter that much in many situations: for instance, if the predictability of symptom A is too low to render an intervention via neighboring symptoms viable, it does not matter that the true predictability is probably lower. The predictability estimate can be improved by any method that reliably replaces our assumption about the direction of edges by estimates about the direction. In cross-sectional data, the direction of edges can under a set of stringent assumptions be inferred via causal search algorithms such as the PC-algorithm (Spirtes et al., 2000)). In time-series data with lagged effects, this problem is circumvented by using the direction of time: if A and B are related and A precedes B in time, then we say that A Granger-causes B and we have a directed edge from A to B (Granger, 1969). The predictability measure we propose here can easily be applied to these time-series models as well (Haslbeck & Waldorp, 2020).

Fourth, the interpretation of predictability of a node as the degree to which it is determined by the node it is connected to is only appropriate if the network model is an appropriate model for the phenomenon at hand. A network model can be problematic or even inappropriate for a number of reasons (see also Haslbeck & Waldorp, 2020; Fried & Cramer, 2017). In the presence of two or more variables that measure the same underlying construct (e.g. several questions about sad mood) we would not interpret connections between those variables as genuine causal relations and hence we also would not interpret predictability as a measure of determination. Another problem is a situation in which variables are deterministically related such as the variables age and age of diagnosis in the paper of Deserno et al. (2017). Clearly in this case, we would not think of a process in which age is causing age of diagnosis or vice versa. Moreover, it is problematic if two or more nodes have a common cause that is not included in the network, because this leads to a spurious edge between these nodes. In all three cases, interpreting edges as genuine cause-effect relationships is incorrect, and the interpretation of predictability as degree of determination by neighboring nodes does not apply. This could be the case for substance abuse and PTSD where substance use and traumatic experiences may be common causes for (parts of) the symptom network (Fried & Cramer, 2017). While this is a major limitation, it applies to any other statistical model as well: for instance, interpret-

ing Cronbach's alpha or factor loadings in factor models makes only sense in case the factor model is the appropriate model for the data.

In sum, predictability is a useful additional characterization of psychopathological networks, may have direct implications for clinical practice, and provides a method to investigate theoretical questions such as the degree of self-determination of a network.

## **Acknowledgements**

The authors would like to thank Kamran Afzali, Denny Borsboom, Gary Brown, Tiago Cabaço, Sara Plakolm, Mijke Rhemtulla and Lourens Waldorp for helpful comments. We would also like to thank all authors who made their data available for this project, and especially the four author teams who shared their data publicly.

---

# MODERATED NETWORK MODELS

---

## Abstract

Pairwise network models such as the Gaussian Graphical Model (GGM) are a powerful and intuitive way to analyze dependencies in multivariate data. A key assumption of the GGM is that each pairwise interaction is independent of the values of all other variables. However, in psychological research this is often implausible. In this chapter, we extend the GGM by allowing each pairwise interaction between two variables to be moderated by (a subset of) all other variables in the model, and thereby introduce a *Moderated Network Model* (MNM). We show how to construct MNMs and propose an  $\ell_1$ -regularized nodewise regression approach to estimate it. We provide performance results in a simulation study and show that MNMs outperform the split-sample based methods Network Comparison Test (NCT) and Fused Graphical Lasso (FGL) in detecting moderation effects. Finally, we provide a fully reproducible tutorial on how to estimate MNMs with the R-package *mgm* and discuss possible issues with model misspecification.

## 5.1 Introduction

Network (or graphical) models are a powerful and intuitive tool to analyze dependencies in multivariate data and their popularity has recently surged in psychological research (e.g., personality psychology (Costantini et al., 2015), social psychology (Dalege et al., 2016), and clinical psychology (Borsboom, 2017; Haslbeck & Fried, 2017; Eidhof et al., 2017; Kendler, Aggen, Flint, Borsboom, & Fried, 2018)). The network models used in this literature model interactions between *pairs* of variables, for instance *mood* and *physical activity*. Examples for such pairwise network models are the Gaussian Graphical Model (GGM) implied by the multivariate Gaussian distribution (Lauritzen, 1996), the Ising model for binary-valued data (Wainwright et al., 2008) and the Mixed Graphical Model (MGM) for data from mixed distributions (S. Chen, Witten, & Shojaie, 2014; Yang et al., 2014a).

A key assumption of these pairwise network models is that there are no moderation effects, which means that the interaction between any pair of variables is independent of the values of *all other* variables in the network. As an example, let's say we have the variable *fatigued* in the model. Then this assumption says that the (possibly positive) interaction between *mood* and *physical activity* does not depend on how fatigued an individual is. This is an empirical question and so seems worth checking. Because psychology studies highly contextualized and multicausal phenomena, the occurrence of such moderation effects is often plausible. Psychological researchers have known this for a long time, which is reflected by the widespread use of moderation in the analysis of psychological data: either explicitly as moderation analysis in the regression framework (e.g., Fairchild & McQuillin, 2010; MacKinnon & Luecken, 2008) or as interaction terms, for instance, in a 2-way Analysis of Variance (e.g., Tabachnick & Fidell, 2007).

Moderation is important in network models for the same reason it is important in these well-known analyses: missing a moderator means getting an average model (over the values of the missed moderator) that is potentially inappropriate for individuals with *any* value on the moderator variable. This averaging can lead to detecting no effect at all, even if there is a strong moderation effect (see Section 5.2.1). In addition, differences in the distributions of the missed moderator variable across studies offer one explanation for contradicting evidence across studies (Collaboration et al., 2015; Collaboration, 2012): for example, study A might have studied mostly well-rested individuals and found a positive relationship between *mood* and *physical activity*, while study B might have studied mostly fatigued individuals and found a negative relationship. Moderators are also important in clinical practice: for example, (Borsboom, 2017) suggested that one can intervene on a symptom network of a mental disorder both on the node- and the network level. Since moderators determine the relationships between variables, they are the natural tool to find possible interventions on the network. More generally, moderators are useful in applications if a medication/treatment only works for a certain group of patients (defined by the moderator variable) and does not work (or even has adverse effects) for other groups of patients. Thus,



studying moderator variables is a key requirement for moving towards personalized medicine and therapy (e.g., Hamburg & Collins, 2010).

One way to check for moderation effects in network models is to split the data set in two parts along the moderator variable, estimate a network model on each of them, and compare them. A more sophisticated version of this procedure is the Network Comparison Test (NCT) (van Borkulo et al., 2016) which performs a permutation test on differences across data sets for each edge-parameter; another procedure is the Fused Graphical Lasso (FGL), which jointly estimates two Gaussian Graphical Models (GGMs) on two data sets, with an additional penalty on differences between the two GGMs (Danaher, Wang, & Witten, 2014). These data-split approaches have two major draw-backs: first, the type of moderation effect they approximate is a step function, with the step placed at the value of the moderator variable at which the data set was split. Such a step function is implausible in many situations. In the above example, this would mean that the relationship between *mood* and *physical activity* remains (possibly positively) constant while increasing the value of *fatigue*. And at some specific value of *fatigue*, the relationship “jumps” to another (possibly negative) constant. Second, splitting the data set in half means losing information, because now both network models have to be estimated on half the data compared to the original data set. This leads to greatly reduced sensitivity to detect both pairwise interactions and moderation effects.

We propose a more direct approach to detecting moderation. Specifically, we extend pairwise network models by a specified set of moderator variables to obtain a *Moderated Network Model* (MNM). This circumvents the above mentioned problems of the split-data approaches by fitting a linear moderation effect and making full use of the data. We do this by using the standard moderation definition from multiple linear regression and extend the multivariate Gaussian distribution with moderation effects. In a similar way one could also extend Ising models and more generally, MGMs, with moderation effects. Here, we take the first step by extending the popular GGM with moderation effects. Specifically, we make the following contributions:

1. We introduce moderation for network models by extending pairwise network models with moderation effects similar to moderation effects in the linear regression framework
2. We suggest a new visualization of moderated network models, based on factor graphs
3. In a simulation study, we investigate the performance of the moderated network model in estimating moderation effects and compare it to the performance of the sampled-split based methods NCT and FGL
4. A fully reproducible tutorial demonstrates how to fit and visualize Moderated Network Models using the R-package *mgm*

In Section 5.2, we briefly review moderation in linear regression (5.2.1) and then show how to construct a Moderated Network Model (5.2.3). In the last two

subsections of Section 5.2 we show how to visualize (5.2.4) and estimate MNMs (5.2.5) using an  $\ell_1$ -regularized nodewise regression approach. In Section 6.3 we report the performance of our estimation approach in estimating MNMs, and compare its performance in recovering moderation effects to the split-sample methods NCT and FGL. Finally, in Section 5.4, we provide a fully reproducible tutorial (6.4) on how to estimate MNMs with the R-package *mgm* and discuss possible issues with model misspecification (5.4.2).

## 5.2 Moderated Network Models

In this section, we first review basic concepts of moderation in multiple regression, which are useful for introducing MNMs (Section 5.2.1). Using these concepts, in Section 5.2.3 we construct the MNM by extending the multivariate Gaussian with 3-way interactions. In Section 5.2.4, we show how to visualize MNMs using factor graphs and in Section 5.2.5, we present an  $\ell_1$ -regularized nodewise regression approach to estimate MNMs.

### 5.2.1 Moderation in Linear Regression

Here we review basic concepts of moderation in multiple regression, which we use to construct MNMs. Readers who are familiar with these concepts can skip directly to Section 5.2.3.

#### 5.2.1.1 Moderation and Interactions in Linear Regression

By moderation we mean that the effect of the predictor  $B$  on response variable  $A$  is a *linear* function of a third variable  $C$ . The simplest possible example to introduce moderation is a linear regression model in which  $A$  is a function of  $B$  and  $C$

$$A = \beta_B B + \beta_C C + \varepsilon, \tag{5.1}$$

where  $\beta_B$  is the effect of  $B$  on  $A$ ,  $\beta_C$  is the effect of  $C$  on  $A$ , and  $\varepsilon$  has a Gaussian distribution with mean  $\mu = 0$  and variance  $\sigma^2$  (e.g., Aiken, West, & Reno, 1991). In this model, both effects  $\beta_B, \beta_C$  are constants and therefore not a function of any variable. This changes when adding the product interaction term  $BC$  with parameter  $\omega_{BC}$  as a predictor to the model

$$A = \beta_B B + \beta_C C + \omega_{BC} BC + \varepsilon \tag{5.2}$$

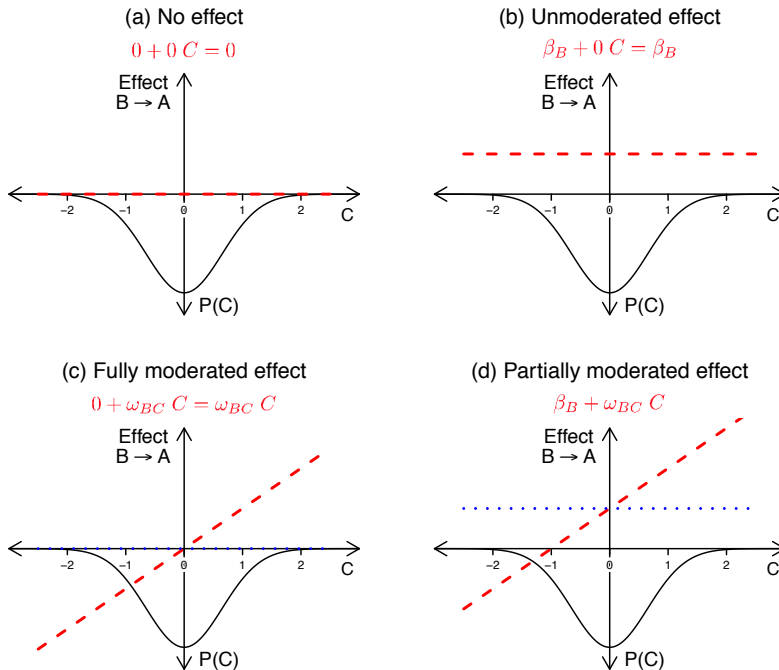
$$= (\beta_B + \omega_{BC} C) B + \beta_C C + \varepsilon \tag{5.3}$$

$$= \beta_B B + (\beta_C + \omega_{BC} B) C + \varepsilon. \tag{5.4}$$

Rewriting the model with interaction the term in (5.2) into (5.3) shows that the effect of  $B$  on  $A$  is now equal to

$$(\beta_B + \omega_{BC} C), \tag{5.5}$$

and therefore a linear function of  $C$ , with constant term  $\beta_B$  and slope  $\omega_{BC}$ . If the effect of  $B$  on  $A$  depends linearly on  $C$ , we say that this effect is linearly moderated by  $C$ . Because we can rewrite (5.2) also into (5.4),  $\omega_{BC}$  can also be interpreted as the moderation effect of  $C$  on the effect of  $B$  on  $A$ . The above rewriting shows that the interaction effect and moderation effects are different interpretations of the same parameter  $\omega_{BC}$ . Throughout the chapter we adopt the moderation perspective, because it is more intuitive and the parameter is easier to interpret.



**Figure 5.1:** The linear function in equation (5.5) determining the main effect of  $B$  on  $A$  for the four possible zero/nonzero combinations of the parameters  $\beta_B$  and  $\beta_{BC}$ , displayed both as equation and visually (dashed red line). The dotted blue line in panels (c) and (d) indicate the effect one would obtain when fitting a regression model without moderation/interaction term.

To develop some intuition for moderation, let's consider an example in which  $\beta_B = 0.2$  and  $\omega_{BC} = 0.1$ . Then, if  $C = 0$  the effect of  $B$  on  $A$  is equal to  $0.2 + 0.1 \cdot 0 = 0.2$ . And if  $C = 1$ , it is equal to  $0.2 + 0.1 \cdot 1 = 0.3$ . In this example the effect of  $B$  on  $A$  is equal to the constant  $\beta_B$  plus  $C$  times  $\omega_{BC}$ . Figure 5.1 shows four possible cases for effects of  $B$  on  $A$ .

The x-axis shows the values of the moderator  $C$ . The y-axis shows both the effect of  $B$  on  $A$  as a function of  $C$  and the density of  $C$ . In panel (a) of Figure 5.1 there is no effect of  $B$  on  $A$ ; in panel (b) there is a constant effect from  $B$  on  $A$  that is independent of  $C$ ; in panel (c) there is an effect of  $B$  on  $A$  that is fully determined (moderated) by  $C$ ; and in panel (d) the effect of  $B$  on  $A$  is equal to a constant plus a dependency on  $C$ . The dotted blue lines in panel (c) and (d) of

Figure 5.1 indicate the parameter one would obtain for the effect of  $B$  on  $A$  with a simple regression model *without* moderation/interaction term. We make two observations: first, the constant parameters (dotted blue line) are a poor description of the true moderated parameters (dashed red line). Second, (c) shows that if  $\beta_B = 0$  and  $|\omega_{BC}| > 0$ , one would entirely miss the presence of an effect when estimating a regression model without moderation.

In the procedure to estimate MNM described in Section 5.2.5 we mean-center all variables before estimation. This is required to ensure that all parameters in the MNM have a meaningful interpretation. We discuss this issue in Appendix B.1, or refer the reader to (Afshartous & Preston, 2011).

### 5.2.1.2 Regression vs. Network Semantics

In the following section we wed the world of regression with the world of networks. Unfortunately, depending on whether one adopts the regression (conditional distribution) or network (joint distribution) perspective, the same parameter is referred to differently. To avoid confusion, we make these differences explicit: In Section 5.2.1, we discussed moderation from the regression (conditional distribution) perspective. In regression,  $\beta_B$  is typically referred to as main effect, or conditional main effect in the presence of moderation, and  $\omega_{BC}$  is referred to as the moderation/interaction effect. From the perspective of the network model,  $\beta_B$  is referred to as a pairwise interaction (because it is associated with the product  $AB$  in the joint distribution), and  $\omega_{BC}$  is referred to as a 3-way interaction (because it is associated with the product  $ABC$  in the joint distribution) or a moderation effect (because it moderates the pairwise interaction  $AB$ ). The two different perspectives will become apparent in the following section, where we show both the joint distribution of the MNM and the conditional distributions. In the remainder of the chapter we adopt the network perspective, except when otherwise stated.

## 5.2.2 Gaussian Distribution and Gaussian Graphical Model

A graphical (or network model) is a statistical model for which an undirected graph/network encodes the conditional dependence structure between random variables (Lauritzen, 1996). A popular network model for continuous data is the multivariate Gaussian distribution. We introduce this distribution here, because we will use it as a basis for constructing Moderated Network Models for continuous data in the following section:

$$P(X = x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}, \quad (5.6)$$

where  $x$  is  $p$ -dimensional vector of random variables,  $\mu$  is a  $p$ -dimensional vector of means,  $\Sigma$  is a  $p \times p$  variance-covariance matrix, and  $|\Sigma|$  denotes the determinant of  $\Sigma$ .

In the case of the multivariate Gaussian distribution, it is easy to obtain the graph/network that encodes the conditional dependence structure between

random variables: if an entry in the inverse variance-covariance matrix  $\Sigma^{-1}$  is nonzero, the two corresponding variables are conditionally dependent (present edge in the network); if the entry is zero, the two corresponding variables are conditionally independent (no edge in the network). The resulting conditional (in-)dependence network is also called Gaussian Graphical Model (GGM). For an accessible introduction to GGMs, their relation to regression and Structural Equation Modeling (SEM), and how to estimate them, we refer the reader to (Epskamp, Waldorp, et al., 2018).

### 5.2.3 Construction of Moderated Network Model

The central goal of this chapter is to construct a joint distribution over  $p$  continuous variables which allows that each pairwise interaction between variables  $X_i$  and  $X_j$  is a linear function of all other variables. Another way of saying this is that each pairwise interaction between  $X_i$  and  $X_j$  is linearly moderated by all other variables. Here we construct such a joint distribution by adding moderation effects to the multivariate Gaussian distribution, which models (unmoderated) linear pairwise interactions between variables.

The density of the multivariate Gaussian distribution (5.6) shown in the previous section can be rewritten in its exponential family form as

$$P(X) = \exp \left\{ \sum_i^p \alpha_i \frac{X_i}{\sigma_i} + \sum_{\substack{i,j \in V \\ i \neq j}} \beta_{i,j} \frac{X_i}{\sigma_i} \frac{X_j}{\sigma_j} + \sum_i^p \frac{X_i^2}{\sigma_i^2} - \Phi(\alpha, \beta) \right\}, \quad (5.7)$$

where  $X \in \mathbb{R}^p$ ,  $V = \{1, 2, \dots, p\}$  is the index set for the  $p$  variables,  $\alpha$  is a  $p$ -vector of intercepts,  $\beta$  is a  $p \times p$  matrix of  $\binom{p}{2}$  partial correlations<sup>1</sup>,  $\sigma_i^2$  is the variance of  $X_i$ , and  $\Phi(\alpha, \beta)$  is the log normalizing constant which ensures that the probability distribution integrates to 1.

To see how the common form of the Gaussian density in equation (5.6) can be written as equation (5.7) above, momentarily assume that all means are equal to zero  $\mu = 0$ . Then the term in the exponential simplifies to  $-\frac{1}{2}x^T \Sigma^{-1}x$ . This inner product is the same as the sum  $\sum_{\substack{i,j \in V \\ i \neq j}} \beta_{i,j} \frac{X_i}{\sigma_i} \frac{X_j}{\sigma_j}$  in (5.7), except that we summed over each  $(i, j)$  combination twice, which is why we multiply with  $\frac{1}{2}$ . With means unequal zero, one can expand the expression in the exponential and gets a similar expression for the interactions, plus an expression for the intercepts  $\alpha$ , which are a function of the means and the interaction parameters.

Now, to extend the Gaussian distribution in (5.7) with all possible moderation effects, we add all 3-way interactions to the model:

<sup>1</sup> $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ , the  $k^{\text{th}}$  binomial coefficient of the polynomial expansion of  $(1+x)^n$ .

$$P(X) = \exp \left\{ \sum_i^p \alpha_i \frac{X_i}{\sigma_i} + \sum_{\substack{i,j \in V \\ i \neq j}} \beta_{i,j} \frac{X_i}{\sigma_i} \frac{X_j}{\sigma_j} + \sum_{\substack{i,j,q \in V \\ i \neq j \neq q}} \omega_{i,j,q} \frac{X_i}{\sigma_i} \frac{X_j}{\sigma_j} \frac{X_q}{\sigma_q} + \sum_i^p \frac{X_i^2}{\sigma_i^2} - \Phi(\alpha, \beta, \omega) \right\}, \quad (5.8)$$

where  $\omega$  is a  $p \times p \times p$  array of  $\binom{p}{3}$  3-way interactions.

How many parameters are introduced by adding 3-way interactions? For  $p = 10$  variables, adding all 3-way interactions means that the number of interaction parameters increases from  $\binom{10}{2} = 45$  to  $\binom{10}{2} + \binom{10}{3} = 45 + 120 = 165$ . Instead of adding all moderation effects (3-way interactions) one can also add single moderation effects, or all moderation effects of a subset of variables. For instance, adding all moderation effects of  $M \in \{1, 2, \dots, p\}$  moderators would result in  $\sum_{m=1}^M \sum_{i=1}^{\min\{0, p-1-m\}} i$  additional moderation parameters.

The distribution of  $X_s$  conditioned on all remaining variables  $X_{\setminus s}$  is given by

$$P(X_s | X_{\setminus s}) = \exp \left\{ \alpha_s \frac{X_s}{\sigma_s} + \sum_{\substack{i \in V \\ i \neq s}} \beta_{i,s} \frac{x_i}{\sigma_i} \frac{X_s}{\sigma_s} + \sum_{\substack{i,j \in V \\ i \neq j \neq s}} \omega_{i,j,s} \frac{x_i}{\sigma_i} \frac{x_j}{\sigma_j} \frac{X_s}{\sigma_s} + \frac{X_s^2}{\sigma_s^2} - \Phi^*(\alpha, \beta, \omega) \right\}, \quad (5.9)$$

where  $\Phi^*(\alpha, \beta, \omega)$  is the log-normalizing constant with respect to the conditional distribution  $P(X_s | X_{\setminus s})$ , and we use lower case letters  $x_i$  to indicate fixed values opposed to random variables  $X_i$ . We now show that (5.9) is a conditional Gaussian distribution. To make the presentation more clear, we set  $\sigma_s = 1$  without loss of generality. If we let

$$\mu_s = \alpha_s + \sum_{\substack{i \in V \\ i \neq s}} \beta_{i,s} x_i + \sum_{\substack{i,j \in V \\ i \neq j \neq s}} \omega_{i,j,s} x_i x_j \quad (5.10)$$

and

$$\Phi^*(\alpha, \beta, \omega) = \frac{\mu_s}{2} - \log\left(\frac{1}{\sqrt{2\pi}}\right)$$

we can rewrite (5.9) into

$$\begin{aligned}
P(X_s|X_{\setminus s}) &= \frac{1}{\sqrt{2\pi}} \exp\left\{\mu_s X_s - \frac{X_s}{2} - \frac{\mu_s}{2}\right\} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{(X_s - \mu_s)^2}{2}\right\},
\end{aligned} \tag{5.11}$$

which is the well-known form of the conditional Gaussian distribution. In Appendix B.2, we provide more intuition for the MNM by deriving the joint distribution in (5.8) from the conditionals as in (5.11) for the case of  $p = 3$  variables.

Since each  $P(X_s|X_{\setminus s})$  is only parameterized by  $\mu_s$  and regression estimates  $\mathbb{E}[X_s|X_{\setminus s}] = \mu_s$ , the above derivation shows that a Moderated Network Model can be estimated with a series of  $p$  regressions that include the appropriate moderation (3-way interaction) effects. Specifically, the equation for the mean  $\mu_s$  of the conditional distribution  $X_s$  in (5.10) has the same form as the moderated linear regression in equation (5.2) in Section 5.2.1, except that it includes more terms. In Section 5.2.5, we show how to estimate the  $p$  conditional Gaussian distributions using  $\ell_1$ -regularized regression and how to combine the resulting estimates to the MNM joint distribution.

The mean (5.10) of the conditional distribution  $P(X_s|X_{\setminus s})$  compared to the joint distribution  $P(X)$  in (5.8) explains the different terminology for interaction effects, depending on whether one adopts the regression- or graph perspective (see Section 5.2.1.2). For example, in the joint distribution, the second term indicates pairwise interactions because two variables are multiplied. In the mean of the conditional, which is estimated in a linear regression model, the second term only contains a single variable and is therefore referred to as a main effect.

Above we discussed that one could include all 3-way interactions or only a subset of 3-way interactions in the model. However, we always include all pairwise interactions. While all pairwise interactions are included in the model, this does not mean that the parameter associated with a pairwise interaction has to be nonzero if the 3-way interaction moderating that pairwise interaction is nonzero. In other words, in the joint MNM (5.8) the presence of a 3-way interaction does not imply the presence of a pairwise interaction. This is in contrast to log-linear models for categorical data, in which a  $k$ -order interaction always implies a  $(k-1)$ -order interaction

The number of parameters of the MNM is much larger than in pairwise network models, especially when  $p$  is large. If the proportion of nonzero 3-way interaction parameters in the true model would be the same as the proportion of nonzero pairwise interaction in the true model, we needed a lot more observations  $n$  to estimate the MNM with similar accuracy as the pairwise model. However, it is highly implausible that *every* variable in the model moderates *every* pairwise interaction. Instead, we would expect that *some* variables moderate *some* pairwise interactions. Under the assumption that a large fraction of moderation effects are equal to zero in the true model, it is possible to estimate the model accurately with a reasonable sample size  $n$ . In Section 5.2.5, we present an  $\ell_1$ -regularized regression procedure to estimate MNMs, which uses this assump-

tion. In Section 6.3, we explicitly show in a simulation study how much data is needed to recover a MNM in realistic situations.

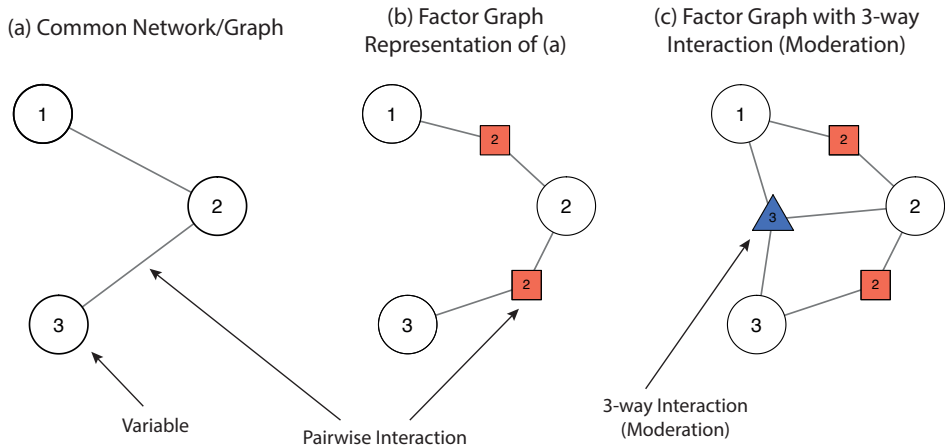
We showed that the MNM joint distribution in (5.8) can be factorized into  $p$  conditional Gaussian distributions. However, the MNM joint distribution is not a multivariate Gaussian distribution, because we added 3-way interactions (moderation effects). For the multivariate Gaussian distribution, all parameters have to be finite and the covariance matrix has to be positive-definite to ensure that the distribution is normalizable. For the MNM proposed here, the constraints to ensure normalizability are unknown. For the class of MGMs, which generalize MNMs, (Yang et al., 2014a) proposes that for normalizability it is sufficient that the sum of unnormalized terms in the exponential in (5.8) are smaller than zero, which will ensure that the unnormalized mass in (5.8) converges to zero. Thus, in order to ensure normalizability, one needs to constrain the parameter space such that this inequality is satisfied. The required constraint is most likely a constraint on the 3-way interactions, and is a function of all other parameters and the variances of the conditional Gaussian distributions. However, working out these constraints is beyond the scope of this chapter, in which we focus on introducing the idea of moderation in network models to an applied audience. In the present chapter, we therefore estimate MNMs with an unconstrained node-wise algorithm. This means that we do not know whether the estimated parameters lead to a normalizable joint distribution. One consequence of having no guarantee that the joint distribution is normalizable is that one cannot apply any global goodness of fit analyses, for example, to select between models with different sets of included moderators. While this is a major limitation of the here proposed MNM, one can perform model selection with out of sample prediction error, which does not require a proper joint distribution.

### 5.2.4 Visualizing Moderated Network Models

Pairwise network models are typically visualized in a network consisting of nodes representing variables and undirected edges representing pairwise interactions. In MNMs, we have additional moderation parameters and therefore need to find a new visualization that allows to include those without giving up clarity. We solve this problem with a factor graph visualization, in which each interaction (pairwise or higher-order) is represented by a *factor node* (see e.g., Koller & Friedman, 2009). We first show how to represent a pairwise network model as a factor graph and then demonstrate how to include moderation parameters in the factor graph visualization.

Figure 5.2 (a) shows the typical network-visualization of a pairwise network model with edges 1-2 and 2-3 indicating pairwise interactions between those two pairs of variables. Panel (b) shows the visualization of the same network model as a factor graph: now each pairwise interaction is represented by a square red factor node which connects to the nodes that are involved in the respective pairwise interaction. In the present example the network model has two pairwise interactions, each of which is now represented by a factor nodes. The label 2 indicates that the interaction is pairwise.





**Figure 5.2:** (a) The typical network-visualization of a network model with pairwise interactions between variables 1-2 and 2-3; (b) the network model in (a) visualized as a factor graph. Edges are now represented by factor nodes (red square nodes) of order 2; edges now indicate which variables are connected to which factor (c) the factor graph visualization of the previous network model with an additional 3-way interaction (blue triangle node).

The factor graph visualization allows us to also include interactions involving three variables (3-way interactions): in panel (c) we add the 3-way interaction 1-2-3 to the pairwise model and visualize it in a factor graph. Again we visualize the two pairwise interactions 1-2 and 2-3 as separate factor nodes. Similarly, we visualize the 3-way interaction as a factor node that connects to the three variables involved in the 3-way interaction. Again, the label 3 indicates that the interaction is a 3-way interaction. We can interpret this 3-way interaction in terms of moderation in three different ways: (i) the moderation effect of 3 on the pairwise interaction 1-2, (ii) the moderation effect of 1 on the pairwise interaction 2-3, and (iii) the moderation effect of 2 on the pairwise interaction 1-3. In Section 5.2.1 we called (i) and (ii) *partially* moderated because there is a pairwise interaction independent of the value of the moderator variable. And we called (iii) *fully* moderated, because the pairwise interaction is fully determined by the value of the moderator variable.

In Figure 5.2 we did not display the value of the two pairwise and the 3-way interaction. In a factor graph this information can either be displayed on the factor nodes or on the edges connecting them to variables. Since researchers are already familiar with displaying the weight of parameters as the width of edges, we chose the latter option. We show such weighted factor graphs in empirical data examples in Section 5.4.

### 5.2.5 Estimation via $\ell_1$ -regularized Nodewise Regression

In this section we show how to estimate the  $p$  conditional distributions of the MNM with  $\ell_1$ -regularized nodewise regression and how to combine the estimates

to the joint MNM. Our approach is similar to the nodewise regression approach for estimating the multivariate Gaussian distribution (Meinshausen & Bühlmann, 2006), except that we estimate the conditionals of the MNM instead of the joint Gaussian distribution.

### 5.2.5.1 Estimate Nodewise Regressions

To estimate the  $p$  regression models, we minimize the squared loss plus the  $\ell_1$ -norm of the parameter vectors  $\beta_{s,\cdot}$  and  $\omega_{s,\cdot}$ , for each variable  $s \in V$

$$\arg_{\beta_{s,\cdot}, \omega_{s,\cdot}} \min \left\{ \sum_{z=1}^n (X_{z,s} - \hat{\mu}_{z,s})^2 + \lambda_s (\|\beta_{s,\cdot}\|_1 + \|\omega_{s,\cdot}\|_1) \right\}, \quad (5.12)$$

where  $X_{z,s}$  is the value of variable  $s$  in row  $z$  of the data matrix,  $\hat{\mu}_{z,s}$  is the predicted mean for row  $z$  (see equation 5.10),  $\beta_{s,\cdot}$  and  $\omega_{s,\cdot}$  are vectors containing parameters associated with pairwise and 3-way interactions, respectively,  $\|\beta_{s,\cdot}\|_1 + \|\omega_{s,\cdot}\|_1 = \sum_{i \neq s} |\beta_{s,i}| + \sum_{i,j \in V, i \neq j \neq s} |\beta_{s,i,j}|$  is the sum of the  $\ell_1$ -norms of both parameter vectors, and  $\lambda_s$  is a tuning parameter that weights the  $\ell_1$ -norm relative to the squared loss. Note that  $\beta_{s,\cdot}$  contains only the pairwise interactions involving variable  $s$  and has therefore less elements than  $\beta$  in the joint distribution. Similarly,  $\omega_{s,\cdot}$  contains only the 3-way interactions involving variable  $s$  and has therefore less elements than  $\omega$  in the joint distribution.

Prior to estimation we mean-center each variable and divide each variable by its standard deviation. This ensures that the penalization of a given parameter does not depend on the standard deviation of its associated (product of) variables and simplifies the model because all intercepts are equal to zero. In addition, recall that mean-centering of variables is necessary to obtain interpretable parameter estimates (see Section B.1).

We chose  $\ell_1$ -regularized (LASSO) regression for three reasons: (1) the number of parameters can be large if  $p$  is large and many moderation effects are included in the model (see Section 5.2.3), which leads to high variance on the parameter estimates (overfitting). The  $\ell_1$ -regularization shrinks parameter estimates towards zero and thereby mitigates this problem; (2) The  $\ell_1$ -penalty sets small parameter estimates to zero, which simplifies the interpretation of the model, especially when interpreted from a network/graph perspective; and (3)  $\ell_1$ -penalization ensures that the model remains identifiable when the number of parameters is larger than the number of observations. When estimating an MNM with  $p = 20$  variables and includes all moderation effects, each nodewise regression has  $p - 1 + \binom{p-1}{2} = 190$  parameters, which means unregularized methods require  $n \geq 190$  observations. The  $\ell_1$ -penalization allows to estimate such a model also with  $n < 190$ .

The main assumption underlying  $\ell_1$ -regularized regression is that most of the parameters in the true model are equal to zero (also called *sparsity* assumption). This seems a reasonable assumption for the MNM in psychological data, since we would not expect that every variable moderates every pairwise interaction. For an excellent discussion of  $\ell_1$ -regularized regression see (Hastie et al., 2015).

In each of the  $p$  regressions, one has to select a tuning parameter  $\lambda_s$  which controls the strength of the penalization. If  $\lambda_s = 0$ , the loss function in (5.12) reduces to squared loss alone, which is the loss function of standard OLS regression. If  $\lambda_s$  is huge, all parameters are set to zero. To select an optimal  $\lambda_s$ , one can use a cross-validation scheme or an information criterion. Foygel and Drton (2010) showed that the Extended Bayesian Information Criterion (EBIC), a modification of the BIC Schwarz et al. (1978) that puts an extra penalty on nonzero parameters, performs well in estimating sparse parameter vectors.

### 5.2.5.2 Combine Estimates to Joint Moderated Network Model

The above estimation procedure leads to two estimates for every pairwise interaction, and three estimates for every 3-way interaction (moderation effect). For example, we obtain an estimate for the pairwise interaction between  $X_i$  and  $X_j$  from the regression of  $X_i$  on  $X_j$ , and another estimate from the regression of  $X_j$  on  $X_i$ . Similarly, we obtain three estimates for any given moderation effect from three regressions: the nodewise estimation procedure returns three estimates for each moderation parameter: (1)  $X_q$  moderating the predictor  $X_s$  in the regression on  $X_j$ , (2)  $X_s$  moderating the predictor  $X_j$  in the regression on  $X_q$ , and (3)  $X_q$  moderating the predictor  $X_j$  in the regression on  $X_s$ . In order to arrive at a single estimate to specify the joint MNM, we either take the arithmetic mean across the two/three values (OR-rule), or take the arithmetic mean across the two/three values if all three values are nonzero and otherwise set the aggregated parameter to zero (AND-rule). The AND-rule is more conservative than the OR-rule. It is even more conservative for 3-way interactions, because now three parameter estimates have to be nonzero to set the aggregate parameter to nonzero. For a more elaborate description of the nodewise regression procedure see (Haslbeck & Waldorp, 2020).

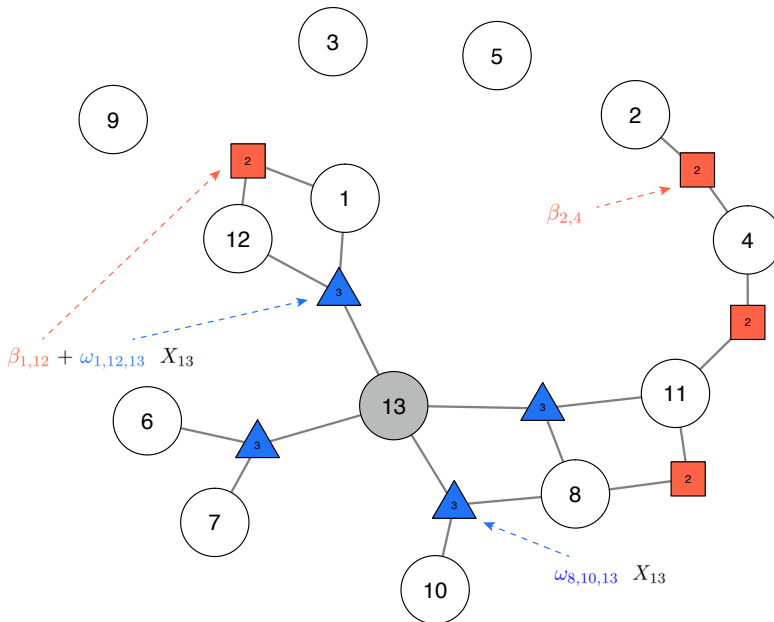
In the following section, we investigate the performance of the  $\ell_1$ -regularized nodewise regression approach in estimating MNMs, and compare its performance in detecting moderation effects to the split sample methods NCT and FGL.

## 5.3 Simulation Study

The goal of this simulation is (a) to investigate the performance of  $\ell_1$ -regularized nodewise regression in estimating moderated network models and (b) compare its performance to detect moderation effects to the split-sample methods Network Comparison Test (NCT) and Fused Graphical Lasso (FGL). Note that since the NCT and FGL can only provide a piecewise constant approximation of the linear moderation effects in MNMs, we expect that they will perform worse than MNMs. However, because the methods differ in several additional characteristics, and to determine the exact performance differences, we map out the differences of NCT, FGL and MNMs. We first describe the data generation (5.3.1) and the estimation procedures (5.3.2). Finally, we report performance results (5.3.3) and discuss them in Section 5.3.4.

### 5.3.1 Data generation

We sample observations from 100 Moderated Network Models that are specified by the following procedure: we begin with an empty graph with  $p = 12$  nodes and randomly add six edges. Of these six edges, two are unmoderated pairwise interactions (e.g. edge 2-4 in Figure 5.3; or panel (b) in Figure 5.1), two are fully moderated pairwise interactions (e.g., 6-7 in Figure 5.3; or panel (d) in Figure 5.1), and two are partially moderated pairwise interactions (e.g. edge 12-1 in Figure 5.3; or panel (d) in Figure 5.1). We enforce that each node has at most 2 edges by resampling the graph until this constraint is met. We do this because sampling from highly connected nodes leads to many rejections in the rejection sampler, which makes sampling unfeasible. After obtaining the graph with six edges, we add an additional variable, which serves as the moderator<sup>2</sup>. The final graph therefore has 13 nodes. We repeat this procedure for 100 iterations, yielding 100 data generating models in the simulation. Figure 5.3 shows the model resulting from this procedure in iteration 2.



**Figure 5.3:** The factor graph used in iteration 2 of the simulation. Circle nodes indicate variables. Square nodes with label 2 indicate 2-way interactions. Triangle nodes with label 3 indicate moderation effects (3-way interactions). Each of the three types of pairwise interactions (unmoderated, partially moderated, fully moderated) appears twice. For one of each of them the formula for the total effect is shown. Node 13 is the moderator variable.

<sup>2</sup>We do not allow the initial six edges to be connected to the moderator, because otherwise for some graph configurations (1) moderation effects can turn into quadratic effects and (2) unmoderated pairwise interactions can turn into moderated pairwise interactions.

We visualize the moderated network model using a factor graph (see Section 5.2.4). There is an unmoderated pairwise interaction between 2-4 and 4-11; a partially moderated pairwise interaction between 1-12 and 8-11; and a fully moderated pairwise interaction between 6-7 and 8-10. All moderated interactions are moderated by variable 13. The factor graph visualization highlights the equivalence between moderation effects and higher order interactions: a pairwise interaction between  $A$  and  $B$  that is partially moderated by  $C$  means that  $A$  and  $B$  are also connected to  $C$  in a 3-way interaction. Take the pairwise interaction 1-12 as an example: the total pairwise interaction is equal to the constant  $\beta_{1,12}$  plus variable  $X_{13}$  weighted by the moderation effect  $\omega_{1,12,13}$ . The presence of a 3-way interaction alone can be seen as full moderation: for example 8-10 are only connected via the 3-way interaction with 13. This means the parameter for the pairwise interaction 8-10, is only a function of  $X_{13}$  weighted by the moderation effect  $\omega_{6,7,13}$ . In an unmoderated pairwise interaction (for example 2-4) the two variables are not involved in the same 3-way interaction. In this case, the total parameter for the pairwise interaction is a constant.

We constructed the joint distribution by factoring  $p$  conditional Gaussians (see Section 5.2.3). For each of the  $p$  conditional Gaussians, we set the standard deviation to one and the intercept to zero. To be able to compare the performance in recovering pairwise and 3-way interactions, we set the value of all nonzero parameters  $\beta_{i,j}$  and  $\omega_{i,j,q}$  to 0.2. The total interaction parameters ( $\beta_{i,j} + \omega_{i,j,q}X_q$ ) can be interpreted as (moderated) partial correlations. To sample cases from the joint distribution, we use a Gibbs sampler on the  $p$  conditional Gaussians with means defined in equation (5.10). As discussed in Section 5.2.3, the constraints of the parameter space under which the joint distribution is normalizable are unknown. We work around this problem by using a rejection sampler (for details see Appendix B.3). With this sampling procedure we obtain  $n = 1808$  cases from each of the 100 MNMs.

To investigate performance as a function of the number of observations  $n$  we create 12 variations on a log scale  $n \in \{30, 46, \dots, 1148, 1808\}$ . This range of  $n$ -values was chosen because it allows us to show the performance transition from detecting no parameters at all to perfectly recovering the model as a function of  $n$ . We always use the first 30, 46, ... observations, which means that in scenario  $n = 46$  we take the samples of the scenario  $n = 30$  and add the next 16. This approach minimizes differences in performance across  $n$ -variations due to sampling variation.

### 5.3.2 Estimation

Here we describe the three different methods for detecting moderation effects that we compare in the simulation study: Moderated Network Models (5.3.2.1), the Network Comparison Test (NCT) (5.3.2.2), and the Fused Graphical Lasso (FGL) (5.3.2.3).

### 5.3.2.1 Moderated Network Models

To investigate the performance of MNMs in recovering moderation effects we estimate MNMs in three different versions: in version (1) we know the true moderator variable (variable 13) and specify only that variable as a moderator. If we do not know the true moderator, one can use two different strategies: in version (2) we estimate  $p$  moderated network models, in each of which we specify a different variable as the moderator. After estimating the  $p$  models, all estimates are combined such that if a parameter was estimated nonzero in at least one of the  $p$  models it is considered to be present in the combined model; note that the sensitivity to discover pairwise interactions for (2) will be larger or equal compared to (1) because we combine 13 estimates for each pairwise interaction. The same is true for moderation effects, but we expect the difference to be smaller, because only 3 estimates are combined (in our model the moderation effect of  $C$  on the interaction between  $B, A$  is the same as the moderation effects of  $B$  on the interaction between  $A, C$  and of  $A$  on the interaction between  $C, B$ ). The reason to include this version of the algorithm is to see how much the precision drops when using such an exploratory approach. The second strategy for the situation in which the true moderators are unknown is version (3), in which we include all moderators at once.

In the nodewise regression algorithm used to estimate all three versions of the moderated network models, we select the tuning parameters  $\lambda_s$  that minimizes the Extended Bayesian Information Criterion (EBIC), which has been shown to perform well in recovering sparse graphs (Foygel & Drton, 2010). The EBIC is an extension of the BIC (Schwarz et al., 1978) in that it puts an additional penalty on the number of nonzero parameters. This additional penalty is weighted by a parameter  $\gamma$ . We set  $\gamma = 0.5$  because this value led to good performance in simulations using a setting similar to ours (Epskamp, 2016).

### 5.3.2.2 Network Comparison Test (NCT)

The Network Comparison Test (NCT) performs a permutation test for each edge parameter to determine whether it is reliably different across two groups (data sets). Since the NCT makes a comparison between two groups it requires to split the dataset in half. Here we split at the median of the moderator variable. In a first step, the NCT estimates a model on each data set and takes the absolute value the differences between the corresponding parameters. These differences serve as test statistics. In a second step, a sampling distribution under the null hypothesis (no difference) is created for each edge comparison by  $B$  times randomly permuting the group membership of data points, estimating the two models and computing the absolute value of all edge differences. This gives sampling distributions for each edge-parameter difference, which can be used to test the significance of the edge-difference from step 1 under the null hypothesis that there is no difference. For estimation, the NCT uses the graphical lasso algorithm (J. Friedman, Hastie, & Tibshirani, 2008a) and selects the regularization parameter  $\lambda$  with using the EBIC with hyperparameter  $\gamma = 0.5$ . We used  $B = 1000$  and set the significance threshold to  $\alpha = 0.05$ . In our simulation, for small sample

sizes  $n = 30, 46$  the covariance matrices for the two groups (each computed from  $n = \frac{30}{2}, \frac{46}{2}$ ) were not positive definite for some of the  $B$  bootstrap samples. To still be able to run the NCT algorithm, we modified the original algorithm by (van Borkulo et al., 2016) in that we project these covariance matrices to the nearest positive definite covariance matrix<sup>3</sup>.

### 5.3.2.3 The Fused Group Lasso (FGL)

The Fused Group Lasso (FGL) (Danaher et al., 2014; Costantini et al., 2017) jointly estimates two GGMs by using two separate  $\ell_1$ -penalties: the first penalty (weighted by  $\lambda_1$ ) includes all parameters of the model (two covariance matrices), which is the standard  $\ell_1$ -penalty (e.g., Hastie et al., 2015). This penalty is similar to the one we use to estimate the moderated network models; the second penalty (weighted by  $\lambda_2$ ) includes the *difference* of the two covariance matrices, and therefore penalizes parameter differences across groups. For a detailed description of the FGL see (Danaher et al., 2014). Following the implementation in the R-package *EstimateGroupNetwork* (Costantini & Epskamp, 2017), we first select  $\lambda_1$  and then  $\lambda_2$ , using the EBIC with  $\gamma = 0.5$ . We do not perform a full grid-search on  $\lambda_1, \lambda_2$  since this is computationally very expensive (Costantini et al., 2017). Because the FGL jointly estimates two GGMs on two data sets, also here we median-split the data set along the moderator variable.

For both the NCT and the FGL methods we run two versions: (1) we create two groups by splitting the dataset at the median value of the moderator variable and then run the NCT/FGL; (2) we create the grouping for each of the  $p$  variables in the data set one by one and compute the NCT/FGL for each of those groupings. And then take the union of detected moderation effects as output. Since no more group differences can be discovered in the additional  $p - 1 = 12$  runs, the sensitivity cannot improve. Again, we include this condition to investigate the drop of precision when using such an exploratory approach. The estimates of NCT/FGL can be seen as a piecewise constant approximation to the linear moderation effect, with two constant functions on the left/right of the median split on the moderator variable.

We estimate moderated network models using the implementation in the R-package *mgm* (Haslbeck & Waldorp, 2020). The NCT is implemented in the R-package *NetworkComparisonTest* (van Borkulo, 2016) and the FGL is implemented in the R-package *EstimateGroupNetwork* (Costantini & Epskamp, 2017). All three packages are open source and freely available on the Comprehensive R Archive Network (CRAN) (<https://cran.r-project.org/>).

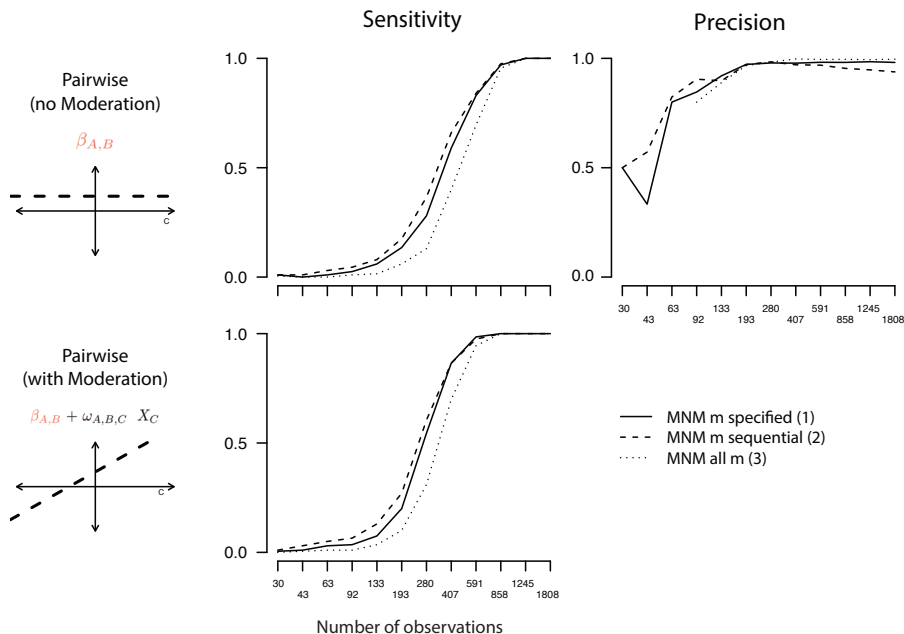
## 5.3.3 Results

We report *sensitivity* (probability of recovering a true parameter) and *precision* (probability that an estimated parameter is a true parameter). For the moderated network models recovering a true (nonzero) parameter means estimating a

<sup>3</sup>We use the implementation of the algorithm of Higham (2002) in the R-package *Matrix* (Bates & Maechler, 2017).

nonzero parameter with positive sign. For NCT and FGL, recovering a moderation effect means that the group difference of a given pairwise interaction is significant (NCT) or nonzero (FGL) and that the parameter estimated on the data set with larger values on the moderator variable is larger (that is, the difference has the correct sign). In the FGL version (2), which runs over all  $\{1, 2, \dots, p\}$  possible moderators, it can happen that a given edge difference is detected with different signs in several runs. We then select the difference with the largest absolute value.

Figure 5.4 shows the average sensitivity and precision of pairwise interactions and the pairwise part in a partially moderated pairwise interaction over 100 iterations. Precision is defined with respect to *all* pairwise interactions (moderated or not) and is therefore only displayed once. Note that the sensitivity for small  $n$  is very low. This implies that precision is undefined in many iterations. We display precision only if precision is defined in at least 5 iterations. We report the performance for the three estimation versions of the moderated network model.



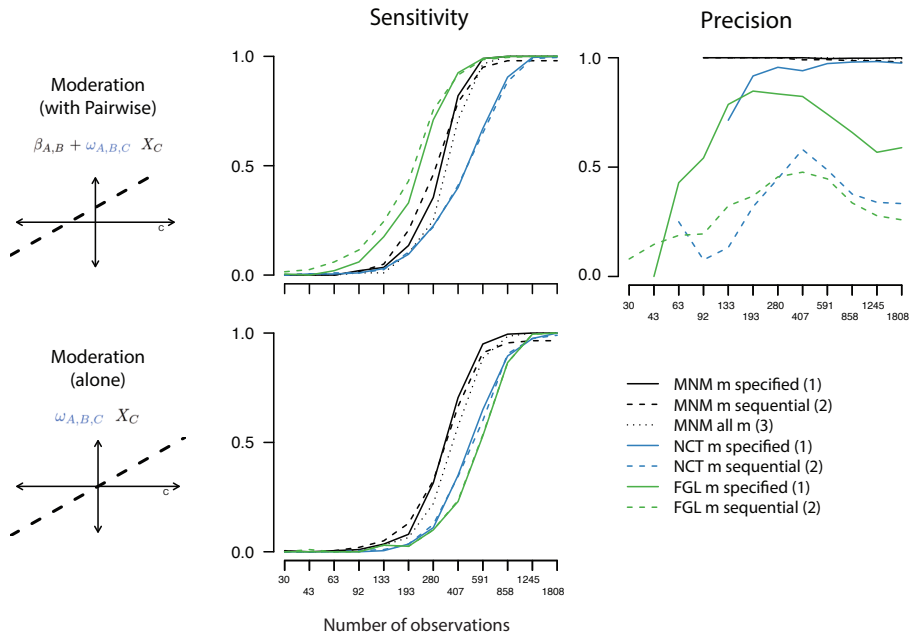
**Figure 5.4:** Sensitivity and precision for the moderated network model estimators for pairwise interactions without moderation (first row) and pairwise interactions with moderation (second row), as a function of  $n$  on a log scale. Precision is defined with respect to all pairwise estimates and is therefore the same for pairwise parameters with/without moderation.

We first turn to the performance of the  $\ell_1$ -regularized nodewise regression in estimating the pairwise parameters in the MNM. The sensitivity of all three versions of the moderated network model seems to converge to 1 when increasing  $n$ . The different versions of the network model stack up as expected: the exploratory sequential version (2) has the highest sensitivity, since combining the standard



version with specified moderator (1) with the estimates of  $p-1$  additional models can only increase the sensitivity. The network model with all possible moderator effects specified at once has the lowest sensitivity. This makes sense, since it has a much larger number of parameters and hence larger regularization parameters  $\lambda_s$  to control the variance of the estimates. Consequently, precision stacks up in reverse order. The precision for versions (1) and (3) seems to converge to 1, while the precision of (2) does not.

Figure 5.5 shows that all of the results described in the previous paragraph also hold for moderation effects. The performance to recover pairwise and moderation effects is similar. The largest difference in performance between parameter types is between the sensitivity to detect the unmoderated pairwise interaction (row 1 in Figure 5.4) and the full moderation (row 2 in Figure 5.5) is *smaller* compared to the sensitivity to detect the pairwise and moderation effects in the partially moderated pairwise interaction (row 2 in Figure 5.4 and row 1 in Figure 5.5, respectively).



**Figure 5.5:** Sensitivity and precision for the moderated network model estimators and the NCT and FGL estimators for the moderation effects in a partially moderated pairwise interaction (first row) and moderation effects in a fully moderated pairwise interaction (second row), as a function of  $n$  on a log scale.

Turning to the NCT, Figure 5.5 shows that its sensitivity seems to converge to 1, but does so slower than all other methods for partial moderation (first row), and comparable to the FGL for full moderation effects (second row). Precision grows slower than for MNMs and is only close to 1 for  $n \geq 591$  observations. The version of the NCT that searches for all  $p$  possible moderators sequentially cannot

improve sensitivity, because there is nothing to detect if an incorrect moderator is specified. The precision of the sequential NCT is low and does not converge to 1 as  $n$  increases.

The FGL shows the highest sensitivity in recovering moderation effects in a pairwise interaction (row 1) and shows the lowest sensitivity (comparable to NCT) for full moderation (row 2). Within FGL, the sequential version shows higher sensitivity while above we claimed that the sensitivity cannot be larger than in the specified version. The precision of the FGL with specified moderator increases up to  $n = 193$  and decreases for larger  $n$ . The sequential FGL has low precision for all  $n$ .

In Appendix B.11 we also provide the results shown in Figures 5.4 and 5.5 in tables.

### 5.3.4 Discussion of Simulation Results

The goal of this simulation was (a) to investigate the performance of  $\ell_1$ -regularized nodewise regression for estimating moderated network models and (b) compare its performance to detect moderation effects to the split-sample methods NCT and FGL.

#### 5.3.4.1 Performance of Moderated Network Models

The MNMs with correctly specified moderator (version 1) and with all moderators specified at once (version 3) are consistent estimators for moderated network models in the setting of our simulation, since both their sensitivity and precision seem to converge to 1 as  $n$  increases. Version 2, which combines results of  $p$  sequential moderated network models showed similar performance, but does not converge to 1 for the sequence of  $n$  investigated in our simulation. The reason is that false-positives accumulate across the  $p$  models. When considering larger  $n$  we would expect that also version 2 converges.

The second important finding is that pairwise interactions and moderation effects (3-way interactions) are roughly equally difficult to estimate. From a  $\ell_1$ -regularized nodewise (LASSO) regression perspective this is what we would expect: moderation effects are just additional predictors that are uncorrelated with the respective main effects (we show that in Appendix B.7). For estimation, moderation effects are therefore in no relevant way different from main effects and hence a different performance in estimating them would be surprising. The fact that moderation effects are just additional predictors in a regularized regression means that we can make use of the large pool of theoretical and simulation results on the performance of  $\ell_1$ -regularized regression in different situations (e.g., Hastie et al., 2015; Bühlmann & Van De Geer, 2011). Theoretical results of the LASSO for nodewise regression require the assumption of sparsity. Graph-sparsity does not apply to moderated network models since their parameters cannot be represented in a  $p \times p$  graph (see Section 5.2.4).

Based on anecdotal evidence we expect that in reality moderation effects are on average smaller than pairwise interactions. If this is true, it will be harder to

recover any moderation effects than recovering any pairwise effects in the trivial sense that smaller effects are harder to recover than larger effects. In the simulation study we kept the size of pairwise interactions and moderation effects equal to investigate the presence of any unexpected effects. This was important so as to verify whether we can use theory on multiple regression to make predictions about the performance of moderated network models in different situations (we can).

The largest performance difference across the four parameter types was between the sensitivity of unmoderated pairwise interaction and full moderation on the one hand, and the pairwise and moderation effects in the partially moderated interaction on the other hand. This difference is explained by the fact that in the latter cases the number of uncorrelated predictors is larger, which leads to a small increase in sensitivity. In Appendix B.4 we provide a figure that directly displays this difference. In Appendices B.5 and B.6 we show with additional simulations that the above explanation is correct. Specifically, we show that this phenomenon is also present in unmoderated network models and thus not specific to MNMs.

In the present simulation we specified the MNM to include only a single moderator variable. We chose to include a single moderator variable to ensure that the simulation setup is easy to understand and to keep the sampling procedure feasible. However, the performance of the estimators with all moderators specified, or all moderators specified sequentially will not change much when including an additional moderator. The reason is that all moderators are already included in the model. Including one additional parameter therefore does not change the estimated model, it merely means that one additional parameter is nonzero in the true MNM. Since we estimate the MNM nodewise, and because higher- and lower-order terms are uncorrelated (see Appendix B.7), adding a moderator in the present situation is similar to fitting a fixed linear regression model and setting one parameter in the true model from zero to nonzero.

#### 5.3.4.2 Performance of NCT and FGL

Sensitivity and precision of the NCT in which we specified the correct moderator seem to converge to 1 as a function of  $n$ . However, both sensitivity and precision are lower than for the moderated network model in all situations included in our simulation. The sequential version of the NCT has low precision for all  $n$  and should therefore not be used in situations similar to the one used in our simulation.

Turning to the FGL version in which we also specified the correct moderator, its sensitivity is comparable to the NCT for full moderation and largest across all methods for moderation in partially moderated interactions. The explanation for this difference in sensitivity is a combination of two factors: first, the EBIC selects models with higher sensitivity if the number of uncorrelated neighbors is larger (we show this in Appendix B.6). In the moderation with pairwise effect (row 1, Figure 5.5) there is an additional predictor (neighbor) compared to the moderation effect alone (row 2). Second, the graphical lasso is more liberal

than nodewise regression. The second factor combined with the first one explains why the sensitivity difference is larger for the FGL compared to the moderated network models. However, since the FGL is not the focus of the present work we did not further investigate this difference. The difference in sensitivity between the standard and the sequential version of the FGL is explained by the fact that by design 4-8 variables are correlated with the moderator 13 (see Figure 5.3). Therefore, splitting by these variables in the sequential version is similar to splitting along the moderator proportional to the correlation between the moderator and the splitting variable. The precision of the FGL with specified moderator increases up to  $n = 193$  and decreases for larger  $n$ . The sequential FGL has low precision for all  $n$ . A partial explanation for this low precision of the FGL could be the low precision of the graphical lasso algorithm for large  $n$  if its assumption of sparsity is violated (Williams & Rast, 2018). Because of its low precision, we do not recommend the FGL for situations similar to the one in our simulation.

So far, we only discussed the version of the NCT and FGL in which the true moderator variable was known and provided to the algorithms. However, this may not be the case in many realistic situations. In version (2) of NCT and FGL we ran the algorithms  $p$  times, each time specifying another variable as the moderator and combining the results (for details see above). Figure 5.5 shows that the sensitivity of this approach is (as expected) a bit higher, however, precision is very low.

### 5.3.4.3 Moderated Network Models vs. Sample-split Methods

For the situations considered in our simulation study, we found that moderated network models are consistent estimators for moderation effects and outperform the split-sample based methods NCT and FGL. We generated the data from a MNM and therefore our model is correctly specified, while the sample-split methods NCT/FGL only approximate the linear moderation effect with a piecewise constant function. We therefore expected that our model would perform better. If the moderator is a Bernoulli random variable and NCT/FGL do not have the disadvantage of only approximating the moderation effect, the performance difference may be smaller. However, this comparison would require moderated MGMs, which we leave for future work (see also Discussion in Section 5.5). Another important *general* advantage of the moderated network approach is that it has much larger sensitivity to detect pairwise interactions, because the method does not require to split the sample in half.

## 5.4 Empirical Data Examples

In this section, we apply Moderated Network Models to empirical data. Specifically, we provide a fully reproducible tutorial on how to fit MNMs to a data set of mood variables using the R-package *mgm* (Haslbeck & Waldorp, 2020) (Section 6.4). We also present different options for visualizing MNMs using factor graphs (see Section 5.2.4). On the basis of the same data set we then discuss possible aspects of model misspecification and tools to detect those (Section 5.4.2).

### 5.4.1 R-Tutorial: Fit Moderated Network Model to Data Set of Mood Variables

We show how to fit a MNM to a cross-sectional data set consisting of  $n = 3896$  observations of the five mood variables *hostile*, *lonely*, *nervous*, *sleepy* and *depressed* with 5 response categories. This data set is a subset of the data set `msq` from the R-package *psych* (Revelle, 2017). To fit the MNM, we use the R-package *mgm*, which implements functions to estimate  $k$ -order Mixed Graphical Models (MGMs), of which GGMs and MNMs are special cases. The package can be installed and loaded in the following way:

```
install.packages("mgm")
library(mgm)
```

In the following two subsections we first show how to fit the MNM to the data and then present possible visualizations of the model.

#### 5.4.1.1 Fit Moderated Network Model to Data

The data set containing the five trait mean scores is automatically loaded with *mgm* and available as the object `msq_p5`:

```
> dim(msq_p5)
[1] 3896  5
> head(msq_p5)
hostile  lonely  nervous  sleepy  depressed
1 -0.4879522  0.7280507  1.0018084 -0.2334325 -0.5998642
2 -0.4879522 -0.6442210 -0.5445646 -0.2334325 -0.5998642
3 -0.4879522 -0.6442210 -0.5445646 -1.1857391 -0.5998642
4 -0.4879522  0.7280507  2.5481814 -0.2334325  0.8672236
5 -0.4879522 -0.6442210 -0.5445646 -0.2334325 -0.5998642
6 -0.4879522 -0.6442210 -0.5445646  0.7188742  0.8672236
```

`dim(msq_p5)` shows that the dataset consists of 3896 rows and 5 columns and `head(msq_p5)` displays the first 6 rows of the data set. The data points have several points after the decimal because each variable was scaled to mean zero and a standard deviation of one.

We provide the data in `msq_p5` to the estimation function `mgm()` of the *mgm* package. Next to the data, we specify the types and levels for each variable. Since we model all variables as Gaussian distributions, we specify "g" for each variable and the number of levels as 1 by convention for continuous variables. This specification is necessary in *mgm*, because the package also allows to model Poisson variables and categorical variables with  $m$  categories. Via the argument `moderator` one specifies the moderators to be included in the model. For instance, if we select `moderators = c(1, 3)` all moderation effects of variables 1 and 3 are included in the model. Here we do not have any prior theory about possible moderators and therefore specify all variables as moderators by setting `moderators = 1:5`. This corresponds to version 3 of the estimator for MNMs in the simulation study (Section 6.3).

The estimation algorithm uses  $p = 5$  nodewise penalized regressions, for each of which an appropriate regularization parameter  $\lambda_s$  has to be selected (see Section 5.2.5). We select the  $\lambda_s$  that minimizes the EBIC with the hyperparameter  $\gamma = 0.5$  by setting `lambdaSel = "EBIC"` and `lambdaGam = .5`. This is the same setup we used in the simulation study. Alternatively one could select  $\lambda_s$  using cross-validation (`lambdaSel = "CV"`). With `scale = TRUE` we specify that all predictors are scaled to mean zero and standard deviation 1. This is a standard procedure in regularized regression and avoids that the penalization of a given parameter depends on the standard deviation of the associated variable (see Section 5.2.5). With `ruleReg = "AND"` we specify that the nodewise regressions are combined with the AND-rule (see Section 5.2.5).

```
mgm_mod <- mgm(data = msq_p5,
               type = rep("g", 5),
               level = rep(1, 5),
               lambdaSel = "EBIC",
               lambdaGam = .5,
               ruleReg = "AND",
               moderators = 1:5,
               scale = TRUE)
```

The main output is stored in `mgm_mod$interactions`. For a detailed description of the output see the help file `?mgm` and the *mgm* paper (Haslbeck & Waldorp, 2020). The list entry `mgm_mod$interactions$indicator` contains a list of all estimated parameters separately for each order (2-way, 3-way, etc.):

```
> mgm_mod$interactions$indicator
[[1]]
[,1] [,2]
[1,]  1   3
[2,]  1   4
[3,]  1   5
[4,]  2   3
[5,]  2   4
[6,]  2   5
[7,]  3   4
[8,]  3   5
[9,]  4   5

[[2]]
[,1] [,2] [,3]
[1,]  1   2   3
[2,]  1   2   4
[3,]  1   3   4
[4,]  3   4   5
```

The first level of this list shows that there are nine pairwise interactions; and the second entry shows that there are four moderation effects (or 3-way interactions). Specifically, the entry `mgm_mod$interactions$indicator[[1]][6, ]`

indicates that there is a nonzero pairwise interaction between variables 2-5. And the entry `mgm_mod$interactions$indicator[[2]][4, ]` indicates that there is a nonzero moderation effect between variables 3-4-5. To obtain more information about a given interaction we use the function `showInteraction()`. One can obtain the parameter for the pairwise interaction 2-5 (*lonely* and *depressed*) via:

```
> showInteraction(object = mgm_mod, int = c(2,5))
Interaction: 2-5
Weight: 0.4318148
Sign: 1 (Positive)
```

This pairwise interaction can be interpreted as in a linear regression: when increasing *lonely* by one unit, *depressed* increases by  $\approx 0.432$  units, when keeping all other variables constant. The parameters for the moderation effect can be obtained similarly: to obtain the moderation effect between *nervous*, *sleepy* and *depressed* we provide the respective column numbers to the `int` argument:

```
> showInteraction(object = mgm_mod, int = c(3,4,5))
Interaction: 3-4-5
Weight: 0.0564465
Sign: 1 (Positive)
```

We can interpret this moderation effect in three different ways: First, the pairwise interaction between *nervous* and *sleepy* is equal to zero when *depressed* is equal to zero (no pairwise interaction between *nervous* and *sleepy*) and increases by  $\approx 0.06$  when increasing *depressed* by one unit. For the other two interpretations we need the parameters of the pairwise interactions between *depressed* and *sleepy*, and between *depressed* and *nervous*:

```
> showInteraction(object = mgm_mod, int = c(4,5))
Interaction: 4-5
Weight: 0.1534387
Sign: 1 (Positive)>
> showInteraction(object = mgm_mod, int = c(3,5))
Interaction: 3-5
Weight: 0.1029161
Sign: 1 (Positive)
```

The second interpretation is: the pairwise interaction between *depressed* and *sleepy* is equal to  $\approx 0.153$  when *nervous* is equal to zero, and increases by  $\approx 0.06$  when increasing *nervous* by one unit. Similarly, the third interpretation is that the pairwise interaction between *depressed* and *nervous* is equal to  $\approx 0.103$ , when *sleepy* is equal to zero, and increases by  $\approx 0.06$  when increasing *sleepy* by one unit. For example, if *sleepy* has the value 2, then the pairwise interaction parameter between *depressed* and *nervous* is equal to  $\approx 0.103 + 2 \times 0.06 = 0.223$

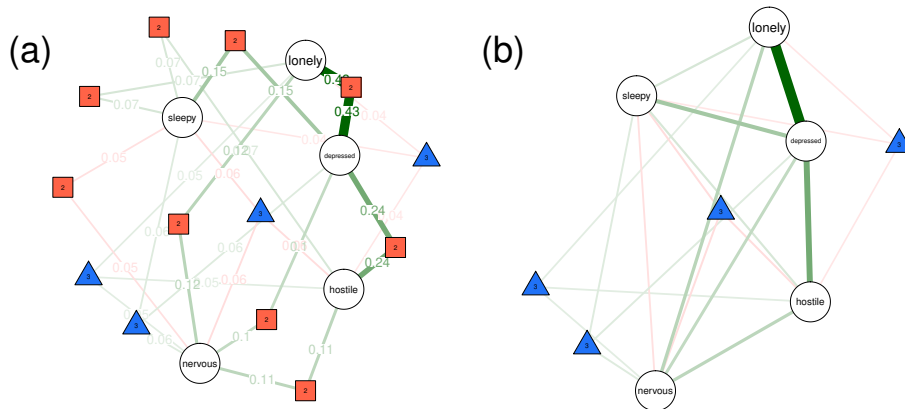
#### 5.4.1.2 Visualize Moderated Network Model as Factor Graph

In many situations it is more convenient to inspect the model parameters graphically. Since the MNM contains more than  $p \times p$  parameters, they cannot be visualized in a standard network with  $p \times p$  edges. Instead, we use a factor graph,

in which we introduce additional nodes for interactions (for details see Section 5.2.4). The function `FactorGraph()` from the *mgm* package draws such factor graphs from the output objects of `mgm()`:

```
FactorGraph(object = mgm_mod,
            edge.labels = TRUE,
            labels = colnames(msq_p5))
```

With `edge.labels = TRUE` we specified that the values of parameters are shown in the visualization. Figure 5.6 (a) shows the resulting plot:



**Figure 5.6:** Two different factor graph visualizations: (a) variable-nodes are displayed as circle nodes, pairwise interactions are displayed as square nodes, and moderation effects (3-way interactions) are displayed as triangles; (b) Only moderation effects (3-way interactions) are displayed as triangle nodes, pairwise interactions are displayed as simple edges. Green edges indicate parameters with positive sign. The widths of edges is proportional to the absolute value of the parameter.

The green (red) edges indicate parameters with positive (negative) sign, and the width of edges is proportional to the absolute value of the parameter. If two variables are connected by a pairwise interaction but not by a 3-way interaction, the pairwise interaction parameter is a partial correlation. If two variables are connected by a pairwise interaction *and* a 3-way interaction, the pairwise interaction parameter is the partial correlation between the two variables if the third variable (in the 3-way interaction) is equal to zero (see Section 5.2.1). The function `FactorGraph()` is a wrapper around the `qgraph()` function from the *qgraph* package (Epskamp et al., 2012) and all `qgraph()` arguments can be passed to customize the visualization.

For models that include many pairwise interactions this visualization may become unclear. To alleviate this problem, the factor nodes representing pairwise interactions can be replaced by simple edges with `PairwiseAsEdge = TRUE`. In addition, we specified with `edge.labels = FALSE` that the parameter values are not shown. The resulting visualization is shown in Figure 5.6 (b). While this



graph is not a traditional factor graph anymore, the visualization contains the same information as the visualization in (a), except the parameter values.

An alternative way to visualize MNMs is to condition on a set of values of the nonzero moderators and visualize the resulting pairwise network. This can be done with the function `condition()` which takes the model object `mgm_mod` and a list assigning a value to each moderator as input. The function outputs a new, conditioned, pairwise model object that can be visualized as a network. Repeating this process for a number of values of the nonzero moderators allows to show the pairwise network as a function of the nonzero moderators. This is especially useful if there is only a single moderator in the model. In the case of a large number of moderators, this approach becomes unfeasible, because the number of values to map out the space of moderator variables (and therefore the number of networks to plot) becomes grows exponentially with the number of moderators.

In Appendix B.8 we provide an additional tutorial in which we recover the MNM used in iteration 2 of the simulation study in Section 6.3.

## 5.4.2 Model Misspecification

Like for any other statistical model, when fitting MNMs to empirical data we assume that these data were generated by the class of MNMs, that is, we assume that the model is correctly specified. Specifically, we constructed the MNM from  $p$  conditional Gaussian distributions and therefore assume that each variable is conditionally Gaussian. In addition, we assume that the mean of each variables is modeled by a regression equation of the form of equation (5.10).

### 5.4.2.1 Types of Model Misspecification

The MNM is estimated via  $p$  moderated multiple regression models and therefore the possible types of model misspecification are the same as in multiple regression with moderation/interactions. The first type of misspecification in these models is the presence of non-linear effects (in pairwise or/and moderation effects); the second type is the presence of conditional distributions that are not Gaussian distributed. Both types of problems are well documented in the regression literature (e.g. (Aiken et al., 1991; Afshartous & Preston, 2011; Hainmueller, Mummolo, & Xu, 2018)), which is why we do not discuss them here in detail.

Instead, we focus on a new type of misspecification that arises from constructing the MNM joint distribution from the  $p$  conditional Gaussian distributions. Specifically, our construction of the MNW implies that the population moderation effect of  $A$  on the pairwise interaction  $B, C$  is the same as the moderation effect of  $B$  on  $A, C$  and  $C$  on  $A, B$  (and the three parameters converge empirically as  $n \rightarrow \infty$ ). It is this equality that justifies aggregating these three parameter estimates using the AND- or OR-rule (see Section 5.2.5). If all variables are generated from a joint distribution that can be factorized into conditional Gaussians, the moderation effects are the same across conditional distributions (nodewise regressions). However, if the data are skewed, it is possible that the moderation

effects are *different* across nodewise regressions and summarizing them in a single parameter would be misleading. This does not mean that the moderation effects always differ across nodewise regressions if the data is skewed. For example, the data used in the previous section were skewed but the moderation effects were roughly the same across nodewise regressions. In the following subsection we give an empirical data example with skewed variables in which moderation effects *are* different across nodewise regressions.

#### 5.4.2.2 Different Moderation Effects across Nodewise Regressions

We illustrate the problem of largely varying moderation effects across nodewise regressions using a data set consisting of  $n = 3896$  observations of the mood variables *afraid*, *ashamed* and *distressed*. These data are also from the data set `msq` from the R-package `psych` (Revelle, 2017) and are loaded with the `mgm` package as the object `msq_p3`. We request the dimensions and the first six rows of the data set with the following code:

```
> dim(msq_p3)
[1] 3896  3
> head(msq_p3)
afraid  ashamed distressed
1  2.1854037 -0.280891  0.8504921
2 -0.2853835 -0.280891 -0.5565216
3 -0.2853835 -0.280891 -0.5565216
4 -0.2853835  2.067502  0.8504921
5 -0.2853835 -0.280891 -0.5565216
6 -0.2853835 -0.280891 -0.5565216
```

The histograms in Figure 5.7 show that all three variables are highly skewed:

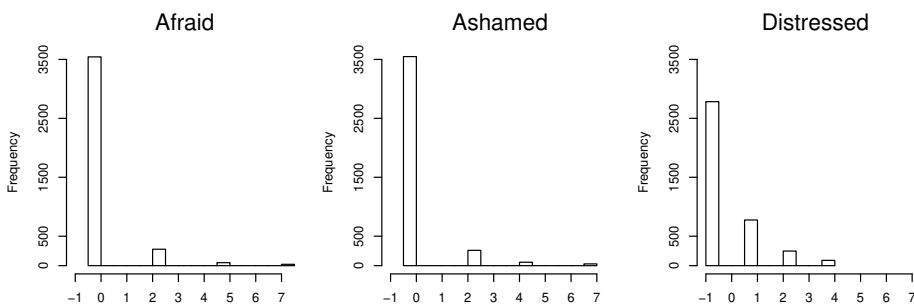


Figure 5.7: Histograms for the scaled variables *afraid*, *ashamed* and *distressed*.

In Appendix B.9, we show conditional scatter plots to provide some intuition for how it is possible that moderation effects differ across nodewise regressions. Here we proceed by estimating a moderated network model with the same specifications as above in Section 6.4:

```

mgm_mod2 <- mgm(data = msq_p3,
  type = rep("g", 3),
  level = rep(1, 3),
  lambdaSel = "EBIC",
  lambdaGam = .5,
  ruleReg = "AND",
  moderators = 1:3)

```

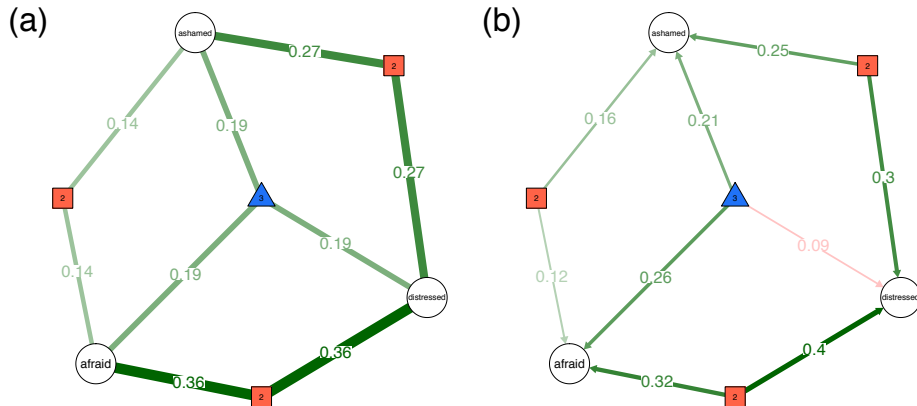
Again similarly to above we use the function `FactorGraph()` to plot the factor graph visualization of the moderated network model

```

FactorGraph(object = mgm_mod2,
  edge.labels = TRUE)

```

which is shown in Figure 5.8 (a):



**Figure 5.8:** (a) Factor graph visualization of the estimated network model with aggregated parameter values; (b) Nodewise factor graph visualization of all estimated nodewise (unaggregated) parameters. The direction of edges indicates the nodewise regression in which the parameter has been estimated.

Panel (a) in Figure 5.8 shows that there is a moderation effect with value  $\approx 0.19$ . We can interpret this moderation effect in the following way: the pairwise interaction between *distressed* and *ashamed* is equal to  $\approx 0.27$  when *afraid* is equal to zero, and increases by  $\approx 0.19$  when increasing *afraid* by one unit. However, we now show that this interpretation is inappropriate, because the nodewise estimates differ widely and the aggregate parameter is therefore misleading.

The nodewise parameter estimates can be accessed via the `mgm()` output object (see `?mgm`). A more convenient way to inspect the unaggregated nodewise estimates is to plot them into a directed version of the factor graph, in which the direction of the edges indicates the nodewise regression in which the parameter has been estimated. This modified Factor graph can be plotted by setting `Nodewise = TRUE` in the `FactorGraph()` function:

```
FactorGraph(object = mgm_mod2,  
            edge.labels = TRUE,  
            Nodewise = TRUE)
```

Figure 5.8 (b) shows the resulting visualization. A directed edge towards a given node always indicates a parameter obtained from the regression on that node. For example, the directed edge towards *afraid* from the order-2 factor node that connects *afraid* with *distressed* indicates the pairwise interaction between *afraid* and *distressed* obtained from the nodewise regression on *afraid*. Importantly, the directionality of the edges shown in Figure 5.8 (b) only stems from the nodewise regression algorithm and does not represent an actual directionality of the effect.

We see that the moderation effect on *distressed* is actually negative. The correct interpretation would therefore be that the pairwise interaction between *distressed* and *ashamed* is equal to  $\approx 0.27$  when *afraid* is equal to zero, and *decreases* by  $\approx 0.09$  when increasing *afraid* by one unit. This is a moderation effect in the opposite direction of the aggregate moderation effect we used in the interpretation above.

What should one do in such a situation? There is no easy answer. Clearly, interpreting the parameters of the moderated network model joint distribution shown in (a) is misleading and therefore no option. A solution would be to reject the joint distribution in (a) and instead report the combined conditional distributions in (b). This has the downside that the model is more complex and that the joint distribution is unknown. The latter is undesirable, because this means that we cannot perform inference on the joint distribution. On the other hand, in many applications of network analysis in psychology no such inference is performed. The principled solution would be to create a joint distribution that incorporates skewed distributions and moderation effects that vary across conditional distributions. However, we expect this to be difficult and far beyond the scope of the present chapter.

## 5.5 Discussion

We introduced Moderated Network Models by using the standard definition of moderation in the regression framework and adding moderation effects to the multivariate Gaussian distribution. We presented a new visualization for Moderated Network Models based on factor graphs and we proposed an  $\ell_1$ -regularized nodewise regression procedure to estimate this model. In a simulation study we reported the performance of this approach to recover different types of parameters in a random graph with moderation and showed that estimating a moderated network model outperforms the split-sample based methods Fused Graphical Lasso (FGL) and the Network Comparison Test (NCT). Finally, we provided a fully reproducible tutorial on how to estimate MNMs with the R-package *mgm* and discuss possible issues with model misspecification.

Three limitations are important to keep in mind. First, as discussed in Section 5.2.3, we do not have an explicit constraint on the parameter space of the MNM joint distribution that ensures that the distribution is normalizable. In order to

sample observations, we worked around this issue with a rejection sampler (see Appendix B.3). But for a given model estimated from data, it is unclear whether the resulting joint distribution is normalizable. This means that while the joint distribution does capture the dependency structure in the data, it might not be possible to define a probability distribution over it. A consequence of the absence of a guaranteed probability distribution is that one cannot use global likelihood ratio tests or goodness of fit analyses to select between models. We expect that an appropriate constraint on the parameter space to be difficult to work out since it involves all parameters of the model and the variances of all conditional Gaussians. While this is an important limitation to keep in mind, all conditional distributions *are* consistently estimated proper distributions. Thus, if inferences are limited to the conditional distribution there is no issue. This implies that predictions for any variable can be computed without any limitation, which allows to perform model selection using out of sample prediction error. Also, it is important to keep in mind that the joint distribution correctly captures the dependency structure. This implies that one can use network statistics such as centrality metrics, modularity or global efficiency to describe the global network structure.

Second, the performance results obtained from our simulation study may be different in other setups. The best way to obtain the performance in setups with larger/smaller parameters or higher/lower sparsity is to run appropriate additional simulation studies. That said, recall that we estimate the MNM using a series of multiple regression models with interaction terms. Since interaction effects can be seen as additional variables in a multiple regression, the performance in recovering is in principle the same as for main effects. This means that one can draw on the rich literature on these models to make predictions about the performance in recovering MNMs in different setups (e.g., Hastie et al., 2015; Bühlmann & Van De Geer, 2011).

Third, in our simulation study we assigned the same size to pairwise interactions and moderation effects. We did this to confirm our prediction from the linear regression framework that pairwise and moderation effects are equally hard to estimate. But in reality moderation effects are often much smaller than pairwise interactions. This means that moderation effects are on average harder to detect than pairwise interactions in the same sense as small pairwise interactions are harder to detect than large pairwise interactions. It is possible to make the estimation procedure more liberal by reducing the hyperparameter  $\gamma$  in the EBIC to detect small moderation effects. However, this would result in higher false positive rates for both pairwise and moderation effects. In sum, moderation effects tend to be smaller than pairwise interaction and are therefore harder to detect by *any* estimation algorithm. However, MNMs will recover moderation effects that have a size comparable to pairwise interactions. And any moderation effect can be recovered if enough data is available.

We see the following extensions for future research. First, in this chapter we suggested a  $\ell_1$ -regularized nodewise regression approach to estimate moderated network models. We chose this estimator because it deals well with the large number of parameters and renders the interpretation of the model easier

by setting small parameters to zero. In addition, the underlying assumption that most parameters are zero is reasonable in the case of MNMs (see Section 5.2.5). But in some situations a different estimator might perform better. One example is the presence of correlated predictors in skewed data. In this situation, the lower-order terms will be correlated with the higher-order terms in the regression models (see Section B.9 for an illustration of that fact). If this correlation becomes too large, the  $\ell_1$ -regularized estimator selects only one of the two terms. This problem is larger, if  $n$  is small and the selected regularization parameter  $\lambda$  large. It would be interesting to map out how problematic this is and whether other estimators, for example based on significance testing, are better suited for such situations.

Second, the MNM we propose includes all pairwise interactions and a set of specified 3-way interactions/moderation effects. Since we freely estimate all specified parameters, it is possible that a moderation effect is estimated to be nonzero, while none of the pairwise interactions between the involved variables is estimated to be nonzero. In some situations, the model may be more meaningful when enforcing a *structural hierarchy* that only allows to estimate moderation effects to be nonzero if all respective pairwise interactions are estimated nonzero (Bien, Taylor, & Tibshirani, 2013). However, such a choice reduces the fit of the model and should be based on substantive grounds.

Third, we use a single regularization parameter  $\lambda_s$  in each of the  $p$  nodewise regressions. The sparsity assumption here is with respect to the entire parameter vector, which includes both pairwise interactions and moderation effects (3-way interactions). However, if one assumes a different level of sparsity for pairwise interactions and moderation effects, separate regularization parameters may be more appropriate. The downside of two regularization parameters is that any model selection procedure needs to search a grid of  $\lambda$ s instead of a sequence, which increases the computational cost. In addition, estimates may become more unstable, because the considered model space is larger.

Fourth, it would be useful to extend the present work to moderated Mixed Graphical Models (MGMs) (S. Chen et al., 2014; Yang et al., 2014a). Here we have shown how to extend the multivariate Gaussian distribution by adding linear moderation effects for one or several moderator variables. This approach of adding terms for moderation effects to the nodewise regression equations can in principle be extended to the more general class of MGMs, in which each variable in the model is a univariate member of the exponential family conditioned on all other variables (S. Chen et al., 2014; Yang et al., 2014a). The only difference is that we perform the nodewise regressions in the GLM framework using the appropriate link-functions (see e.g., Nelder & Baker, 1972). However, depending on which types of variables are involved in an interaction, both pairwise interactions and moderation effects are captured by *sets* of varying numbers of parameters. It would be useful to give detailed treatment of all possible (moderated) interaction types between different types of variables, a description of how to interpret them and provide performance results for estimating different types of moderation effects in different situations. A special case of moderated MGMs would be an MGM with a single categorical variable as a moderator. This would allow to

investigate group differences across two or *several* groups in a principled way and is likely to out-perform group-split based methods on this task. Note that these analyses are implemented in the R-package *mgm* (Haslbeck & Waldorp, 2020) which we used in the application section of this chapter. However, a full treatment of moderated MGMs is beyond the scope of the present chapter.

Fifth, it could be useful to extend the notion of centrality to factor graphs and obtain a *moderator centrality*. A naive way of doing that would be to simply add up the moderation effect of a given variable to find out which variable has the strongest influence on the pairwise interactions in the network model. But one could come up with more sophisticated measures that take the structure of the network into account. This would especially be interesting in data sets that include contextual variables, because it allows to identify which of them have the strongest influence on a network model of psychological variables or symptoms.

In sum, Moderated Network Models relax the assumption of Gaussian Graphical Models that each pairwise interaction is independent of the value of all other variables by allowing that each pairwise interaction is *moderated* by (potentially) all other variables. This allows more precise statements about the sign and value of a given interaction parameters in a given situation, which may reduce the presence of seemingly contradictory research outcomes and provides a step towards more accurate models for subgroups and individuals.

## Acknowledgements

We would like to thank George Aalbers, Fabian Dablander, Peter Edelsbrunner, Sacha Epskamp, Eiko Fried, Oisín Ryan for helpful comments on earlier versions of this chapter.





# TIME-VARYING VAR MODELS

---

## Abstract

Time series of individual subjects have become a common data type in psychological research. These data allow one to estimate models of within-subject dynamics, and thereby avoid the notorious problem of making within-subjects inferences from between-subjects data, and naturally address heterogeneity between subjects. A popular model for these data is the Vector Autoregressive (VAR) model, in which each variable is predicted as a linear function of all variables at previous time points. A key assumption of this model is that its parameters are constant (or stationary) across time. However, in many areas of psychological research time-varying parameters are plausible or even the subject of study. In this tutorial paper, we introduce methods to estimate time-varying VAR models based on splines and kernel-smoothing with/without regularization. We use simulations to evaluate the relative performance of all methods in scenarios typical in applied research, and discuss their strengths and weaknesses. Finally, we provide a step-by-step tutorial showing how to apply the discussed methods to an openly available time series of mood-related measurements.

## 6.1 Introduction

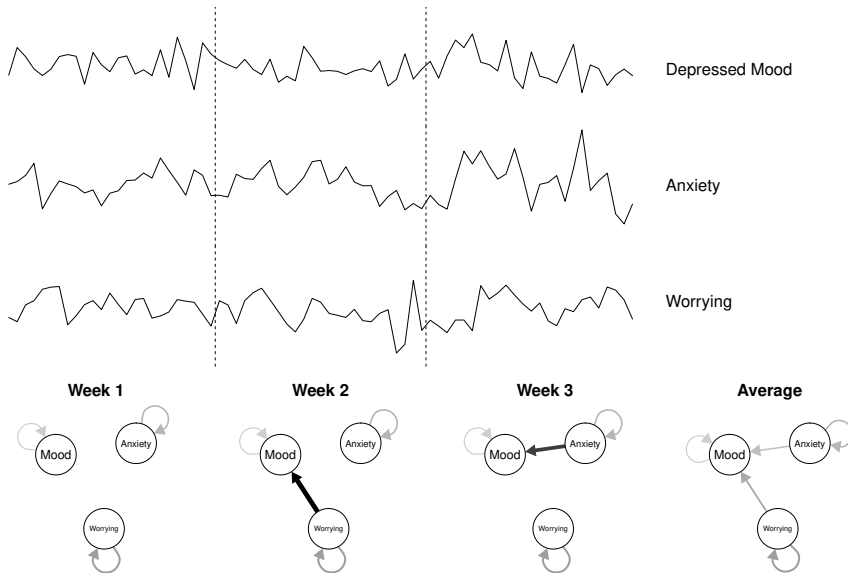
The ubiquity of mobile devices has led to a surge in time series (or intensive longitudinal) data sets from single individuals (e.g., Bringmann et al., 2013; Kramer et al., 2014; Hartmann et al., 2015; Kroeze et al., 2016; van der Krieke et al., 2017; Bak, Drukker, Hasmi, & van Os, 2016; Snippe et al., 2017; Fisher, Reeves, Lawyer, Medaglia, & Rubel, 2017; Groen et al., 2019). This is an exciting development because these data allow one to model within-subject dynamics, which avoids the notorious problem of making within-subjects inferences from between-subjects data, and naturally addresses heterogeneity between subjects (Fisher, Medaglia, & Jeronimus, 2018; Molenaar, 2004). The ability to analyze within-subjects data therefore promises to be a major leap forward both for psychological research and applications in (clinical) practice.

A key assumption of all standard time series models is that all parameters of the data generating model are constant (or stationary) across the measured time period. This is called the *assumption of stationarity*<sup>1</sup>. While one often assumes constant parameters, changes of parameters over time are often plausible in psychological phenomena. As an example, take the repeated measurements of the variables *Depressed Mood*, *Anxiety* and *Worrying*, modeled by a time-varying first-order Vector Autoregressive (VAR) model shown in Figure 6.1. In week 1, there are no cross-lagged effects between any of the three variables. However, in week 2 we observe a cross-lagged effect from *Worrying* on *Mood*. A possible explanation could be a physical illness in week 2 that moderates the two cross-lagged effects. In week 3, we observe a cross-lagged effect from *Anxiety* on *Mood*. Again, this could be due to an unobserved moderator like a stressful period at work. The fourth visualization shows the average of the previous three models, which is the model one would obtain by estimating a stationary VAR model on the entire time series. In this situation, the stationary model is clearly inappropriate because it is different to the true model across *all* intervals of the time series.

Time-varying models are of central interest when studying psychological phenomena from a within-person perspective. For example, in the network approach to psychopathology, it is suggested that mental disorders arise from causal interactions among symptoms (see also Borsboom & Cramer, 2013; Schmittmann et al., 2013; Robinaugh, Hoekstra, & Borsboom, 2019). This means that the interactions between symptoms are different for healthy and unhealthy individuals (Pe et al., 2015; van Borkulo et al., 2015) and that the interactions in an individual change when she or he transitions from a healthy to an unhealthy state (or vice versa). Time-varying models are able to capture this change. Next to detecting these changes, they may also shed light on why those changes occurred. For example, one could correlate time-varying parameters with contextual factors such as elevated stress levels, social setting or major life events and thereby possibly uncover conditions and events that predict the onset of mental disorder.

---

<sup>1</sup>We use this definition of stationarity, because for VAR models with eigenvalues within the unit circle, which we focus on in this paper, it is equivalent to definitions based on the moments of distributions. This implies that we do not consider diverging VAR models (with eigenvalues outside the unit circle) which have a non-stationary distribution while its parameters are constant across time.



**Figure 6.1:** Upper panel: hypothetical repeated measurements of *Depressed Mood*, *Anxiety* and *Worrying*, generated from a time-varying lag 1 VAR model. Lower panel: the time-varying VAR-model generating the data shown in the upper panel. It consists of three models, one for each week. The fourth model (left to right) indicates the average of the three models, which is what one obtains when estimating a stationary VAR model on the entire time series.

ders. Time-varying models can also be used to study how parameters change in response to interventions. For example, in Section 6.4 we will fit a time-varying VAR model on ESM measurements during a double-blind medication reduction study (Wichers et al., 2016).

Time-varying models are also central to the idea of Early Warning Signals (EWS; Scheffer et al., 2009). For example, Wichers et al. (2016) suggested to anticipate phase-transitions between healthy and unhealthy states with EWS such as time-varying autocorrelation and variance (see also van de Leemput et al., 2014a). Time-varying VAR models are an extension of these EWS to multivariate time-series. Anticipating the sensitive periods around phase transitions is interesting, because during those periods treatment may be more efficient (Olthof et al., 2019). This means that time-varying models could be used as a tool to monitor patients and determine periods during which treatment is most promising.

In this tutorial paper we provide an introduction to how to estimate a time-varying version of the Vector Autoregressive (VAR) model, which is arguably the simplest multivariate time series model for temporal dependencies in continuous data, and is used in many of the papers cited above. We will focus on two sets of methods recently proposed by the authors to estimate such time-varying VAR models: Bringmann, Ferrer, Hamaker, Borsboom, and Tuerlinckx (2018) presented a method based on splines using the Generalized Additive Modeling (GAM) framework, which estimates time-varying parameters by modeling

them as a spline function of time; and Haslbeck and Waldorp (2020) suggested a method based on penalized kernel-smoothing (KS), which estimates time-varying parameters by combining the estimates of several local models spanning the entire time series. While both methods are available to applied researchers, it is unclear how well they and their variants (with/without regularization or significance testing) perform in situations that are typical in applied research. We aim to improve this situation by making the following contributions:

1. We report the performance of GAM based methods with and without significance testing, and the performance of KS based methods with and without regularization in situations that are typical for Experience Sampling Method (ESM) studies.
2. We discuss the strengths and weaknesses of all methods and provide practical recommendations for applied researchers
3. We compare time-varying methods to their corresponding stationary counterparts to address the question of how many observations are necessary to identify the time-varying nature of parameters.
4. We provide tutorials on how to estimate time-varying VAR models using both methods on an openly available intensive longitudinal dataset using the R-packages *mgm* and *tvvarGAM*.

The paper is structured as follows. In Section 6.2.1 we define time-varying VAR models, which are the focus of this paper. We next present two sets of methods to recover such models: one method based on splines with and without significance testing (Section 6.2.2), and one method based on kernel estimation with and without regularization (Section 6.2.3). In Sections 6.3.1 and 6.3.2 we report two simulation studies that investigate the performance of these two models and their stationary counterparts. In Section 6.4 we provide a fully reproducible tutorial on how to estimate a time-varying VAR model from an openly available time series data set collected in ESM studies using the kernel smoothing method using the R-package *mgm* (we repeat the same tutorial with the GAM method in the appendix). Finally, in Section 6.5 we discuss possible future directions for research on time-varying VAR models.

## 6.2 Estimating Time-Varying VAR Models

We first introduce the notation for the stationary first-order VAR model and its time-varying extension (Section 6.2.1) and then present the two methods for estimating time-varying VAR models: the GAM-based method (Section 6.2.2) and the penalized kernel-smoothing-based method (Section 6.2.3). We discuss implementations of related methods in Section 6.2.4.

### 6.2.1 Vector Autoregressive (VAR) Model

In the first-order Vector Autoregressive (VAR(1)) model, each variable at time point  $t$  is predicted by all variables (including itself) at time point  $t - 1$ . Next to a set of intercept parameters, the VAR(1) model is comprised by autoregressive effects, which indicate how much a variable is predicted by itself at the previous time point, and cross-lagged effects, which indicate how much a variable is predicted by all other variables at the previous time point.

Formally, the variables  $\mathbf{X}_t \in \mathbb{R}^p$  at time point  $t \in \mathbb{Z}$  are modeled as a linear combination of the same variables at  $t - 1$

$$\mathbf{X}_t = \boldsymbol{\beta}_0 + \mathbf{B}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon} = \begin{bmatrix} X_{t,1} \\ \vdots \\ X_{t,p} \end{bmatrix} = \begin{bmatrix} \beta_{0,1} \\ \vdots \\ \beta_{0,p} \end{bmatrix} + \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,p} \\ \vdots & \ddots & \vdots \\ \beta_{p,1} & \cdots & \beta_{p,p} \end{bmatrix} \begin{bmatrix} X_{t-1,1} \\ \vdots \\ X_{t-1,p} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}, \quad (6.1)$$

where  $\beta_{0,1}$  is the intercept of variable 1,  $\beta_{1,1}$  is the autoregressive effect of  $X_{t-1,1}$  on  $X_{t,1}$ , and  $\beta_{p,1}$  is the cross-lagged effect of  $X_{t-1,1}$  on  $X_{t,p}$ , and we assume that  $\boldsymbol{\varepsilon} = \{\epsilon_1, \dots, \epsilon_p\}$  are independent (across time points) samples drawn from a multivariate Gaussian distribution with variance-covariance matrix  $\Sigma$ . In this paper we do not model  $\Sigma$ , however, it can be obtained from the residuals of the model and used to estimate the inverse covariance matrix (see e.g., Epskamp, Waldorp, et al., 2018).

Throughout the paper we deal with first-order VAR models in which all variables at time point  $t$  are a linear function of all variables at time point  $t - 1$ . In the interest of brevity we will therefore refer to this first-order VAR model (or VAR(1) model) as a VAR model. More lags can be included by adding further parameter matrices and lagged variable vectors  $\mathbf{X}_{t-k}$  (for a lag of  $k$ ) to the model in (6.1). Note that while we focus on VAR(1) models in the this paper, the presented methods can be used to estimate time-varying VAR models with any set of lags. For a detailed description of VAR models we refer the reader to Hamilton (1994).

In both the GAM and the KS method we estimate (6.1) by predicting each of the variables  $X_{t,i}$  for  $i \in \{1, \dots, p\}$  separately. Specifically, we model

$$X_{t,i} = \beta_{0,i} + \boldsymbol{\beta}_i \mathbf{X}_{t-1} + \epsilon_i = \beta_{0,i} + \begin{bmatrix} \beta_{i,1} & \cdots & \beta_{i,p} \end{bmatrix} \begin{bmatrix} X_{t-1,1} \\ \vdots \\ X_{t-1,p} \end{bmatrix} + \epsilon_i, \quad (6.2)$$

for all  $i \in \{1, \dots, p\}$ , where  $\boldsymbol{\beta}_i$  is the  $1 \times p$  vector containing the lagged effects on  $X_{t,i}$ . After estimating the parameters in each equation, we combine all estimates to the VAR model in (6.1).

In order to turn the stationary VAR model in (6.1) into a time-varying VAR model, we introduce a time index for the parameter matrices

$$\mathbf{X}_t = \boldsymbol{\beta}_{0,t} + \mathbf{B}_t \mathbf{X}_{t-1} + \boldsymbol{\varepsilon}. \quad (6.3)$$

This allows a different parameterization of the VAR model at *each time point* and thereby allows the model to vary across time. Throughout this paper we

assume that the time-varying parameters are smooth deterministic functions of time. We define a smooth function as a function for which the first derivative exists everywhere. In the following two subsections we introduce two different ways to estimate such a time-varying VAR model.

The VAR model has often been discussed and visualized as a network model (Epskamp, Waldorp, et al., 2018), and also here we will use both statistical and network/graph terminology. To avoid confusion between the two terminologies, we explicitly state how the terms in the two terminologies correspond to each other. From the statistical perspective there are two types of lagged effects between pairs of variables: autocorrelations (e.g.,  $X_{t-1} \rightarrow X_t$ ) and cross-lagged effects (e.g.,  $X_{t-1} \rightarrow Y_t$ ). In the network terminology variables are nodes, and lagged effects are represented by directed edges. An edge from a given node on itself is also called a self-loop, and represents autocorrelation effects. The value of lagged effects is represented in sign and the absolute value of the edge-weights of the directed edges. If an edge-weight between variables  $X_t$  and  $Y_{t-1}$  is nonzero, we say that the edge from  $X_t$  and  $Y_{t-1}$  is present. *Sparsity* refers to how strongly connected a network is: if many edges are present, sparsity is low; if only few edges are present, sparsity is high. On a node-level, sparsity is captured by the *indegree* (how many edges point towards a node) and *outdegree* (how many edges point away from a node). In statistical terminology indegree is the number of incoming lagged effects on variable  $X$ , and outdegree the number outgoing lagged effects from variable  $X$ .

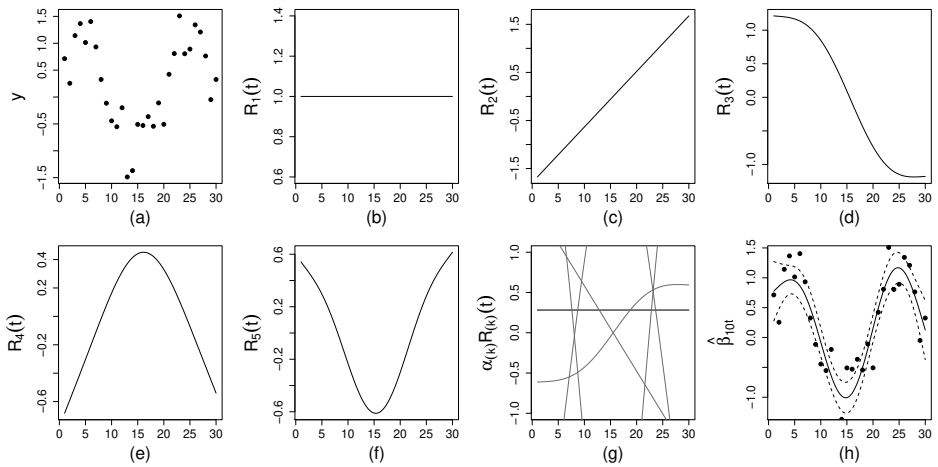
### 6.2.2 The GAM Method

In this section we explain how to estimate a time-varying VAR model using the Generalized Additive Model (GAM) framework, which allows for non-linear relationships between variables (see also Bringmann et al., 2017, 2018). We leverage the GAM framework for the estimation for time-varying models by using it to define each parameter as a function of time. Because GAMs are able to represent non-linear functions, this allows us to recover non-linear time-varying parameters. In what follows we illustrate how this approach works for the simplest possible example, a model consisting only of a time-varying intercept parameter,  $y = \beta_{0,t} + \varepsilon$ .

Panel (a) of Figure 6.2 shows that the values of  $y$  are varying over time, so the intercept will have to be time-varying as well, if the intercept-only model is supposed to fit the data well. This is achieved by summing the following five basis functions

$$\hat{\beta}_{0,t} = \hat{\alpha}_1 R_1(t) + \hat{\alpha}_2 R_2(t) + \hat{\alpha}_3 R_3(t) + \hat{\alpha}_4 R_4(t) + \hat{\alpha}_5 R_5(t), \quad (6.4)$$

which are displayed in panels (b) - (f) in Figure 6.2. Panel (g) overlays all used basis functions, and panel (h) displays the estimate of the final smooth function  $\hat{\beta}_{0,t}$ , which is obtained by adding up the weighted basis functions ( $\hat{\alpha}$ ) (see panel (g) and (h) of Figure 6.2). The optimal regression weights are estimated using standard linear regression techniques. The same rationale is applied to every time-varying parameter in the model.



**Figure 6.2:** An example of the basis function for a time-varying parameter  $\hat{\beta}_{0,t}$ . In panel (a) the data are shown. In panel (b)-(f) the estimated 5 basis functions are given and panel (g) shows the weighted basis functions. In the last panel (h) the final smooth function is illustrated with credible intervals around the smooth function.

There are several different spline bases such as cubic, P-splines, B-splines, and thin plate splines. The advantage of thin plate splines, which is the basis used here, is that one does not have to specify knot locations, resulting therefore in fewer subjective decisions that need to be made by the researcher (Wood, 2006). The basis functions in Figure 6.2 exemplify the thin plate spline basis. In the figure, panels (b)-(f) show that each additional basis function ( $R$ ) increases the nonlinearity of the final smooth function. This is reflected in the fact that every extra basis function is more “wiggly” than the previous basis functions. For example, the last basis function in panel (f) is “wigglier” than the first basis function in panel (b). The spline functions used here are smooth up to the second derivative. Thus, a key assumption of the GAM method is that all true time-varying parameter functions are smooth as well. This assumption is also called the assumption of *local stationarity*, because smoothness implies that the parameter values that are close in time are very similar, and therefore locally stationary. This would be violated by, for example, a step function, where the GAM method would provide incorrect estimates around a “jump” (but would still give good estimates for the two constant parts).

As the number of basis functions determines the nonlinearity of the smooth function (e.g.,  $\hat{\beta}_{0,t}$ ), a key problem is how to choose the optimal number of basis functions. The final curve should be flexible enough to be able to recover the true model, but not too flexible as this may lead to overfitting (Andersen, 2009; Keele, 2008). The method used here to find the optimal number of basis functions is penalized likelihood estimation (Wood, 2006). Instead of trying to select the optimal number of basis functions directly, one can simply start by including more basis functions than would be normally expected, and then adjust for too

much wiggleness with a *wiggleness penalty* (Wood, 2006).

Thus, the problem of selecting the right number of basis functions is reduced to selecting the right wiggleness penalty. This is achieved using generalized cross-validation (Golub, Heath, & Wahba, 1979), where the penalty parameter with the lowest Generalized Cross-Validation (GCV) value is expected to give a good bias-variance trade-off. Specifically, the penalization decreases the influence of the basis functions ( $R$ ) by reducing the values of their regression coefficients ( $\hat{\alpha}$ ). Therefore, smoothness is imposed on the curve both through the choice of the number of basis functions and the final level of penalization on these basis functions.

To estimate time-varying VAR models with the GAM method, we use the *tv-varGAM* package in *R* (Bringmann & Haslbeck, 2017), which is a wrapper around the *mgcv* package (Wood, 2006). As the wiggleness penalty is automatically determined, the user only needs to specify a large enough number of basis functions. The default settings are the thin plate regression spline basis and 10 basis functions, which although an arbitrary number, is often sufficient (see the simulation results in Bringmann et al., 2017). The minimum number is in most models three basis functions. In general, it is recommended to increase the number of basis functions if it is close to the effective degrees of freedom (edf) selected by the model. The effective degrees of freedom is a measure of nonlinearity. A linear function has an edf of one, and higher edf values indicate wigglier smooth functions (Shadish, Zuur, & Sullivan, 2014).

The GAM function in the *mgcv* package outputs the final smooth function, the GCV value and the edf. Furthermore, the uncertainty of the smooth function is estimated with 95% Bayesian credible intervals (Wood, 2006). In the remainder of this manuscript we refer to this method as the GAM method. We refer to a variant of the GAM method, in which we set those parameters to zero whose 95% Bayesian credible interval overlaps with zero, with `GAM(st)`, for “significance thresholded”. With GLM we refer to the standard unregularized VAR estimator.

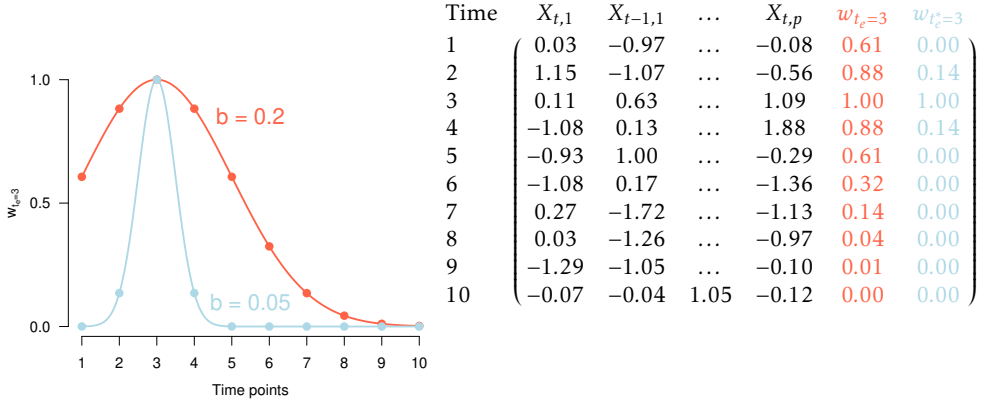
After the model is estimated, it is informative to check if the smooth functions were significantly different from zero (at some point over the whole time range), and if each smooth function had enough basis functions. Significance can be examined using the  $p$ -values of each specific smooth function, which indicates whether the smooth function is significantly different from zero. To see whether there are enough basis functions, the edf of each smooth function can be examined. The edf value should be well below the maximum possible edf or the number of basis functions for the smooth function (or term) of interest (in our case 10, see Wood, 2006). When the edf turns out to be too high, the model should be refitted with a larger (e.g., double) number of basis functions.

### 6.2.3 The Kernel-smoothing Method

In the kernel-smoothing method one obtains time-varying parameters by estimating and combining a sequence of local models at different time points across the time series. A local model is estimated by weighting all observations depending on how close they are to the time point at which the local model is estimated.



In Figure 6.3 we show an example in which a single local model is estimated at time point  $t_e = 3$ . We do this by giving the time points close to  $t_e$  a high weight and time points far away from  $t_e$  a very small or zero weight. If we estimate models like this on a sequence of equally spaced estimation points across the whole time series and take all estimates together, we obtain a time-varying model.



**Figure 6.3:** Illustration of the weights defined to estimate the model at time point  $t_e = 3$ . Left panel: a kernel function defines a weight for each time point in the time series. Right panel: the weights shown together with the VAR design matrix constructed to predict  $X_{t,1}$ .

Specifically, we use a Gaussian kernel  $\mathcal{N}(\mu = t_e, \sigma^2 = b^2)$  function to define a weight for each time point in the time series

$$w_{j,t_e} = \frac{1}{\sqrt{2\pi}b^2} \exp\left\{-\frac{(j-t_e)^2}{2b^2}\right\}, \quad (6.5)$$

where  $j \in \{1, 2, \dots, n\}$ , which is the local constant or Nadaraya-Watson estimator (Fan & Gijbels, 1996).

For the example shown in Figure 6.3 this means that the time point  $t_e = 3$  gets the highest weight, and if the distance to  $t_e$  increases, the weight becomes exponentially smaller. The same idea is represented in the data matrix in the right panel of Figure 6.3: each time point in the multivariate time series is associated with a weight defined by the kernel function. The smaller we choose the bandwidth  $b$  of the kernel function, the smaller the number of observations we combine in order to estimate the model at  $t_e$ : when using a kernel with bandwidth  $b = 0.2$  (red curve), we combine more observations than when using the kernel with  $b = 0.05$  (blue curve). The smaller the bandwidth the larger the sensitivity to detect changes in parameters over time. However, a small bandwidth means that less data is used and therefore the estimates are less reliable (e.g., only three time points when  $b = 0.05$ ; see right panel of Figure 6.3).

Since we combine observations close in time to be able to estimate a local model, we have to assume that the models close in time are also similar. This is equivalent to assuming that the true time-varying parameter functions are

smooth, or locally stationary. Thus, the key assumption of the kernel-smoothing approach is the same as in the spline approach. For the kernel-smoothing method, we need the additional assumption that the chosen bandwidth is small enough to capture the time-varying nature of the true model. For example, if the parameters of the true model vary widely over time, but the bandwidth is so large that at any estimation point almost the entire time series is used for estimation, it is impossible to recover the true time-varying function.

The weights  $w_{j,t_e}$  defined in (6.5) enter the loss function of the  $\ell_1$ -regularized regression problem we use to estimate each of the  $p$  time-varying versions of the model in (6.2)

$$\hat{\beta}_{t_e} = \arg_{\beta_{t_e}, \beta_{0,t_e}} \min \left\{ \frac{1}{n} \sum_{j=2}^n w_{j,t_e} (X_{i,j} - \beta_{0,t_e} - \beta_{t_e} X_{j-1})^2 + \lambda_i \|\beta_{t_e}\|_1 \right\}, \quad (6.6)$$

where  $X_{i,j}$  is the  $j^{\text{th}}$  time point of the  $i^{\text{th}}$  variable in the design matrix,  $\|\beta_{t_e}\|_1 = \sum_{i=1}^p \sqrt{\beta_{i,t_e}^2}$  is the  $\ell_1$ -norm of  $\beta_{t_e}$ , and  $\lambda_i$  is a parameter controlling the strength of the penalty. Note that the indices  $i$  and  $t_e$  are fixed in (6.6) because we estimate the time-varying VAR model equation by equation, separately for each estimation point  $t_e$ .

For each of the  $p$  regressions, we select the  $\lambda_i$  that minimizes the out-of-sample deviance in 10-fold cross validation (J. Friedman et al., 2010). In order to select an appropriate bandwidth  $b$ , we choose the  $\hat{b}$  that minimizes the out of sample deviance *across* the  $p$  regressions in a time stratified cross validation scheme (for details see Section 6.3.1.2). We choose a constant bandwidth for all regressions so we have a constant bandwidth for estimating the whole VAR model. Otherwise the sensitivity to detect time-varying parameters and the trade-off between false positives and false negatives differs between parameters, which is undesirable.

In  $\ell_1$ -penalized (LASSO) regression the squared loss is minimized together with the  $\ell_1$ -norm of the parameter vector. This leads to a trade-off between fitting the data (minimizing squared loss) and keeping the size of the fitted parameters small (minimizing  $\ell_1$ -norm). Minimizing both together leads to small estimates being set to exactly zero, which is convenient for interpretation. When using  $\ell_1$ -penalized regression, we assume that the true model is sparse, which means that only a small number of parameters  $k$  in the true model are nonzero. If this assumption is violated, the largest true parameters will still be present, but small true parameters will be incorrectly set to zero. However, if we keep the number of parameters constant and let  $n \rightarrow \infty$ ,  $\ell_1$ -regularized regression also recovers the true model if the true model is not sparse. For an excellent treatment on  $\ell_1$ -regularized regression see Hastie et al. (2015).

As noted above, the larger the bandwidth  $b$ , the more data is used to estimate the model around a particular estimation point. Indeed, the data used for estimation is proportional to the area under the kernel function or the sum of the weights  $N_{\text{util}} = \sum_{j=1}^n w_{j,t_e}$ . Notice that  $N_{\text{util}}$  is smaller at the beginning and end of

the time series than in the center, because the kernel function is truncated. This necessarily leads to a smaller sensitivity to detect effects at the beginning and the end of the time series. For a more detailed description of the kernel smoothing approach see Haslbeck and Waldorp (2020). In the remainder of this manuscript we refer to this method as KS(L1). With GLM(L1) we refer to the stationary  $\ell_1$ -penalized estimator.

### 6.2.4 Related methods

Several implementations of related models are available as free software packages. The R-package *earlywarnings* (Dakos & Lahti, 2013) implements the estimation of a time-varying AR model using a moving window approach. The R-package *MARSS* (E. Holmes, Ward, & Wills, 2013; E. E. Holmes, Ward, & Wills, 2012) implements the estimation of (time-varying) state-space models, of which the time-varying VAR model is a special case. While the state-space model framework is very powerful due to its generality, it requires the user to specify the way parameters are allowed to vary over time, for which often no prior theory exists in practice (Belsley & Kuti, 1973; Tarvainen, Hiltunen, Ranta-aho, & Karjalainen, 2004). In parallel efforts, Casas and Fernandez-Casal (2018) developed the R-package *tvReg*, which estimates time-varying AR and VAR models, as well as IRF, LM and SURE models, using kernel smoothing similar to the kernel smoothing approach described in the present paper, however does not offer  $\ell_1$ -regularization. Furthermore, the R-package *bvarsv* (Krueger, 2015) allows one to estimate time-varying VAR models in a Bayesian framework.

The R-package *dynr* (Ou, Hunter, & Chow, 2019) provides an implementation for estimating regime switching discrete time VAR models, and the R-package *tsDyn* (Fabio Di Narzo, Aznarte, & Stigler, 2009) allows to estimate the regime switching Threshold VAR model (Tong & Lim, 1980; Hamaker, Grasman, & Kamphuis, 2010). These two methods estimate time-varying models that switch between piece-wise constant regimes, which is different to the methods presented in this paper, which assume that parameters change smoothly over time.

Another interesting way to modeling time-varying parameters is by using the fused lasso (Hastie et al., 2015). However, to our best knowledge this method is currently only implemented for the estimation of Gaussian Graphical Models: R. Monti (2014) provide a Python implementation of the SINGLE algorithm (R. P. Monti et al., 2014), and (Gibbert, 2017) provide a Python implementation of the (group) fused-lasso based method as presented in Gibberd and Nelson (2017).

## 6.3 Evaluating Performance via Simulation

In this section we use two simulations to evaluate the performance of the methods introduced in Section 6.2 in estimating time-varying VAR models. In the first simulation (Section 6.3.1) we generate time-varying VAR models based on a random graph with fixed sparsity, which is the natural choice in the absence of any knowledge about the structure of VAR models in a given application. This

simulation allows us to get a rough overview of the performance of all methods and their strengths and weaknesses. In the second simulation (Section 6.3.2), we generate time-varying VAR models in which we vary the level of sparsity. This simulation allows us to discuss the strengths and weaknesses of all methods in more detail, specifically, we can discuss in which situations methods with/without regularization or thresholding perform better. Finally, in Section 6.3.3 we discuss the combined results of both simulations, and provide recommendations for applied researchers.

### 6.3.1 Simulation A: Random Graph

In this simulation we evaluate the performance of all methods in estimating time-varying VAR models that are generated based on a random graph. We first describe how we generate these time-varying VAR models (Section 6.3.1.1), discuss details about the estimation methods (Section 6.3.1.2), report the results (Section 6.3.1.3), and provide a preliminary discussion (Section 6.3.1.4).

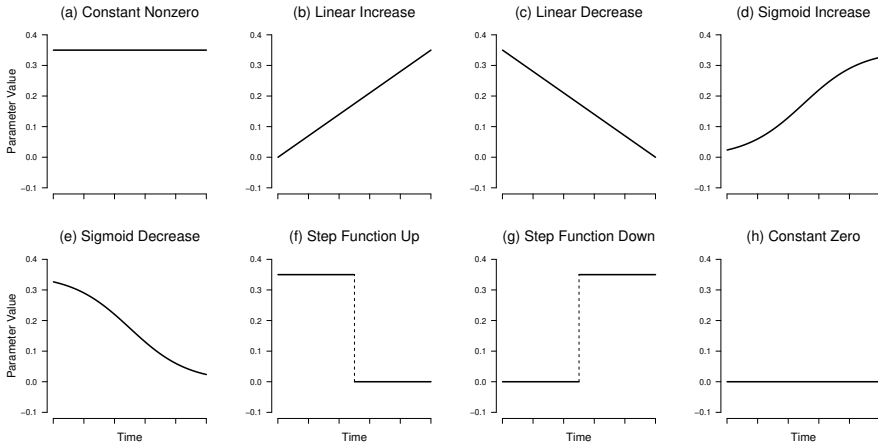
#### 6.3.1.1 Data generation

We generated time-varying VAR models by first selecting the structure of a stationary VAR model and then turning this stationary VAR model into a time-varying one. Specifically, we used the following procedure to specify whether a parameter in the time-varying VAR(1) model is nonzero: we choose all our VAR models to have  $p = 10$  variables, which is roughly the number of variables measured in typical ESM studies. We start out with an empty  $p \times p$  VAR parameter matrix. In this matrix we set all  $p$  autocorrelations to be nonzero, since autocorrelations are expected to be present for most phenomena and are observed in essentially any application (e.g., aan het Rot, Hogenelst, & Schoevers, 2012; Snippe et al., 2017; Wigman et al., 2015). Next, we randomly set 26 of the  $p \times p - p = 90$  off-diagonal elements (the cross-lagged effects) to be present. This corresponds to an edge probability of  $P(\text{edge}) \approx 0.29$ <sup>2</sup>. This approach returns an initial  $p \times p$  matrix with ones in the diagonal and zeros and ones in the off-diagonal.

In a second step we use the structure of this VAR model to generate a time-varying VAR model. Specifically, we randomly assign to each of the nonzero parameters one of the sequences (a) - (g) in Figure 6.4. If an edge is absent in the initial matrix, all entries of the parameter sequence are set to zero (panel (h) in Figure 6.4). Note that only the time-varying parameter functions (a - e) and (h) in Figure 6.4 are smooth functions of time. Therefore, the two methods presented in this paper are only consistent estimators for those types of time-varying parameters. They cannot be consistent estimators for the step-functions (f) and (g), however, we included them to investigate how closely the methods studied in this paper can approximate the step function as a function of  $n$ .

---

<sup>2</sup>We set a fixed number of elements to nonzero instead of using draws with  $P(\text{edge}) = 0.2$ , because we resample the VAR matrix until it represents a stable VAR model (the absolute value of all eigenvalues is smaller than 1). By fixing the number of nonzero elements we avoid biasing  $P(\text{edge})$  through this resampling process. Thus, all VAR matrices in each iteration and at each time point has no eigenvalue with absolute value greater than 1.



**Figure 6.4:** The eight types of time-varying parameters used in the simulation study: (a) constant nonzero, (b) linear increase, (c) linear decrease, (d) sigmoid increase, (e) sigmoid decrease, (f) step function up, (g) step function down and (h) constant zero.

The maximum parameter size of time-varying parameters is set to  $\theta = 0.35$  (see Figure 6.4). The noise is drawn from a multivariate Gaussian with variances  $\sigma^2 = \sqrt{0.10}$  and all covariances being equal to zero. Hence the signal/noise ratio used in our setup is  $S/N = \frac{0.35}{0.10} = 3.50$ . All intercepts are set to zero and the covariances between the noise processes assigned to each variable are zero.

Using these time-varying VAR model, we generate 12 independent time series with lengths  $n = \{20, 30, 36, 69, 103, 155, 234, 352, 530, 798, 1201, 1808\}$ . We chose these values because they cover the large majority of scenarios applied researchers typically encounter. Each of these time-varying models covers the full time period  $[0, 1]$  and is parameterized by a  $p \times p \times n$  parameter array  $B_{i,j,t}$ . For example, the  $B_{1,2,310}$  indicates the cross-lagged effect from variable 2 on variable 1 at the 310th measurement point, which occurs then at time point  $310/530 \approx 0.59$ , if there are in total 530 measurements. Importantly, in this setting increasing  $n$  does not mean that the time period between the first and the last measurement of the time series becomes larger. Instead, we mean by a larger  $n$  that more evenly spaced measurements are available in the same time period. This means that the larger  $n$ , the smaller the time interval between two adjacent measurements. That is, the data density in the measured time period increases with  $n$ , which is required to consistently estimate time-varying parameters (Robinson, 1989). This makes sense intuitively: if the goal is to estimate the time-varying parameters of an individual in January, then one needs sufficient measurements in January, and it does not help to add additional measurements from February.

We run 100 iterations of this design and report the mean absolute error over iterations. These mean errors serve as an approximation of the expected population level errors.

### 6.3.1.2 Estimation

To estimate time-varying VAR models via the GAM method we use the implementation in the R-package *tvvarGAM* (Bringmann & Haslbeck, 2017) version 0.1.0, which is a wrapper around the *mgcv* package (version 1.8-22). The tuning parameter of the spline method is the number of basis functions used in the GAM. Previous simulations have shown that 10 basis functions give good estimates of time-varying parameters (Bringmann et al., 2018). To ensure that the model is identified, for a given number of basis functions  $k$  and variables  $p$ , we require at least  $n_{\min} > k(p+1)$  observations. In our simulation, we used this constraint to select the maximum number of basis functions possible given  $n$  and  $p$ , but we do not use less than 3 or more than 10 basis functions. That is, the selected number of basis functions  $k_s$  is defined as

$$k_s = \max \left\{ 3, \min \left\{ \max \left\{ k; k > \frac{n}{p+1} \right\}, 10 \right\} \right\}. \quad (6.7)$$

If  $k_s$  satisfies the above constraint, the time-varying VAR model can be estimated with the spline-based method. With this constraint the model cannot be estimated for  $n = \{20, 30\}$ . We therefore do not report results for GAM and GAM(st) for these sample sizes.

In principle it would be possible to combine  $\ell_1$ -regularization with the GAM-method, similarly as in the KS-method. However, an implementation of such a method is currently not available and we therefore cannot include it in our simulation.

We estimated the time-varying VAR model via the KS and KS(L1) methods using the R-package *mgm* (Haslbeck & Waldorp, 2020) version 1.2-2. As discussed in Section 6.2.3, these methods require the specification of a bandwidth parameter. Therefore, the first step of applying these methods is to select an appropriate bandwidth parameter by searching the candidate sequence  $\mathbf{b} = \{0.01, 0.045, 0.08, 0.115, 0.185, 0.22, 0.225, 0.29, 0.325, 0.430, 0.465, 0.5\}$ . For  $n \leq 69$  we omit the first 5 values in  $\mathbf{b}$ , and for  $n > 69$  we omit the last 5 values. We did this to save computational cost because for small  $n$ , small  $b$  are never selected, and analogously for large  $n$ , large  $b$  values are never selected. To select an appropriate bandwidth parameter we use a cross-validation-like scheme, which repeatedly divides the time series in a training and a test set, and in each repetition fits time-varying VAR models using the bandwidths in  $\mathbf{b}$ , and evaluates the prediction error on the test set for each bandwidth. More concretely, we define a test set  $S_{\text{test}}$  by selecting  $|S_{\text{test}}| = \lceil (0.2n)^{2/3} \rceil$  time points stratified equally across the whole time series. Next, we estimate a time-varying VAR model for each variable  $p$  at each time point in  $S_{\text{test}}$  and predict the  $p$  values at that time point. After that we compute for each  $b$  the  $|S_{\text{test}}| \times p$  absolute prediction errors and take the arithmetic mean. Next, we select the bandwidth  $\hat{b}$  that minimizes this mean prediction error. Finally, we estimate the model on the full data using  $\hat{b}$  and  $\hat{\lambda}$  at 20 equally spaced time points, where we select an appropriate penalty parameter  $\hat{\lambda}_i$  with 10-fold cross-validation for each of the  $p$  variables (for more details see Haslbeck & Waldorp, 2020).

We also investigate the performance of the kernel-smoothing method without  $\ell_1$ -regularization. We refer to this method as KS. In order to compare the  $\ell_1$ -regularized time-varying VAR estimator to a stationary  $\ell_1$ -regularized VAR estimator, we also estimate the latter using the *mgm* package. We call this estimator GLM(L1).

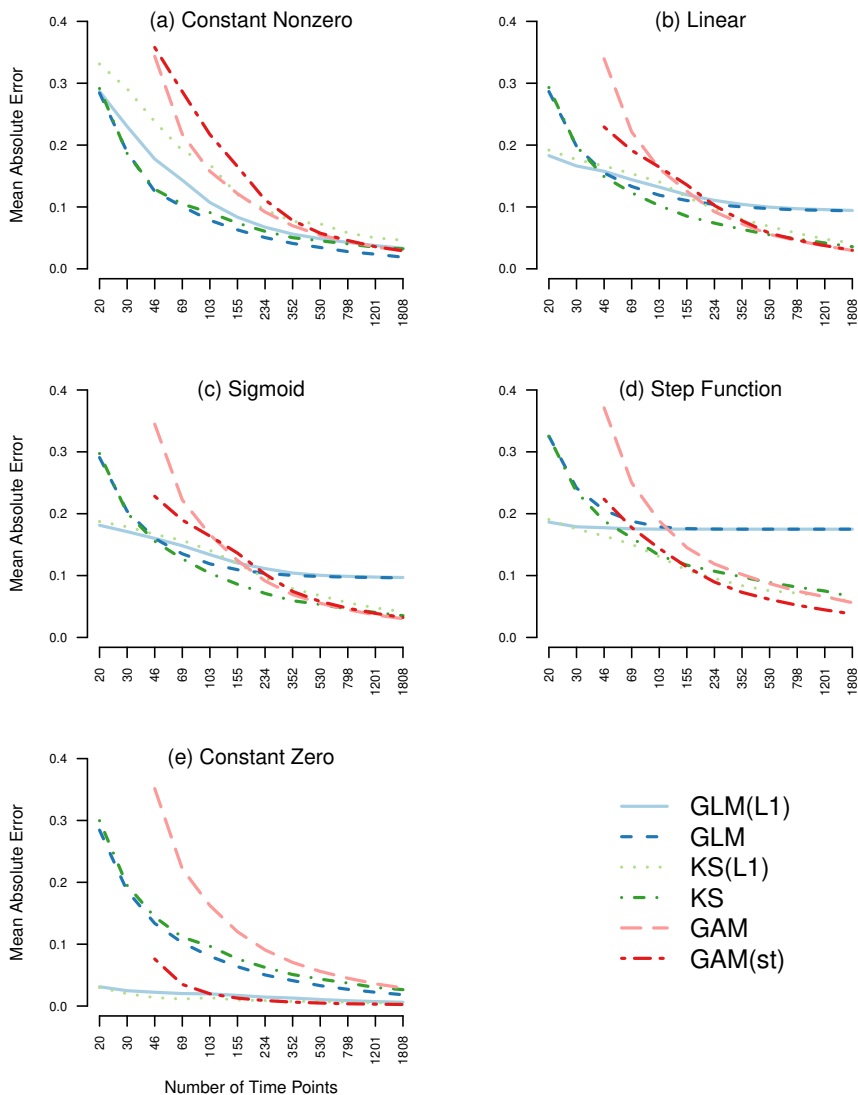
Both time-varying estimation methods are consistent if the following assumptions are met; (a) the data is generated by a time-varying VAR model as specified in equation (6.1), (b) all parameters are smooth functions of time, (c) with the eigenvalues of the VAR matrix being within the unit circle at all time points, (d) and the error covariance matrix is diagonal. We also fit a standard stationary VAR model using linear regression to get the unbiased stationary counter-part of the GAM methods. Specifically for the KS-method, it is additionally required that we consider small enough candidate bandwidth values. We do this by using the sequence  $\mathbf{b}$  specified above.

### 6.3.1.3 Results

We first report the performance of the GLM, GLM(L1), KS, KS(L1), GAM and GAM(st) methods in estimating different time-varying parameters by evaluating the estimation error *averaged across time*. Next, we zoom in on the performance *across time*, for the constant and the linear increasing parameter function, and finally examine the performance in structure recovery of all methods.

**Absolute Error Averaged over Time** Figure 6.5 displays the absolute estimation error, averaged over time points, iterations, and time-varying parameter functions of the same type, as a function of sample size  $n$ . Since the linear increase/decrease, sigmoid increase/decrease, and step function increase/decrease are symmetric, we collapsed them into single categories to report estimation error. The absolute error on the y-axis can be interpreted as follows: let's say we are in the scenario with  $n = 155$  observations and estimate the constant function in Figure 6.5 (a) with the stationary  $\ell_1$ -regularized regression GLM(L1). Then the *expected* average (across the time series) error of the constant function is  $\pm 0.09$ .

Figure 6.5 (a) shows that, for all methods, the absolute error in estimating the constant nonzero function is large for small  $n$  and seems to converge to zero as  $n$  increases. The GLM method has a lower estimation error than its  $\ell_1$ -regularized counterpart, GLM(L1). Similarly, the KS method outperforms the KS(L1) method. The stationary GLM method also outperforms all time-varying methods, which makes sense because the true parameter function is not time-varying.



**Figure 6.5:** The five panels show the mean absolute estimation error averaged over the same type, time points, and iterations as a function of the number of observations  $n$  on a log scale. We report the error of six estimation methods: stationary unregularized regression (blue), stationary  $\ell_1$ -regularized regression (light blue), time-varying regression via kernel-smoothing (green), time-varying  $\ell_1$ -regularized regression via kernel-smoothing (light green), time-varying regression via GAM (pink), and time-varying regression via GAM with thresholding at 95% CI (red). Some data points are missing because the respective models are not identified in that situation (see Section 6.3.1.2).

For the linearly increasing/decreasing time-varying parameter in Figure 6.5 (b), the picture is more complex. For very small  $n$  from 20 to 46 the regularized methods GLM(L1) and KS(L1) perform best. This makes sense because, for



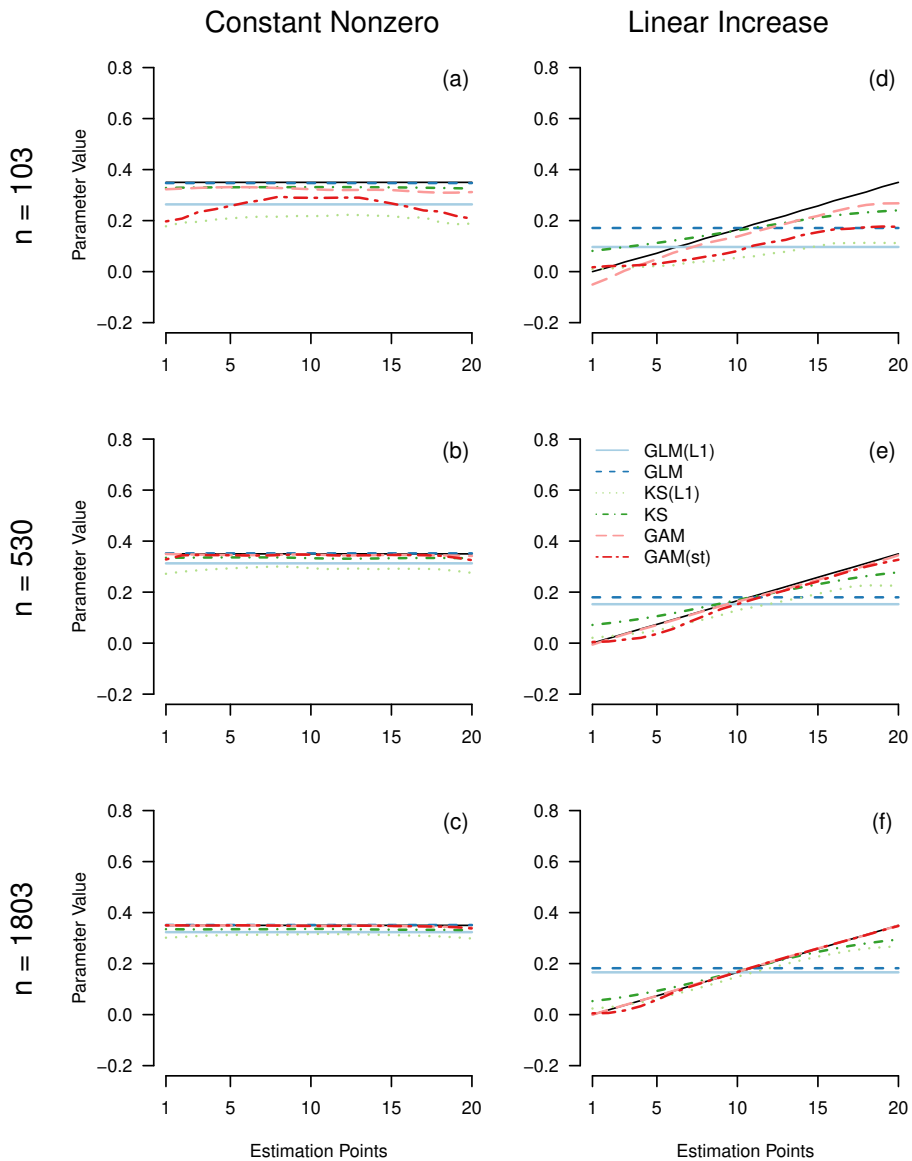
such small  $n$ , the estimates of all other methods suffer from huge variance. For sample sizes from 46 to 155 the unregularized methods perform better: now the bias of the regularized methods outweighs the reduction in variance. From sample sizes between 155 and 352 the time-varying methods start to outperform the two stationary methods. Interestingly, until around  $n = 530$  the KS methods outperforms all other time-varying methods. For  $n > 530$  all time-varying methods perform roughly equally. Overall, the error of all time-varying methods seem to converge to zero, as we would expect from a consistent estimator. The error of the stationary methods converges to  $\approx 0.088$ , which is the error resulting from approximating the time-varying function with the optimal constant function with value  $\frac{0.35}{2}$ . Since the sigmoid increase/decrease functions in panel (c) are very similar to the linear increase/decrease functions, we obtain qualitatively the same results as in the linear case.

In the case of the step function we again see a similar qualitative picture, however here the time-varying methods outperform the stationary methods already at a sample size of around  $n = 69$ . The reason is that the step function is more time-varying in the sense that here the best constant function is a worse approximation than in the linear and the sigmoid case. Another difference is that the GAM(st) method seems to outperform all other methods by a small margin if the sample size is large.

Finally, the absolute error for estimating the constant zero function is lowest for the regularized methods and the thresholded GAM method. This is what one expect since these methods bias estimates towards zero, and the true parameter function is zero across the whole time period.

In Figure 6.5 we reported the mean population errors of the six compared methods in various scenarios. These mean errors allow one to judge whether the *expected* error of one method will be larger than the one of another method. However, it is also interesting to inspect the population sampling variance around these mean errors. This allows one to gauge with which probability one method will be better than another for a given sample. We show a version of Figure 6.5 that includes the 25% and 95% quantiles of the absolute error in Appendix C.1.

**Absolute Error over Time for Constant and Linear Increasing Function** To investigate the behavior of the different methods in estimating parameters across the time interval, Figure 6.6 displays the mean absolute error for each estimation point (spanning the full period of the time series) for the constant nonzero function and the linear increasing function for  $n = \{103, 530, 1803\}$ . Note that these results were already shown in aggregate form in Figure 6.5: for instance, the average (across time) of estimates of the stationary  $\ell_1$ -regularized method in Figure 6.6 (a) corresponds to the single data point in Figure 6.5 (a) of the same method at  $n = 103$ .



**Figure 6.6:** Mean and standard deviations of estimates for the constant parameter (left column), and the linear increasing parameter (right column), for  $n = 103$  (top row),  $n = 530$  (second row) and  $n = 1803$  (bottom row) averaged over iterations, separately for the five estimation methods: stationary  $\ell_1$ -regularized regression (red), unregularized regression (blue), time-varying  $\ell_1$ -regularized regression via kernel-smoothing (green), time-varying regression via GAM (pink), and time-varying regression via GAM with thresholding at 95% CI (orange).

Panel (a) of Figure 6.6 shows the average parameter estimates of each method for the constant function with  $n = 103$  observations. In line with the aggregate results in Figure 6.5, the stationary methods outperform the time-varying methods,

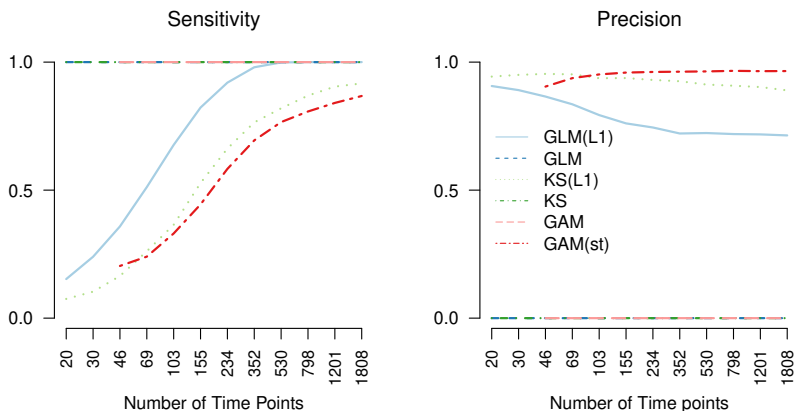
and the unregularized methods outperform the regularized methods. We also see that the KS(L1) and the GAM(st) methods are biased downwards at the beginning and the end of the time series. The reason is that less data is available at these points, which results in stronger bias toward zero (KS(L1)) and more estimates being thresholded to zero. When increasing  $n$ , all methods become better at approximating the constant nonzero function. This is what we would expect from the results in Figure 6.5, which suggested that the absolute error of all methods converges to zero as  $n$  grows.

In the case of the linear increase with  $n = 103$  (d), we see that the time-varying methods follow the form of the true time-varying parameter, however, some deviations exist. With larger  $n$ , the time-varying methods recover the linearly increasing time-varying parameter with increasing accuracy. In contrast, the stationary methods converge to the best-fitting constant function. We also see that the average estimates of the regularized methods are closer to zero than the estimates of the unregularized methods. However, note that, similar to panel (e) in Figure 6.5, the regularized methods would perform better in recovering the constant zero function.

Here we only presented the mean estimates of each method, which displays the bias of the different methods as a function of sample size. However, it is equally important to consider the variance around estimates. We display this variance in Figure C.2 in Appendix C.2. This figure shows that — as expected — the variance is very large for small  $n$ , but approaches 0 when  $n$  becomes large.

**Performance in Structure Recovery** In some situations the main interest may be to recover the *structure* of the VAR model, that is, we would like to know which parameters in the VAR parameter matrix are nonzero. We use two measures to quantify the performance of structure recovery. Sensitivity, the probability that a parameter that is nonzero in the true model is estimated to be nonzero; and precision, the probability that a nonzero estimate is nonzero in the true model. While higher values are better for both sensitivity and precision, different estimation algorithms typically offer different trade-offs between the two. Figure 6.7 shows this trade-off for the five estimation methods.

The unregularized stationary GLM method, the unregularized KS method, and the unthresholded time-varying GAM method have a sensitivity of 1 and a precision of 0 for all  $n$ . This is trivially the case because these methods return nonzero estimates with probability 1, which leads to a sensitivity of 1 and a precision of 0. Consequently, these methods are unsuitable for structure estimation. For all remaining methods, sensitivity seems to approach 1 when increasing  $n$ , while GLM(L1) has the highest sensitivity followed by KS(L1) and GAM(st). As expected, the precision of these methods is stacked up in reverse.



**Figure 6.7:** Sensitivity and precision for the five estimation methods across all edge-types for different variations of  $n$ . The lines for the unthresholded GAM(st) method and the stationary GAM method overlap completely, since they do not return estimates that are exactly zero. Some data points are missing because the respective models are not identified in that situation (see Section 6.3.1.2).

**Computational Cost** In Appendix C.3 we report the computational cost of the time-varying methods. The main take away from these results is that computation time is not a major concern for typical psychological applications.

#### 6.3.1.4 Discussion

The first simulation showed how the different methods perform in recovering a VAR model with  $p = 10$  variables based on a random graph, with linear, sigmoid, step and constant parameter functions, with sample sizes that cover most applications in psychology. The compared methods differ in the dimensions stationary vs. time-varying methods, unregularized vs. regularized methods, and GAM- vs. KS-based methods. Since all these dimensions interact with each other and with the type of time-varying parameter function they aim to recover, we discuss these interactions separately for each parameter function.

**Constant Nonzero Function** In the case of the constant nonzero function the stationary and unregularized GLM performed best, followed by the unregularized time-varying KS method. It makes sense that GLM performs best, because the true parameter function in this case is nonzero and constant across time. The KS method performs similarly especially for small  $n$ , because the bandwidth selection will select a very high bandwidth, which leads to a weighting that is almost equal for all time points, which leads to estimates that are very similar to the ones of the GLM method. The next best method is the stationary regularized GLM(L1) method. This is because the regularization decreases performance if the true parameter function is nonzero, however, it uses the correct assumption that the true parameter function is constant across time. Since the ability to estimate time-varying parameters is no advantage when estimating the constant

nonzero function, the KS(L1) method performs worse than the GLM(L1) method. Interestingly, the unregularized GAM function performs much worse than the unregularized KS method. The significance-thresholded GAM(st) method performs worse than the GAM method, because if the true parameter function is nonzero, thresholding it to zero can only increase estimation error.

**Linear and Sigmoid Functions** The results for the linear increasing/decreasing function are similar to the constant nonzero function, except that that all time-varying methods have a lower absolute error than the stationary methods from  $n > 234$ . The KS method is already better from  $n > 46$ . A difference to the constant nonzero function is that the two regularized methods GLM(L1) and KS(L1) perform best if the sample size is very small ( $n < 46$ ). A possible explanation for this difference is that the bias toward zero of the regularization is less disadvantageous for the linear increasing/decreasing functions, because its parameter values are on average only half as large as for the constant nonzero function. Within time-varying functions, the KS method performs better than the KS(L1) methods, which makes sense because the true parameter function is nonzero. For the same reason, the GLM method outperforms the GAM(st) method. The KS methods perform better than the GAM methods for sample sizes up to  $n = 530$ . The reason is that the estimates of the GAM methods have a larger sampling variance (see Figure C.1 in Appendix C.1). The errors in estimating the sigmoid function are very similar to the linear increasing/decreasing functions, since their functional forms are very similar.

**Step Function** The errors in estimating the step function are again similar to the linear and the sigmoid case, except for two differences: first, the time-varying methods become better than the stationary methods already between  $n = 46$  and  $n = 69$ . And second, the regularized KS(L1) performs better than KS, and the thresholded GAM(st) method performs better than the GAM method. The reason is that in half of the time series the parameter value is zero, which can be recovered exactly with the KS(L1) and the GAM(st) methods. This advantage seem to outweigh the bias these methods have in the other half of the time series in which the parameter function is nonzero.

**Constant Zero Function** In the case of the constant zero function the errors are roughly stacked up the reverse order as in the constant nonzero function. The regularized GLM(L1) and KS(L1) do best, followed by the thresholded GAM(ks) method. Among the unregularized methods the GLM and KS methods perform quite similarly, with the GLM method being slightly better, because the true parameter function is constant. Interestingly, the GAM method performs far worse, which is again due to its high variance (see Figure C.1 in Appendix C.1).

**Summary** We saw that stationary methods outperform time-varying methods when the true parameter function is a constant, and time-varying methods outperform stationary methods if the true parameter function is time-varying, and if

the sample size is large enough. The sample size at which the time-varying methods become better depends on how time-varying the true parameter is: the more time-varying it is, the smaller the sample size  $n$  at which time-varying methods become better than stationary ones. Within time-varying methods, the KS methods outperformed the GAM methods for smaller sample sizes, while the GAM based methods became better with very large sample sizes ( $n > 530$ ).

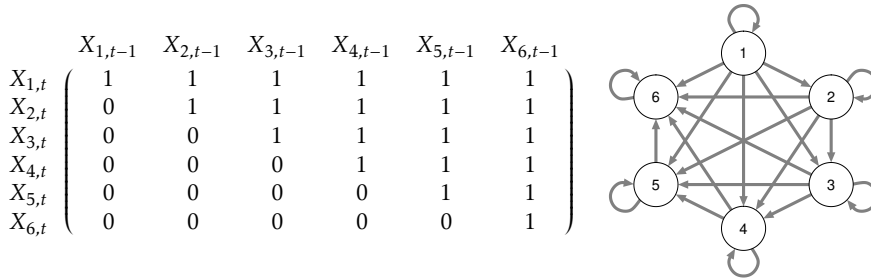
Finally, we saw that regularized methods perform better if the true parameter function is zero, while unregularized methods perform better if the true parameter function is nonzero, as expected. In order to choose between regularized and unregularized methods, one therefore needs to judge how many of the parameters in the true time-varying VAR model are nonzero. Given the expected sparsity of the true VAR model, one could compute a weighted average of the errors shown in this section in order to determine which method has the lowest overall error. However, to evaluate the performance of the different methods for different levels of sparsity more directly, we performed a second simulation study in which we vary the sparsity of the VAR model.

### 6.3.2 Simulation B: Varying Sparsity

In this simulation we evaluate the absolute estimation error of all methods for the different parameter functions and for the combined time-varying VAR model, as a function of sparsity. Specifically, we evaluate the estimation error of recovering the time-varying predictors of a given variable in the VAR model, depending on how many predictors are nonzero. From a network perspective the number of predictors on a given node is equal to its indegree. We will vary the indegree from 1 to 20. The average indegree in Simulation A was  $1 + 9 \times P(\text{edge}) = 2.61$ .

#### 6.3.2.1 Data Generation

We vary sparsity by specifying the structure of the initial VAR matrix to be upper-triangular. We show the structure of such a matrix, and the corresponding directed network in Figure 6.8. In such a model, the first variable has one predictor (itself at  $t - 1$ ), the second variables has two predictors (itself and variable 1 at  $t - 1$ ), the third variable has three predictors, etc. and the last variable has  $p$  predictors. As defined in Section 6.2, the number of nonzero predictor variables (or the indegree from a network perspective) is a local (i.e. for some variable  $X$ ) measure of sparsity. In the simulation we use the same initial VAR matrix, except that we use a VAR model with  $p = 20$  variables. All additional steps of the data generation (see Section 6.3.1.1, and the specification of the estimation methods (Section 6.3.1.2) are the same as in Simulation A.



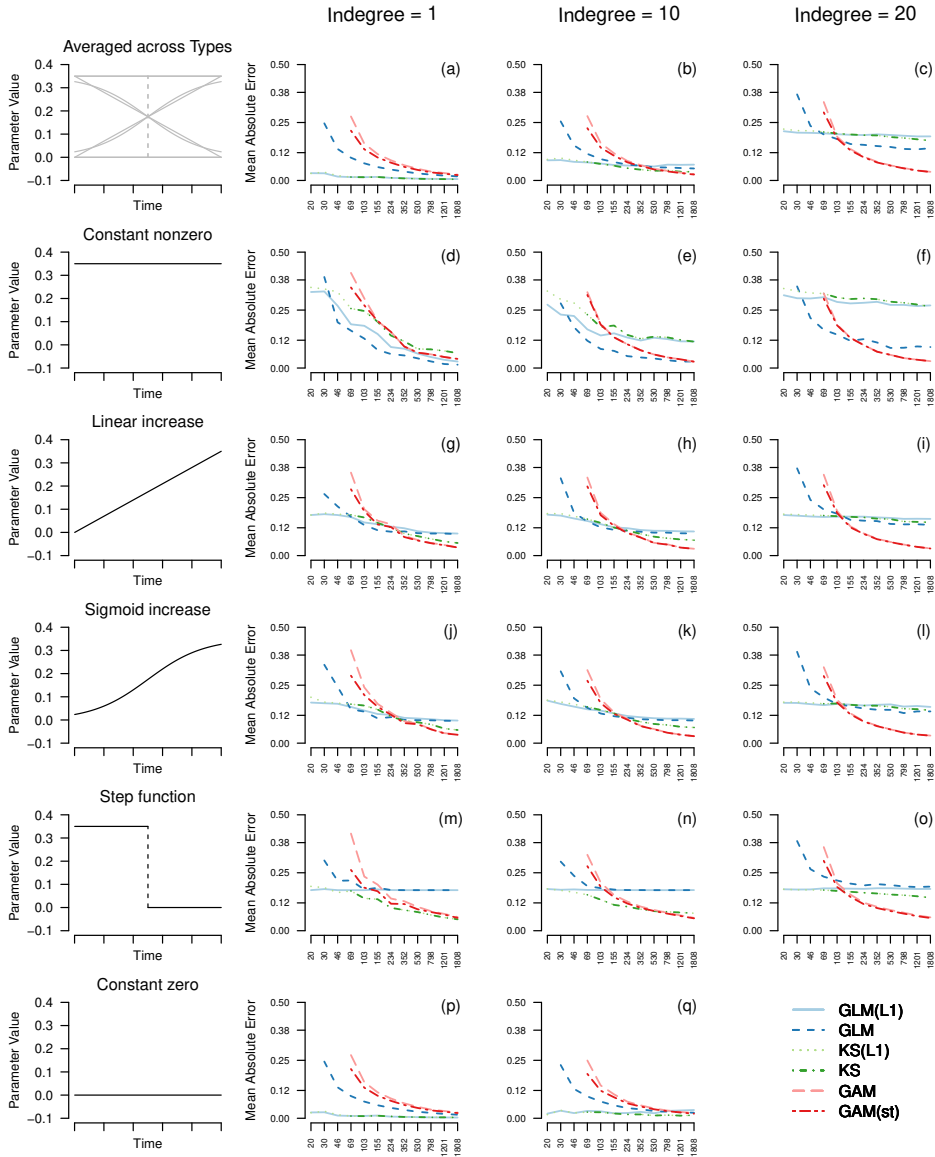
**Figure 6.8:** Left: the upper-diagonal pattern of nonzero parameters used in the time-varying VAR model in the second simulation, here shown for six variables. The row sums are equal to the indegree of the respective nodes, which results in a frequency of one for each indegree value. Right: visualization of the upper-diagonal pattern as a directed graph. The graph used in the simulation has the same structure but is comprised of 20 nodes.

### 6.3.2.2 Results

Figure 6.9 displays the mean absolute error separately for the five different time-varying parameter functions and for indegrees 1, 10, 20. Similarly to Simulation A, we collapsed symmetric increasing and decreasing functions into single categories and report their average performance. The first row of Figure 6.9 shows the performance averaged over time points and types of time-varying parameters for indegree 1, 10 and 20. The most obvious result is that all methods become worse when increasing the indegree. This is what one would expect since more parameters are nonzero and more predictors are correlated. In addition, there are several interactions between indegree and estimation methods. First, the regularized methods perform best when indegree is low, and worst when indegree is high. This makes sense: the bias toward zero of the regularization is more beneficial if almost all parameter functions are zero. However, if most parameter functions are nonzero, a bias toward zero leads to high estimation error. Second, we see that the drop in performance is lower for the GAM based methods compared to the KS based methods. The combined results in the first row are the weighted average of the remaining rows. The estimation errors for the time-varying functions show a similar pattern as in Figure 6.5 of Simulation A, except that the GAM methods perform better for indegree values 10 and 20.

### 6.3.2.3 Discussion

The results of Simulation B depicted the relative performance of all methods as a function of sparsity, which we analyzed locally as indegree. As expected, regularized methods perform better when indegree is low and worse if indegree is high. Interestingly, among the time-varying methods, the GAM based methods perform better than the KS based methods when indegree is high.



**Figure 6.9:** The mean average error for estimates of the upper-triangular model for all five estimation methods for the same sequence of numbers of time points  $n$  as in the first simulation. The results are conditioned on three different indegrees (1, 10, 20) and shown averaged across (a - c) and separately for the time-varying parameter types (d - q).

### 6.3.3 Overall Discussion of Simulation Results

Here we discuss the overall strengths and weaknesses of all considered methods in light of the results of both simulations.



**Stationary vs. Time-Varying Methods** We saw that stationary methods outperform time-varying methods if the true parameter function is constant, as one would expect. If the parameter function is time-varying, then the stationary methods are better for very small sample sizes, but for larger sample sizes, the time-varying methods become better. The exact sample size  $n$  at which time-varying methods start to perform better depends on how strongly the true parameters vary with time: the stronger the variation, the smaller the  $n$ . For the choice of true parameter functions in our simulations, we found that the best time-varying method outperformed the stationary methods at already  $n > 46$ .

**Unregularized vs. Regularized Methods** The results in both simulations showed that if most true parameter functions are zero (high sparsity), regularized methods and the thresholded GAM(st) method performed better compared to their unregularized/unthresholded counterparts. On the other hand, if most true parameter functions are nonzero (low sparsity), the unregularized/unthresholded functions perform better. In Simulation B we specifically mapped out the performance of methods as a function of sparsity and found that unregularized methods are better at an indegree of 10 or larger.

**Kernel-smoothing vs. GAM Methods** If sparsity is high, that is, if most parameter functions are zero, the KS based methods outperformed the GAM based methods for most sample size regimes. Only if the sample size is very large the GAM based methods showed a performance that is equal or slightly better than the KS based methods. However, if sparsity is low, the GAM based methods outperformed the KS based methods.

Accordingly, applied researchers should choose the KS based methods when they expect the time-varying VAR model to be relatively sparse and if they only have a moderate sample size ( $n < 200$  to  $300$ ). If one expects that only few parameter functions are nonzero, the KS based method should be combined with regularization. If one expects the parameter functions of the time-varying VAR model to be largely nonzero, and if one has a large sample size, the GAM based methods are likely to perform better.

**Limitations** Several limitations of the simulation studies require discussion. First, the signal to noise ratio  $S/N = \frac{\theta}{\sigma} = 3.5$  in parameter values could be larger or smaller in a given application and the performance results would accordingly be better or worse. Similarly, the signal to noise ratio would be smaller if we increased the number of variables  $p$ , because more parameters have to be estimated. However, note that  $S/N$  is also a function of  $n$ . Hence if we assume a lower  $S/N$  this simply means that we need more observations to obtain the same performance, while all qualitative relationships between time-varying parameters, structure in the VAR model and estimators remain the same.

Second, the time-varying parameters could be more time-varying. For example, we could have chosen functions that go up and down multiple times instead of being monotone increasing/decreasing. However, for estimation purposes, the

extent to which a function is time-varying is determined by how much it varies over a specified time period *relative* to how many observations are available in the time period. Thus the  $n$ -variations can also be seen as a variation of the extent to which parameters are varying over time: From this perspective, the time-varying parameter functions with  $n = 20$  are very much varying over time, while the parameter functions with  $n = 1808$  are hardly varying over time. Since we chose  $n$ -variations stretching from unacceptable performance ( $n = 20$ ) to very high performance ( $n = 1808$ ), we simultaneously varied the extent to which parameters are time-varying.

Third, we only investigated time-varying VAR models with  $p = 10$  variables and a single lag. In terms of the performance in estimating (time-varying) VAR parameters, adding more variables or lags boils down to increasing the indegree of a VAR model with a single lag and fixed  $p$ . In general, the larger the indegree and the higher the correlations between the predictors, the harder it is to estimate the parameters associated with a variable. Part of the motivation for Simulation B in Section 6.3.2 was to address this limitation.

Finally, we would like to stress that all statements with respect to sample size refer to the effective sample size available to estimate the VAR model. We mention this because the effective sample size that is used to estimate a VAR model is often considerably lower than the number of measurement points in an ESM study. This is both because of planned (e.g., at the day/night shift) and unplanned missing values. For example, if an ESM study has five measurements a day with a measurement interval of 3h and the fourth measurement is missing, then the effective sample size is only three, because only for three time points (2, 3, and 4) a measurement 3h before is available.

## 6.4 Estimating time-varying VAR model on Mood Time Series

In this section we provide a step-by-step tutorial on how to estimate a time-varying VAR model on a mood time series using the KS(L1) method. In addition, we show how to compute time-varying prediction errors for all nodes, how to assess the reliability of all estimates, and how to visualize some aspects of the estimated time-varying VAR model. Finally, we briefly discuss how to select between stationary and time-varying models. All analyses are performed using the R-package *mgm* (version 1.2-8) (Haslbeck & Waldorp, 2020) and R-version 3.6.0, and the code below can also be found as an R-file on Github: [https://github.com/jmbh/tvvar\\_paper](https://github.com/jmbh/tvvar_paper). In Appendix C.5 we show how to fit the same model with the GAM(st) method using the R-package *tvvarGAM*.

### 6.4.1 Data

We illustrate how to fit a time-varying VAR model on a symptom time series with 12 variables related to mood measured on 1476 time points during 238 consecutive days from an individual diagnosed with major depression (Wichers et al.,

2016). The measurements were taken at 10 pseudo-randomized time intervals with average length of 90 minutes between 07:30 and 22:30. During the measured time period, a double-blind medication dose reduction was carried out, consisting of a baseline period, the dose reduction, and two post assessment periods (See Figure 6.10, the points on the time line correspond to the two dose reductions). For a detailed description of this data set see Kossakowski et al. (2017).

## 6.4.2 Load R-packages and Dataset

The above described symptom dataset automatically available when loading the R-package *mgm*. After loading the package, we subset the 12 mood variables contained in this dataset:

```
library(mgm) # Version 1.2-8

mood_data <- as.matrix(symptom_data$data[, 1:12]) # Subset variables
mood_labels <- symptom_data$colnames[1:12] # Subset variable labels
colnames(mood_data) <- mood_labels
time_data <- symptom_data$data_time
```

The object `mood_data` is a  $1476 \times 12$  matrix with measurements of 12 mood variables:

```
> dim(mood_data)
[1] 1476 12

> head(mood_data[,1:7])
Relaxed Down Irritated Satisfied Lonely Anxious Enthusiastic
[1,] 5 -1 1 5 -1 -1 4
[2,] 4 0 3 3 0 0 3
[3,] 4 0 2 3 0 0 4
[4,] 4 0 1 4 0 0 4
[5,] 4 0 2 4 0 0 4
[6,] 5 0 1 4 0 0 3
```

The matrix `time_data` contains information about the time stamps of each measurement. This information is needed for the data preprocessing in the next section.

```
> head(time_data)
date      dayno beepno beeptime resptime_s resptime_e time_norm
1 13/08/12 226 1 08:58 08:58:56 09:00:15 0.000000000
2 14/08/12 227 5 14:32 14:32:09 14:33:25 0.005164874
3 14/08/12 227 6 16:17 16:17:13 16:23:16 0.005470574
4 14/08/12 227 8 18:04 18:04:10 18:06:29 0.005782097
5 14/08/12 227 9 20:57 20:58:23 21:00:18 0.006285774
6 14/08/12 227 10 21:54 21:54:15 21:56:05 0.006451726
```

For a sizable number of measurement points the individual did not provide a response. The *mgm* package takes care of this automatically, by only using those time points to estimate a VAR(1) model for which a measurement at the previous time point is available.

Some of the variables in this data set are highly skewed, which can lead to unreliable parameter estimates. Here we deal with this issue by computing bootstrapped confidence intervals (KS method) and credible intervals (GAM method), to judge how reliable the estimates are. Since the focus in this tutorial is on estimating time-varying VAR models, we do not investigate the skewness of variables in detail. However, in practice the marginal distributions should always be inspected before fitting a (time-varying) VAR model. A possible remedy for skewed variables is to transform them, typically by taking a root, the log, or transformations such as the nonparanormal transform (Liu, Lafferty, & Wasserman, 2009). A disadvantage of this approach is that the parameters are more difficult to interpret. For example, if applying the log-transform to  $X$ , then the cross-lagged effect  $\beta_{X,Y}$  of  $Y$  on  $X$  is interpreted as “When increasing  $Y$  at  $t - 1$  by 1 unit, the log of  $X$  at  $t$  increases by  $\beta_{X,Y}$ , when keeping all other variables at  $t - 1$  constant”.

### 6.4.3 Estimating Time-Varying VAR Model

Here we describe how to use the function `tvmvar()` of the *mgm* package to estimate a time-varying VAR model. A more detailed description of this function can be found in the help file `?tvmvar`. After providing the data via the `data` argument, we specify the type and levels of each variable. The latter is necessary because *mgm* allows one to estimate models including different types of variables. In the present case we only have continuous variables modeled as conditional Gaussian distributions, and we therefore specify `type = rep("g", 12)`. By convention the number of levels for continuous variables is specified as one `level = rep(1, 12)`.

Via the argument `estpoints` we specify that we would like to have 20 estimation points that are equally spaced across the time series (for details see `?tvmvar`). The number of estimation points can be chosen arbitrarily large, however at some point adding more estimation points becomes useless because adjacent estimation points become indistinguishable. Via the argument `timepoints` we provide a vector containing the time point of each measurement. The time points are used to distribute the estimation points correctly on the time interval. If no `timepoints` argument is provided, the function assumes that all measurement points are equidistant. See Section 2.5 in Haslbeck and Waldorp (2020) for a more detailed explanation how the time points are used in *mgm* and an illustration of the problems following from incorrectly assuming equidistant measurement points.

Next, we specify the bandwidth parameter  $b$ , which determines how many observations close to an estimation point are used to estimate the model at that point. Here we select  $b = 0.34$ , which we obtained by searching a candidate sequence of bandwidth parameters, and selected the value that minimized the out-of-bag cross-validation error. The latter is implemented in the function

`bwSelect()` (for details on this time-stratified cross-validation scheme see Section 6.3.1.2). Since `bwSelect()` repeatedly fits time-varying VAR models with different bandwidth parameters, the specification of `bwSelect()` and the estimation function `tvmlvar` are very similar. We therefore refer the reader for the code to specify `bwSelect()` to Appendix C.4.

After that we provide the number of the notification on a given day and the number of the day itself via the arguments `beepvar` and `dayvar`, respectively. This information is used to exclude cases from the analysis which do not have sufficient previous measurements to fit the specified VAR model. This can be both due to randomly missing data, or because of missingness by design. In the present dataset we have both: within a given day the individual did not always answer at all 10 times. And by design, there is a break between day and night. When not considering the correct successiveness the estimated parameters do not only reflect effects from  $t_{t-1}$  on  $t$  but also effects over (possibly) many other time-lags (for instance 10h over night instead of the intended 1h30).

Via the argument `lags = 1` we specify to fit a first order VAR model and specify with the argument `lambdaSel = "CV"` to select the penalty parameters  $\lambda$  with cross-validation. Finally, with the argument `scale = TRUE` we specify that all variables should be scaled to mean zero and standard deviation 1 before the model is fit. This is recommended when using  $\ell_1$ -regularization, because otherwise the strength of the penalization of a parameter depends on the variance of the predictor variable (see Hastie et al., 2015, p. 9). Since the cross-validation scheme uses random draws to define the folds, we set a seed to ensure reproducibility.

```
set.seed(1)
tvvar_obj <- tvmlvar(data = mood_data,
                    type = rep("g", 12),
                    level = rep(1, 12),
                    lambdaSel = "CV",
                    timepoints = time_data$time_norm,
                    estpoints = seq(0, 1, length = 20),
                    bandwidth = 0.34,
                    lags = 1,
                    beepvar = time_data$beepno,
                    dayvar = time_data$dayno,
                    scale = TRUE)
```

Before looking at the results we check how many of the 1476 time points were used for estimation, which is shown in the summary that is printed when calling the output object in the console:

```
> tvvar_obj
mgm fit-object
```

```
Model class: Time-varying mixed Vector Autoregressive (tv-mVAR) model
Lags: 1
Rows included in VAR design matrix: 876 / 1475 ( 59.39 %)
```

Nodes: 12

Estimation points: 20

This means that the VAR design matrix that is used for estimation has 876 rows. One of the removed time points is the first time point, since it does not have a previous time point. Other time points were excluded because of (a) missing measurements during the day or (b) the day-night break. As an example, from the six rows of the time stamps shown above, we could use three time points, since a measurement at the previous time point is available.

The absolute values of the estimated VAR coefficients are stored in the object `tvvar_obj$wadj`, which is an array of dimensions  $p \times p \times \text{lags} \times \text{estpoints}$ , `lags` is the number of lags, and `estpoints` is the number of estimation points. For example, the array entry `tvvar_obj$wadj[1, 3, 1, 9]` returns the cross-lagged effect of variable 3 on variable 1 with the first specified lag size (here 1) at estimation point 9. Due to the large number of estimated parameters, we do not show this object here but instead visualize some aspect of it in Figure 6.10. The signs of all parameters are stored separately in `tvvar_obj$signs`, because signs may not be defined in the presence of categorical variables (which is not the case here). The intercepts are stored in `tvvar_obj$intercepts`.

### 6.4.4 Assessing Reliability of Parameter Estimates

To judge the reliability of parameter estimates, we approximate the sampling distribution of all parameters using the nonparametric block bootstrap. The function `resample()` implements this bootstrap scheme and returns the sampling distribution and a selection of its quantiles of each parameter. First we provide the model object `object = tvvar_obj` and the data `data = mood_data`. `resample()` then fits the model specified as in `tvvar_obj` on 50 (`nB = 50`) different block bootstrap samples, where we specify the number of blocks via `blocks`. The argument `seeds` provides a random seed for each bootstrap sample and `quantiles` specifies the quantiles shown in the output.

```
res_obj <- resample(object = tvvar_obj,
  data = mood_data,
  nB = 50,
  blocks = 10,
  seeds = 1:50,
  quantiles = c(.05, .95))
```

The  $p \times p \times \text{lags} \times \text{estpoints} \times nB$  array `res_obj$bootParameters` contains the sampling distribution of each parameter. For instance, the array entry `res_obj$bootParameters[1, 3, 1, 9, ]` contains the sampling distribution of the cross-lagged effect of variable 3 on variable 1 with the first specified lag size (here 1) at time point 9. Due to its size, we do not show this object here but show the 5% and 95% quantiles of the empirical sampling distribution of three time-varying parameters in Figure 6.10. Also note that the resampling procedure is computationally expensive. With 50 bootstrap samples as specified above, the `resample()` runs approximately 10 minutes.

It is important to keep in mind that the quantiles of these bootstrapped sampling distributions are not confidence intervals around the true parameter. The reason is that the  $\ell_1$ -penalty biases all estimates and hence the whole sampling distribution towards zero which implies that the latter is not centered on the true parameter value.

### 6.4.5 Computing Time-Varying Prediction Error

Here we show how to compute time-varying nodewise prediction errors. Node-wise prediction errors indicate how well the model fits the data on an absolute scale and is therefore useful to judge the practical relevance of (parts of) a VAR model. See Haslbeck and Waldorp (2018) for a detailed description of node-wise prediction error (or predictability) in the context of network models and Haslbeck and Fried (2017) for an analysis of predictability in 18 datasets in the field of psychopathology.

The function `predict()` computes predictions and prediction errors from a given *mgm* model object. We first provide the model object `object = tvvar_obj` and the data `data = mood_data`. We then specify the desired types of prediction, here `R2` for the proportion of explained variance and `RMSE` for the Root Mean Squared Error. `tvMethod = "weighted"` specifies how to combine all time-varying models to arrive at a single prediction for each variable across the whole time series (for details see `?predict`). Finally, we provide `consec = time_data$beepno` for the same reasons as above.

```
pred_obj <- predict(object = tvvar_obj,
                   data = mood_data,
                   errorCon = c("R2", "RMSE"),
                   tvMethod = "weighted",
                   consec = time_data$beepno)
```

The predictions are stored in `pred_obj$predicted` and the error of the predictions of all time-varying models combined are in `pred_obj$errors`:

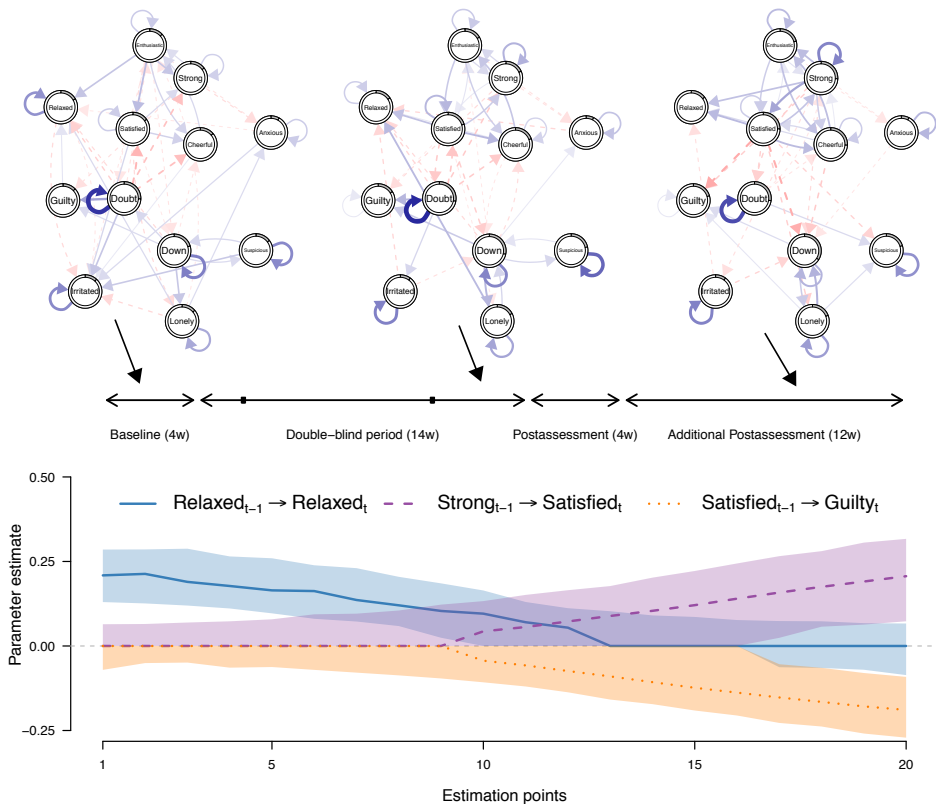
```
> pred_obj$errors
Variable Error.RMSE Error.R2
1      Relaxed      0.939  0.155
2        Down      0.825  0.297
3    Irritated      0.942  0.119
4    Satisfied      0.879  0.201
5      Lonely      0.921  0.182
6     Anxious      0.950  0.086
7 Enthusiastic      0.922  0.169
8  Suspicious      0.818  0.247
9   Cheerful      0.889  0.200
10     Guilty      0.928  0.175
11     Doubt      0.871  0.268
12     Strong      0.896  0.195
```

The prediction errors of each time-varying model separately are stored in `pred_obj$tverrors`. Note that here we weight the errors using the same weight

vector as used for estimation (see Section 6.2.3). For details see `?predict.mgm`. In the following section we visualize the time-varying nodewise estimation error for a subset of estimation points.

### 6.4.6 Visualizing Time-Varying VAR model

Figure 6.10 visualizes a part of the time-varying VAR parameters estimated above.



**Figure 6.10:** Top row: visualization of VAR(1) models at estimation points 2, 10 and 18. Blue solid arrows indicate positive relationships, red dashed arrows indicate negative relationships, and the width of the arrows is proportional to the absolute value of the corresponding parameter. The self-loops indicate autocorrelations. The colored parts of the ring around each node represents the respective within sample proportion of explained variance ( $R^2$ ). Bottom row: three parameters plotted as a function of time; the points are the point estimate obtained from the full dataset, the shaded areas indicate the 5% and 95% quantiles of the bootstrapped sampling distribution at each estimation point.

The top row shows visualizations of the VAR parameters for the estimation points 2, 10 and 18. Blue solid arrows indicate positive relationships, red dashed arrows indicate negative relationships. The width of the arrows is proportional



to the absolute value of the corresponding parameter. The grey part of the ring around each node indicates the proportion of explained variance of each variables by all other variables in the model. Comparing the VAR estimates across the three shown estimation points reveals that some parameters are strongly time-varying. For example, there is an autocorrelation effect of Relaxed at estimation point 2, which becomes smaller at estimation point 10 and vanishes at estimation point 18. On the other hand, the cross lagged effects  $\text{Strong}_{t-1} \rightarrow \text{Satisfied}_t$  and  $\text{Satisfied}_{t-1} \rightarrow \text{Guilty}_t$  are equal to zero at estimation point 2 and become larger in absolute value at estimation points 10 and 18. To better evaluate the time-varying nature of those three parameters we plot them as a line graph in the lower panel of Figure 6.10. Aligning time-varying parameter functions with additional information available about an individual may allow one to explain the changes in parameters. For example, we see that the three time-varying parameters in the lower panel show their largest change after the second reduction of the antidepressant medication. This suggests that the medication reduction could be part of the explanation for this change in parameters. Next to individual interaction parameters, possible analyses can also focus on the changes in intercepts or aggregates of several parameters. For example, one could investigate how the density of the entire or parts of the VAR model changes across time. The code to fully reproduce Figure 6.10 is not shown here due to its length, but can be obtained from Github (<https://github.com/jmbh/tvvar-paper>).

#### 6.4.7 Selecting between Stationary and Time-varying Models

While model selection between stationary and time-varying models is not the topic of this paper and requires a separate treatment to be addressed adequately, we briefly comment on this issue in relation to the methods presented here. One possible way to select between a stationary and a time-varying (VAR) model is to divide the time series into a training and test set. Then one can fit each model on the training set and evaluate on the test set which model has the lower prediction error. In fact, this is the procedure that is implemented in the function `bwSelect()` which we used in Appendix C.4 to select an appropriate bandwidth parameter, and which we described in detail in Section 6.3.1.2. Thus, if one includes large bandwidths ( $b > 1$ ) that are essentially leading to the same estimates as a stationary model, this bandwidth selection procedure includes a model selection procedure between stationary and time-varying models. However, selecting a (roughly) stationary model with this procedure does not necessarily imply that the data generating process is stationary. The reason is that the procedure strikes a balance between stability of estimates and sensitivity to estimate time-varying parameters. If the sample size is low, the procedure will therefore select a stationary model even if the data generating process is time-varying.

Another possibility is to rely on information criteria such as the AIC (see e.g., Bringmann et al., 2018). Finally, one could construct a hypothesis test with the null hypothesis that the data generating process is stationary VAR model. This could be done by estimating a stationary VAR model on the data set at hand, and then generating  $B$  time series of the same length as the original time se-

ries from this model. Then one fits a time-varying VAR model to each of those data sets and records a mean (over variables) prediction error. This way we obtain the sampling distribution of the prediction error under the null hypothesis, and we can perform a hypothesis test using the prediction error of the time-varying VAR model on the actual data as the test-statistic. We could for instance set  $\alpha = 0.05$ , which would mean that we would accept the time-varying model if its error is smaller than the 5% quantile of the sampling distribution. For the data in this tutorial this leads to the rejection of the null-hypothesis, which means that the data generating mechanism is not a stationary VAR model and it is therefore more appropriate to fit a time-varying VAR model. We provide the code to reproduce this test on in the supplementary materials and Github [https://github.com/jmbh/tvvar\\_paper](https://github.com/jmbh/tvvar_paper).

### 6.5 Discussion

We compared the performance of GAM and kernel-smoothing (KS) based methods in combination with and without regularization in estimating time-varying VAR models in situations that are typical for psychological applications. Our simulation results allow researchers to select the best method amongst the ones we considered based on sample size and their assumptions about the sparsity of the true VAR model. In addition, we provided step-by-step tutorials for the KS based method using the R-package *mgm* (Section 6.4) and for the GAM based method using the R-package *tvvarGAM* (Appendix C.5).

Next to assessing the relative performance of different methods, our paper also provides the first overview of how many observations are roughly necessary to estimate time-varying VAR models. For the time-varying functions studied in our paper, already for  $n > 46$  the best time-varying method outperformed stationary methods, suggesting that time-varying methods can be applied to typical ESM data. However, it is important to keep in mind that if the sample size is low, the time-varying methods return very similar estimates as their stationary counterparts. Thus, if the true parameter function is heavily depending on time, and the sample size is small, time-varying methods will not be able to recover most of this dependency on time.

There are several interesting avenues for future research on time-varying VAR models. First, in the present paper we focused on frequentist methods. However, time-varying VAR models can also be estimated in a Bayesian framework (Krueger, 2015). It would be interesting to compare the performance of these methods to the methods presented in this paper. Second, the methods presented here could be extended to beyond the standard VAR models. Examples are mixed VAR models, which allow to jointly model variables defined on different domains (Haslbeck & Waldorp, 2020), unified Structural Equation Models (SEM) that allow an extension of SEM models to different domains (J. Kim, Zhu, Chang, Bentler, & Ernst, 2007), or the graphical VAR model (Abegaz & Wit, 2013), which estimates both the VAR parameters and the residual structure  $\Sigma$  (see Section 6.2.1). In this model, identifying time-varying parameters is especially impor-

tant, because spurious relations in the residual structure can be induced by time-varying parameters. Third, all methods discussed in this paper are based on the assumption that the true parameters are smooth functions of time. However, in some situations it might be more appropriate to assume different kinds of local stationarity, for example piece-wise constant functions (e.g., Gibberd & Nelson, 2017; Bringmann & Albers, 2019). It would be useful to make those alternative estimation methods available to applied researchers, and possibly combine them with the methods presented here. Fourth, the Gaussian kernel in the KS method could be replaced by kernels with finite domains such as the box car function, in order to improve the computational efficiency of the algorithm. Finally, in this paper we focused on the population performance of the two presented methods in a variety of settings. However, we did not discuss in detail how to select between models (for example stationary vs. time-varying) in a practical application. Bringmann et al. (2018) analyzed the performance of information criteria for selecting between stationary and time-varying VAR models with two variables. We believe that a conclusive discussion of different model selection strategies in a variety of realistic situations would be an important avenue for future work.

## Acknowledgements

We would like to thank Denny Borsboom, Fabian Dablander, Marie Deserno, Sacha Epskamp and Oisín Ryan and for their useful comments on earlier versions of this manuscript. We would like to thank Simon Wood for answering several inquiries about his R-package *mgcv*.



# SELECTING BETWEEN AR AND VAR MODELS

---

## Abstract

Time series of individual subjects have become a common data type in psychological research. The Vector Autoregressive (VAR) model, which predicts each variable by all variables including itself at previous time points, has become a popular modeling choice for these data. However, the number of observations in typical psychological applications is often small, which puts the reliability of VAR coefficients into question. In such situations it is possible that the simpler AR model, which only predicts each variable by itself at previous time points, is more appropriate. Bulteel, Mestdagh, Tuerlinckx, and Ceulemans (2018) used empirical data to investigate in which situations the AR or VAR models are more appropriate and suggest a rule to choose between the two models in practice. We provide an extended analysis of these issues using a simulation study. This allows us to (1) directly investigate the relative performance of AR and VAR models in typical psychological applications, (2) show how the relative performance depends both on  $n$  and characteristics of the true model, (3) quantify the uncertainty in selecting between the two models, and (4) assess the relative performance of different model selection strategies. We thereby provide a more complete picture for applied researchers about when the VAR model is appropriate in typical psychological applications, and how to select between AR and VAR models in practice.

---

This chapter has been adapted from: Dablander F.\*, Ryan O.\* & Haslbeck J. M. B.\* (under revision). Choosing between AR(1) and VAR(1) Models in Typical Psychological Applications. Preprint: <https://psyarxiv.com/qgewy/>

## 7.1 Introduction

Time series of individual subjects have become a common data type in psychological research since collecting them has become feasible due to the ubiquity of mobile devices. First-order Vector Autoregressive (VAR) models, which predict each variable by all variables including itself at the previous time point, are a natural starting point for the analysis of dependencies across time in such data and are already used extensively in applied research (e.g., Bringmann et al., 2013; Pe et al., 2015; Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019).

An acute question that arises when using these models is: how reliable are the estimates of the single-subject VAR model, given the typically short time series in psychological research (i.e.,  $n \in [30, 200]$ )? To be more precise, we would like to know how large the *estimation error* is in this setting. Estimation error is defined as the distance between the estimated parameters and the parameters in the true model, assuming that the true model has the same parametric form as the estimated model. If estimation error is large, it might be possible to obtain a smaller estimation error by choosing a simpler model, even though it is less plausible than the more complex model (J. H. Friedman, 1997). A possible simpler model in this setting is the first-order Autoregressive (AR) model, in which each variable is predicted only by itself at the previous time point. While the AR model introduces a strong bias by setting all interactions *between* variables to zero, it can have a lower estimation error when the number of available observations is small. When analyzing time series in psychological research it is therefore important to know (a) in which settings the AR or the VAR model has a lower estimation error, and (b) how to choose between the two models in practice.

Bulteel et al. (2018) identified these important and timely questions, and offered answers to both. They investigated question (a) regarding the relative performance of AR and VAR models by selecting three empirical time series data sets, each consisting of a number of individual time series with the same data structure. For each of these data sets, they approximate the out-of-sample prediction error with out-of-bag cross-validation error for both the AR and the VAR model and their mixed model versions. The authors make a valuable contribution by assessing which of the many cross-validation schemes available for time series approximates prediction error best in this context. Using the approximated prediction error obtained via cross-validation, they find that the prediction error for AR is smaller than for VAR, and that the prediction error of mixed AR and mixed VAR is similar. In a last step, they link prediction and estimation error by stating that “[...] the number of observations  $T$  [here  $n$ ] that is needed for the VAR to become better than the AR is the same for the prediction MSE [mean squared error] as well as for the parameter accuracy [estimation error]” (Bulteel et al., 2018, p. 10). Although the latter statement implies that the estimation error of mixed AR and mixed VAR models are similar, Bulteel et al. (2018) conclude that “[...] it is not meaningful to analyze the presented typical applications with a VAR model” (p. 14) when discussing both mixed and single-subject models.

Using their statement about the link between prediction error and estimation error, together with a preference towards parsimony, Bulteel et al. (2018) also of-

fer an answer to question (b) on how to choose between the AR and VAR models in practice: they suggest using the “1 Standard Error Rule”, according to which one should select the AR model if its prediction error is not more than one standard error above the prediction error of the VAR model, and select the model with lowest prediction error otherwise (Hastie, Tibshirani, & Friedman, 2009, p. 244).

In this commentary, we provide an extended analysis of the problems studied by Bulteel et al. (2018). First, regarding question (a) on the relative performance of AR and VAR models: when the goal is to determine the estimation error in a given setting, one can obtain it directly with a simulation study. A simulation study allows for a more extensive analysis of this problem for three reasons. First, we do not need to make any claim about the relation between prediction error and estimation error, which — as we will show — turns out to be non-trivial. Second, in a simulation study we can average over sampling variance which allows us to compute the expected value of estimation (and prediction) error. While the approach of Bulteel et al. (2018) in using three empirical datasets has the benefit of ensuring the models considered mirror data from psychological applications, these empirical datasets are naturally subject to sampling variation. And third, a simulation study allows us to map out the space of plausible VAR models and base our conclusions on this large set of VAR models instead of the VAR models estimated from the three data sets used by Bulteel et al. (2018). In Section 7.2 we perform such a simulation study, which allows us to give a direct answer to the question of how large the estimation errors of AR and VAR models are in typical psychological applications.

Regarding question (b) on choosing between AR and VAR models in practice, Bulteel et al. (2018) base their “1 Standard Error Rule” (1SER) on the idea that the  $n$  at which the estimation errors of the AR and VAR models cross is (approximately) the same  $n$  at which the prediction errors cross, combined with a preference towards the more parsimonious model. While the 1SER is used as a heuristic used in the statistical learning literature (Hastie et al., 2009), it is not clear why this heuristic would perform better in the present problem than simply selecting the model with the lowest prediction error. In Section 7.3 we show that when choosing between AR and VAR models, the  $n$  at which prediction errors become equal is not necessarily the same as the  $n$  at which estimation errors become equal: in fact, there is a substantial degree of variation in how the prediction and estimation errors of both models cross. Using the relationship between estimation and prediction error we are able to show via simulation when the 1SER is expected to perform better than selecting the model with lowest prediction error. This extended analysis of the problem studied by Bulteel et al. (2018) provides a more complete picture for applied researchers about when the VAR model is appropriate in typical psychological applications, and how to select between AR and VAR models in practice.

## 7.2 When does VAR outperform AR?

In this section we report a simulation study which directly answers the question of how large the estimation errors of AR and VAR models are in typical psychological applications. This allows the reader to get an idea of how many observations  $n_e$  one needs, on average, for the VAR model to outperform the AR model. In addition, we will decompose the variance around those averages in sampling variation and variation due to differences in the VAR parameter matrix  $\Phi$ . Finally, explaining the latter type of variation allows us to obtain  $n_e$  conditioned on characteristics of  $\Phi$ .

### 7.2.1 Simulation Setup

Since the AR model is nested under the more complex VAR model, we focus solely on the VAR as the true data-generating model. To obtain realistic VAR models, we use the following approach: first, we estimate a mixed VAR model to the “MindMaastricht” data (Geschwind, Peeters, Drukker, van Os, & Wichers, 2011), which consists of 52 individual time series with on average  $n = 41$  measurements on  $p = 6$  variables, and is the only publicly available data set used by Bulteel et al. (2018). In a second step, we sample stationary VAR models with a diagonal error covariance matrix from this mixed model.

We expect that the estimation (and prediction) errors of the AR and VAR model depend not only on the number of observations  $n$ , but also on the characteristics of the underlying  $p \times p$  VAR model matrix  $\Phi$ . We therefore stratify the sampling process from the mixed model by two characteristics of  $\Phi$ . This procedure allows us to obtain a better picture of how the performance of AR and VAR may differ depending on the characteristics of the data generating model.

The first characteristic is based on the relative size of the auto-regressive ( $\Phi_{ii}$ ) and cross-lagged effects ( $\Phi_{ij}$ ,  $i \neq j$ ), operationalized as the ratio

$$R = \frac{1}{p} \sum_{i=1}^p |\Phi_{ii}| / \frac{1}{p(p-1)} \sum_{i=1}^p \sum_{j \neq i}^p |\Phi_{ij}| .$$

We expect that true VAR models with a large  $R$  value (i.e., large auto-regressive effects and small cross-lagged effects) result in a low estimation error for AR models, since these VAR models are very similar to an AR model. In contrast, if  $R$  is small, we expect that the estimation error of the AR model is large, because it sets the large cross-lagged effects in the true VAR model to zero.

The second characteristic we consider is based on the eigenvalues  $\lambda$  of the  $\Phi$  matrix, which are commonly used to describe VAR models in the time-series literature (Hamilton, 1994). The absolute value of the eigenvalues encodes information about the dynamics described by  $\Phi$ , with higher absolute eigenvalues denoting a relatively larger carry-over of information from one time-point to the next. We summarize the information contained in  $\lambda$  by taking the mean of their absolute values, also referred to here as the dimensionality  $D$ , given as



$$D = \frac{1}{p} \sum_{i=1}^p |\lambda_i| ,$$

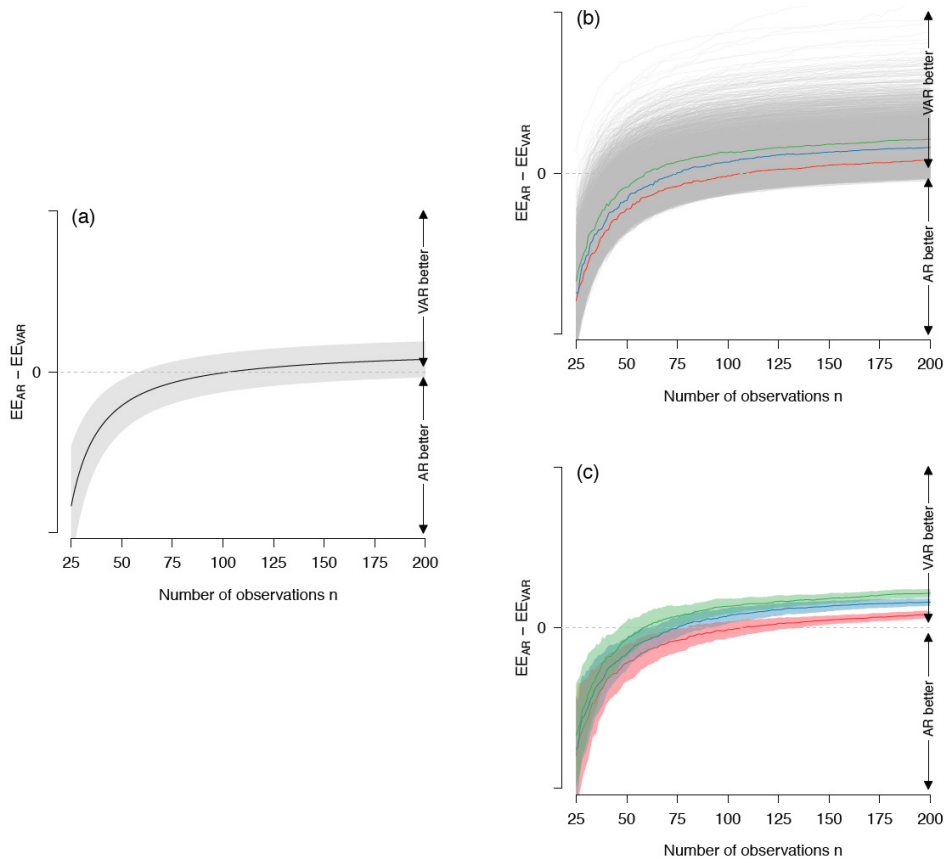
where  $|\lambda_i|$  is the absolute value of the  $i^{\text{th}}$  eigenvalue of  $\Phi$ , and where, to ensure stationarity, only  $\Phi$  matrices with  $|\lambda| < 1$  are included in our analysis. As the eigenvalues are a function of the auto-regressive and cross-lagged parameters,  $R$  and  $D$  are distinct yet correlated characteristics of the VAR model.

Ideally, we would stratify by sampling a fully crossed grid of  $R$  and  $D$  values. However, this is not possible since both measures are correlated and thus some combinations have an extremely small probability. We therefore adopt the following approach: we define a grid of cells with width/height 0.5 on the  $R$ - $D$  plane (see Figure 7.4) and sample 1000 VAR models. We then include only those cells in the design in which *at least one* VAR model has been sampled (see Figure 7.2 in Appendix D.1). This procedure returned 60 non-empty cells. We then sample those 60 cells until each of them contains 100 VAR models. We keep the cell size constant to render the results comparable across cells (see Appendix D.1 for a detailed description of this procedure).

This procedure returns a set of  $60 \times 100 = 6000$  VAR models that includes essentially any stationary VAR model with  $p = 6$  variables, and allows us to describe each model in the dimensions  $R$  and  $D$ . For each of these VAR models, we generate 100 independent time series with  $n = 500$  observations, with a burn-in of  $n_{\text{burn}} = 100$ . We then estimate both the AR and the VAR model on  $n = \{8, 9, \dots, 499, 500\}$  observations. For each model, and each  $n$ , we compute the expected estimation error for both the AR model ( $EE_{\text{AR}}$ ) and the VAR model ( $EE_{\text{VAR}}$ ) model by averaging over the 100 independent time series. This means that while  $EE_{\text{AR}}$  and  $EE_{\text{VAR}}$  have different values depending on  $n$  and the underlying model, we have averaged over the sampling variation.

### 7.2.2 Simulation Results

The simulation described above allows us to investigate the relative performance of AR and VAR models across different samples, sample sizes, and data-generating models. We define the estimation error as the mean squared error of the estimated parameters to the true parameters, and quantify the *relative* performance with two measures: the difference between the estimation errors of the AR and VAR models at a particular sample size,  $EE_{\text{Diff}} = EE_{\text{AR}} - EE_{\text{VAR}}$ ; and,  $n_e$ , the sample size at which the VAR model outperforms the AR model ( $EE_{\text{AR}} > EE_{\text{VAR}}$ ). In the following we examine the mean and variance of  $EE_{\text{Diff}}$  and subsequently study  $n_e$  and its dependence on the characteristics of the true VAR model.

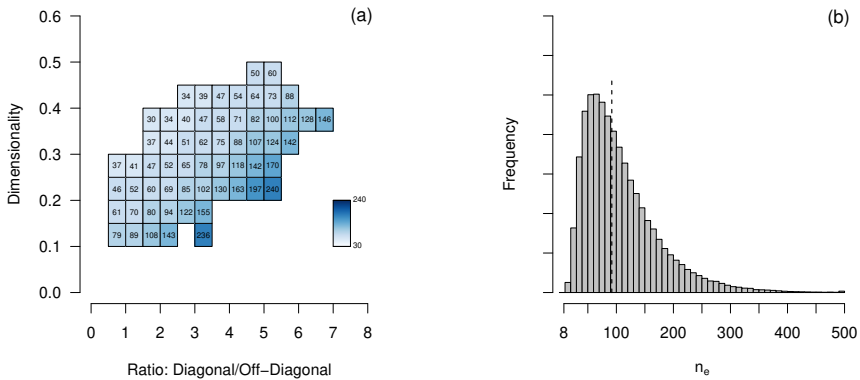


**Figure 7.1:** Difference in estimation error of AR and VAR models ( $EE_{Diff}$ ) across  $n$  on three different levels of aggregation. Panel (a) shows  $EE_{Diff}$  averaged over iterations and models, and the band shows the standard deviation over iterations and models; panel (b) shows  $EE_{Diff}$  for each model averaged across iterations; and panel (c) shows the  $EE_{Diff}$  averaged over iterations for three specific models, and the bands show the standard deviation across 100 iterations (sampling variation).

Figure 7.1 (a) shows the mean and standard deviation of  $EE_{Diff}$  as a function of  $n$ , across all 6000 VAR models and 100 iterations. The dashed line at  $EE_{Diff} = 0$  indicates the point at which the estimation errors of the two models are equal. Below that line, the AR model performs better, that is, its parameter estimates are closer to the parameters of the true VAR model than the parameter estimates of the VAR model. We see that, across all models,  $n_e = 92$  (median). Note that, out of all 600,000 simulated data sets, in only 327 cases the estimation error curves did not yet cross with an  $n$  of 500. Notably, the variance around the difference in estimation error is substantial for all  $n$ . In the following we decompose this variance in variance due to sampling error, and variance due to differences in VAR matrices.

Panel (b) of Figure 7.1 displays the mean  $EE_{Diff}$  for each of the 6000 VAR

models, averaged across 100 iterations. We see that the lines differ considerably and that  $n_e$  substantially depends on the characteristics of the true VAR model. This shows that one cannot expect reliable recommendations with respect to  $n_e$  that ignore the characteristics of the generating model. To illustrate the extent of the sampling variation of the models, we have chosen three particular VAR models (see colored lines). Figure 7.1 (c) shows that they exhibit considerable sampling variation. Note that, as the variance in (b) is due to differences in mean performance across VAR models, it does not decrease with  $n$ . In contrast, the variance in (c) depends on  $n$  as it pertains to the sampling variance of a single VAR model, which decreases with the square root of the number of observations. While the mean  $EE_{Diff}$  (shown in Figure 7.1 (a)) gives a clear answer to the question of which  $n$  is required for the VAR model to outperform the AR model *on average*, both types of variations (see Figure 7.1 (b) and (c)) show that for any *particular* VAR model it is difficult to determine which model performs better with the sample sizes typically available in psychological applications. However, we see that the sampling variation is smaller than the variation across VAR models for most  $n$ . This means that one can make much more precise statements about the relative performance if one specifies the data generating model.



**Figure 7.2:** Left:  $n_e$ , the  $n$  at which estimation error becomes lower for the VAR than for the AR model, as a function of  $D$  and  $R$ . Right: sampling distribution of  $n_e$ , the  $n$  at which the expected estimation error of the VAR model becomes lower than the expected estimation error of the AR model. The dashed line indicates the median of 92.

The large degree of variation around  $EE_{Diff}$  also highlights the potential pitfalls of generalizing the findings of Bulteel et al. (2018) beyond the empirical data sets, which consist of 28, 52, and 95 individual time-series with an average number of 41, 70 and 70 time points, analyzed by the original authors. This is because (i) it is unlikely that their (in total) 175 time series appropriately represent the population of all plausible VAR matrices, (ii) their sample is subject to a substantial amount of sampling variation, and (iii) the absence of systematic variations

of  $n$  does not allow a comprehensive answer to how relative performance relates to sample sizes in principle.

Above we suggested that the relative performance of AR and VAR models (quantified by  $EE_{\text{Diff}}$ ) depends on the characteristics  $R$  and  $D$  of the true VAR parameter matrix. In Figure 7.2 (a) we show the average (across models in cells)  $n$  at which the estimation error of VAR becomes smaller than the estimation of AR (i.e.,  $EE_{\text{Diff}} > 0$ ). We see that the larger the dimensionality  $D$ , and the smaller the ratio  $R$ , the lower the  $n$  at which VAR outperforms AR. This is what one would expect: for large  $R$  values, which reflects small off-diagonal elements, the true VAR model is actually very close to an AR model. In such a situation, the bias introduced by the AR model by setting the off-diagonal elements to zero leads to a relatively small estimation error. It therefore takes a considerable amount of observations until the variance of the VAR estimates becomes small enough to outperform the AR model.

In contrast, if  $R$  is small, which reflects large off-diagonal elements, the bias of the AR model leads to a comparatively larger estimation error. In this situation the VAR model, regardless of the high variance of its parameter estimates, can already outperform the AR model. This trade-off between a simple model with high bias but low variance and a more complex model with low bias but high variance is well-known in the statistical literature as the *bias-variance trade-off* (Hastie et al., 2009). Note that the characteristics  $R$  and  $D$  explain the vertical variation of the estimation error curves shown in Figure 7.1 (a): the curves on top (small  $n_e$ ) have high  $D$  and low  $R$ , while the curves at the bottom (large  $n_e$ ) have low  $D$  and high  $R$ . Figure 7.2 (b) collapses across these values and illustrates the sampling distribution of  $n_e$ , taking into account the probability of any particular VAR matrix (as specified by the mixed model estimated from the “MindMaastricht” data).

In summary, we used a simulation study to investigate the relative performance of AR and VAR models in a much larger space of plausible data-generating VAR models in psychological applications than considered by Bulteel et al. (2018). Next to investigating the *average* relative performance as a function of  $n$ , we also looked into the variation around averages. We showed that there is substantial variation both due to sampling error and differences in VAR matrices, which means that for a particular time series at hand it is difficult to select between AR and VAR with the  $n$  available in typical psychological applications. Finally, we confirmed the intuition that the relative performance depends on the characteristics of the true VAR matrix.

### 7.3 Choosing between VAR and AR based on Prediction Error

In the previous section, we directly investigated the estimation errors of the AR and the VAR model in typical psychological applications and showed that the  $n$  at which VAR becomes better than AR depends substantially on the characteristics of the true model. In practice, the true model is unknown, so we can neither

look up the  $n$  at which VAR outperforms AR in the above simulation study, nor can we compute the estimation error on the data at hand. We therefore have to resort to prediction error, which we can approximate using the data at hand, for instance by using a cross-validation scheme as suggested by Bulteel et al. (2018). However, since we are interested in estimation error, we require a link between prediction error and estimation error. In the remainder of this section we investigate this important link between prediction and estimation error. We describe the implications of this link for the model selection strategy suggested by Bulteel et al. (2018), who use the “1 Standard Error Rule” (1SER) to select the model with lowest estimation error. Finally, we use our simulation study from above to directly compare the performance of the 1SER with model selection based only on the minimum prediction error.

### 7.3.1 The Relation between Prediction Error and Estimation Error

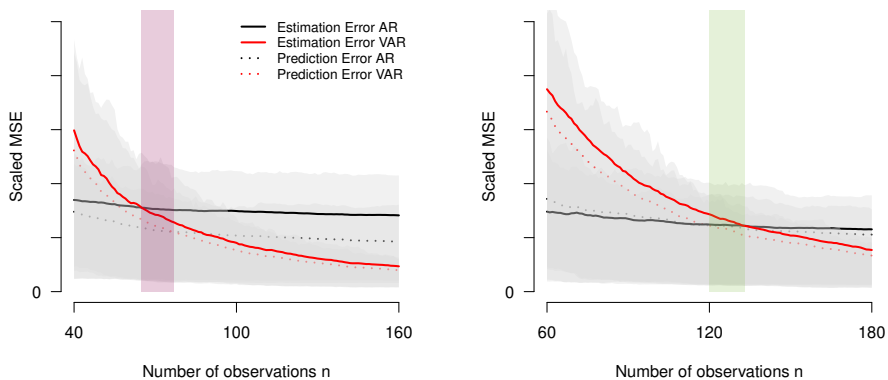
Bulteel et al. (2018) suggest that the link between prediction error and estimation error is straightforward: “[...] the number of observations  $T$  [here  $n$ ] that is needed for the VAR to become better than the AR is the same for the prediction MSE [mean squared error] as well as for the parameter accuracy [estimation error]” (Bulteel et al., 2018, p. 10). More formally, this claim states that if  $n_e$  is the number of observations at which the estimation errors of the AR and VAR model are equal, and if  $n_p$  is the number of observation at which the prediction errors of the AR and VAR model are equal, and  $n_{\text{gap}} = n_e - n_p$ , then  $n_{\text{gap}} = 0$ . Bulteel et al. (2018) do not specify the exact conditions under which this statement should hold, and elsewhere in the text suggest that this should be considered an approximate rather than exact relationship. If this relationship were indeed approximate, it would still be interesting to study in which settings  $n_{\text{gap}} > 0$  or  $n_{\text{gap}} < 0$ , as this bears on model selection, and so we will here focus our investigation quantifying  $n_{\text{gap}}$  and investigating any systematic deviations from zero this quantity exhibits through simulation.

Clearly, it would be unreasonable to expect that  $n_{\text{gap}} = 0$  for *any* data set, since the observations in a given data set are subject to sampling error. We therefore interpret the statement of Bulteel et al. (2018) as a statement about the *expectation* over errors of *any* given VAR model. Assuming momentarily that indeed  $n_{\text{gap}} = 0$  on average for all VAR models, this would mean that if the prediction errors of both models are the same (similar), then the estimation of both models are the same (similar). However, after observing that the single-subject AR outperforms the single-subject VAR and that the mixed AR and mixed VAR have similar prediction errors, Bulteel et al. (2018) conclude that “[...] it is not meaningful to analyze the presented typical applications with a VAR model” (p. 14). Since meaningfulness can only refer to estimation error in the present context, the authors seem to overgeneralize: while this statement is true for the single-subject case, it does not hold in the mixed case. In particular, from  $n_{\text{gap}} = 0$  it follows that if the prediction errors of both models are the same, then their estimation errors are the same, and hence both models are equally meaningful. Note that one could

choose the AR model over the VAR model in this situation by applying a bias towards the more parsimonious model. However, we question whether invoking the principle of parsimony is justifiable when choosing between two models of which the more parsimonious model is theoretically implausible (as the original authors state themselves). In any event, the VAR model would certainly not be more meaningless than the AR model, since their estimation errors would be the same.

### 7.3.2 Assessing $n_{\text{gap}}$ through simulation

We now use the results of the simulation study from the previous section to check whether indeed  $n_{\text{gap}} = 0$  on average for all VAR models. For each of the 6000 models, and for each  $n$ , we compute the prediction error averaged across the 100 models estimated from the 100 data sets generated from the true VAR model, evaluated on a separate time series with  $n_{\text{test}} = 2000$  observations, with a burn-in of  $n_{\text{burn}} = 100$ . This is the out of sample prediction error (i.e., the expected generalization error) that Bulteel et al. (2018) approximate with out-of-bag cross-validation error. We define prediction error as the mean squared error (MSE) of the predicted values relative to the true values in the test data set.

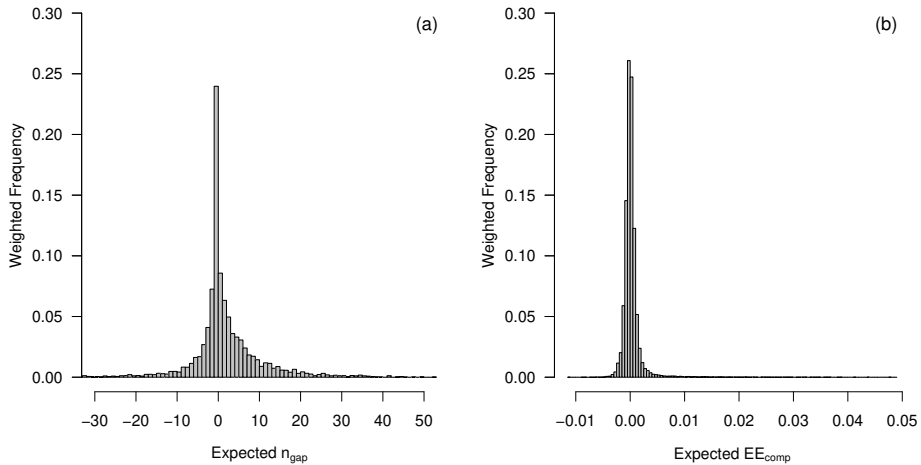


**Figure 7.3:** Scaled Mean Squared Error (MSE) of estimation (solid lines) and prediction errors (dashed lines) for both the AR (black lines) and VAR (red lines) models as a function of  $n$ , separately for model A with  $R = 1.47$  and  $D = 0.24$  (left panel) and model B with  $R = 6.66$  and  $D = 0.36$  (right panel). The red and green shaded area indicates the median  $n_{\text{gap}}$ , and the grey shaded area shows the 20% and 80% quantiles across the 100 iterations per model.

Figure 7.3 shows the estimation (solid lines) and prediction (dashed lines) errors for both the AR (black lines) and VAR (red lines) models as a function of  $n$ , averaged across the iterations, for model A with  $R = 1.47$  and  $D = 0.24$  (left panel) and model B with  $R = 6.66$  and  $D = 0.36$  (right panel). For model A, we see that  $n_{\text{gap}} < 0$ , which shows that  $n_{\text{gap}} = 0$  for all VAR models is incorrect.

What consequences does this gap have for model selection? The negative gap implies that if the prediction errors for the AR and VAR model are the same, the VAR model should be selected, because its estimation error is smaller. In contrast, for model B we observe  $n_{\text{gap}} > 0$ . In this situation, if the prediction errors are equal, one should select the AR model because it incurs smaller estimation error. Clearly,  $n_{\text{gap}}$  differs between the two models, and this difference matters for model selection.

So far we only investigated  $n_{\text{gap}}$  for two individual VAR models. Figure 7.4 (a) shows the distribution of the expected  $n_{\text{gap}}$  across all VAR models, computed by averaging over 100 iterations. Note that for 97 out of 6000 models the curves of prediction errors and estimation errors did not cross within  $n \in \{8, 9, \dots, 499, 500\}$ . The results in Figure 7.4 are therefore computed on 5903 models.



**Figure 7.4:** Panel (a) displays the distribution of the expected  $n_{\text{gap}}$  across all 6000 VAR models, computed by averaging over 100 iterations, and weighted by the probability defined by the original mixed model. Panel (b) shows the distribution of *non-zero*  $EE_{\text{comp}}$  across all  $n$ , 6000 VAR models, averaged across iterations and weighted by the probability defined by the original mixed model.

Each of the data points in the histogram in Figure 7.4 (a) corresponds to the *expected*  $n_{\text{gap}}$  of one of the 6000 models. We see that the expected  $n_{\text{gap}}$  has a right skewed distribution with a mode at zero. This allows us to make a precise statement regarding the crossing of estimation and prediction errors described above: while the most common value of  $n_{\text{gap}}$  is zero, not all expected  $n_{\text{gap}}$  are zero. In fact,  $n_{\text{gap}}$  shows substantial variation across different VAR models. Explaining the variance of  $n_{\text{gap}}$  is interesting, because  $n_{\text{gap}}$  has direct consequences for model selection. If we can relate the  $n_{\text{gap}}$  to characteristics of the  $\Phi$  matrix, it is possible to make more specific statements with respect to when to apply a bias towards the AR or VAR model, when the prediction errors are the same or very similar.

Note that such a function from  $\Phi$  to  $n_{\text{gap}}$  must exist, because the only way the 6000 models differ is in their entries of the VAR parameter matrix  $\Phi$ . However, this function may be very complicated. For example, the correlation of  $n_{\text{gap}}$  with  $R$  and  $D$  are 0.20 and 0.18, respectively. Predicting  $n_{\text{gap}}$  by  $R$  and  $D$  including the interaction term with linear regression achieves  $R^2 = 0.043$ . This shows that a simple linear model including  $R$  and  $D$  is not sufficient to describe the relationship between  $n_{\text{gap}}$  and  $\Phi$ . Future research could look into better approximations of this relationship. If successful, one could build new model selection strategies on reliable predictions of  $n_{\text{gap}}$  from empirical data.

### 7.3.3 Performance of the “1 Standard Error Rule”

Bulteel et al. (2018) propose, in the words of Hastie et al., to “[...] choose the most parsimonious model whose error is no more than one standard error above the error of the best model.” (Hastie et al., 2009, p. 244). This model selection criteria is known as the “1 Standard Error Rule” (1SER) and is suggested by Hastie and colleagues as a method of choosing a model with the minimal out-of-sample prediction error (which is typically unknown), on the basis of out-of-bag prediction error (acquired with cross-validation techniques).

Making inferences from prediction error to estimation error requires a link between the two. Bulteel et al. (2018) provide this link by suggesting that  $n_{\text{gap}} = 0$  (or  $n_{\text{gap}} \approx 0$ ). However, they do not provide justification for why the 1SER should outperform simply selecting the model with the lowest prediction error. Above we showed that  $n_{\text{gap}} = 0$  does not hold for all VAR models. In fact, it is this result that explains why the 1SER can perform better than selecting the model with the lowest prediction error. Specifically, this is the case when  $n_{\text{gap}} > 0$ , which characterizes the situation that the prediction error for VAR is lower than for AR while at the same time the estimation error of VAR is higher than for AR. In such a situation, a bias towards the AR model can be favorable. In contrast, if  $n_{\text{gap}} < 0$  and the prediction error of AR is lower than for VAR, even though the estimation error of VAR is lower than for AR, such a bias would be unfavorable. In the following, we assess the relative performance of the 1SER and simply selecting the model with lowest prediction error, both on average and as a function of  $n$ .

In order to quantify the relative performance of both model selection strategies, we take the prediction and estimation errors of the 6000 VAR models estimated on  $n \in \{8, 9, \dots, 499, 500\}$  and for each model, and each  $n$ , select between the AR and VAR model in two different ways: (1) by simply selecting the model with the lowest prediction error, and (2) by applying the 1SER (using the standard-deviation of the out-of-sample prediction error across 100 training sets). For each of the two strategies, we then subtract the estimation error of the selected model ( $EE_{\text{sel}}$ ) from the estimation error of the model with the lowest estimation error ( $EE_{\text{best}}$ ). The difference  $EE_{\text{diff}} = EE_{\text{best}} - EE_{\text{sel}}$  equals zero if the model with lower estimation error has been selected, and is negative if the model with higher estimation error has been selected. Subsequently, we compute

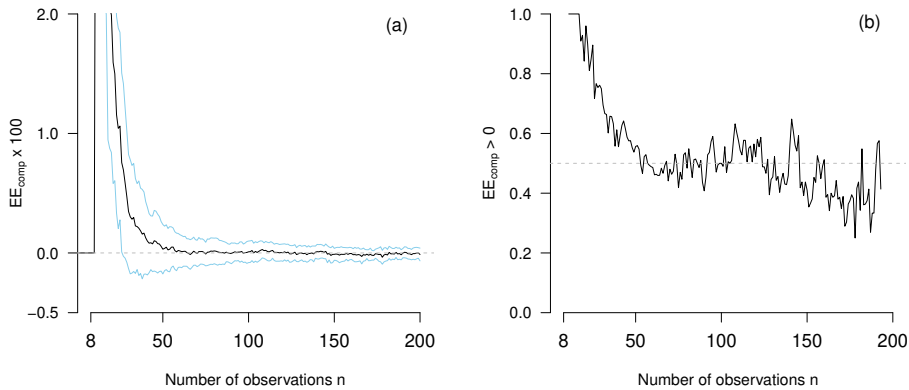
$$EE_{\text{comp}} = EE_{\text{diff}}^{(2)} - EE_{\text{diff}}^{(1)},$$



where  $EE_{diff}^{(2)}$  is the difference obtained using (2), and  $EE_{diff}^{(1)}$  is the difference obtained using (1). The resulting value of  $EE_{comp}$  allows us to compare the performance of the two model selection strategies. That is, if  $EE_{comp} < 0$ , simply selecting the model with lowest prediction error performs better, and if  $EE_{comp} > 0$ , the 1SER performs better.

Figure 7.4 (b) shows the distribution of  $EE_{comp}$  across all 6000 VAR models, averaged over iterations, and weighted by the probability given by the original mixed model. The only interesting cases when comparing model selection procedures are the cases in which they disagree. Therefore, we analyze only those cases for which  $EE_{comp} \neq 0$ . We find that using the 1SER is better in 50.1% of cases (obtained using on a log-spine density estimate on the weighted histogram). This would suggest that it makes essentially no difference whether we use the 1SER or select the model with lowest prediction error. However, these proportions average over the number of observations  $n$  and therefore cannot reveal differences in relative performance for different sample sizes.

Figure 7.5 (a) shows  $EE_{comp}$  as a function of  $n$ , averaged across all 6000 models.



**Figure 7.5:** Panel (a) displays  $EE_{comp}$  averaged across 6000 models as a function of  $n$  (black line) and the standard deviation around the average (blue line). Panel (b) displays, for each  $n$ , the proportion of times that  $EE_{comp} > 0$  across 6000 models (i.e., the proportion of 1SER performing better).

Because the VAR prediction error is huge for very small  $n$ , both model selection strategies choose the same model, resulting in  $EE_{comp} = 0$  for those  $n$ . However, from around  $n = 10$  on until around  $n = 60$ ,  $EE_{comp}$  is substantially positive, indicating that the 1SER outperforms simply selecting the model with the lowest prediction error by a large margin. However, for  $n > 60$  we see that  $EE_{comp}$  approaches zero and then becomes slightly negative. The latter is also illustrated in panel (b), which displays the weighted proportion of models in which the 1SER is better (i.e.,  $EE_{comp} > 0$ ). The explanation of this curve has three parts. First,  $n_{gap}$  tends to be *larger* if the gap is located at a *small*  $n$  (Pearson correlation

$r = -0.16$ ). If  $n_{\text{gap}}$  is large (and therefore positive), the AR model has lower estimation error than the VAR model, even though the prediction errors are the same (compare Figure 7.5 (b)). In such situations, biasing model selection towards selecting the AR model is advantageous. Since the 1SER constitutes a bias towards the AR model, it performs better for small  $n$ . Second, this also explains why the 1SER performs worse than simply selecting the model with lowest prediction error for large  $n$ : here the gap is small (negative), indicating that if the prediction errors are the same, the VAR model performs better. Clearly, in such a situation, providing a bias towards AR is disadvantageous. Therefore, the 1SER performs worse. Finally, why does the curve get closer and closer to zero? The reason is that the standard error converges to zero with (the square root of) the number of observations, and therefore the probability that both rules select the same model approaches 1 as  $n$  goes to infinity.

To summarize, we found that the 1SER is better than simply selecting the model with the lowest prediction error only in 50.1% of the cases in which the two rules do not select the same model. However, when looking at the relative performance as a function of  $n$ , we found that the 1SER is better than selecting the model with lowest prediction error until around  $n = 60$ , and worse above. Finally, we were able to explain the dependence of the relative performance on  $n$  with the fact that  $n_{\text{gap}}$  is larger when it occurs at a smaller  $n$ . For applied researchers these results suggest that, for VAR models with  $p = 6$  variables, the 1SER should be applied for  $n < 60$ .

## 7.4 Discussion

In this paper we provided an extended analysis of the problem studied by Bulteel et al. (2018) by using a simulation study to (a) map out the relative performance of AR and VAR models in typical psychological applications as a function of the number of observations  $n$ , and (b) investigate how to choose between AR and VAR models in practice. We found that, averaged over all models considered in our simulation, the VAR model outperforms the AR model for  $n > 92$  observations in terms of estimation error. In addition, we show that and explain why the 1SE rule proposed by Bulteel et al. (2018) performs better than selecting the model with the lowest prediction error when  $n$  is small.

Next to the *average* estimation errors of AR and VAR models, we also investigated the *variance* around those averages. We decomposed this variance in variance due to different true VAR models, and variance due to sampling. The variance across different VAR models showed that the relative performance, that is, the  $n$  at which VAR becomes better than AR ( $n_e$ ) depends on the characteristics of the true VAR parameter matrix  $\Phi$ . For example, if the true VAR model is very close to an AR model, it takes more observations until the VAR model outperforms the AR model. This shows that one cannot expect reliable recommendations with respect to  $n_e$  that ignore the characteristics of the generating model. The size of the sampling variation indicates that, for many of the considered sample sizes, whether the VAR or AR model will have lower estimation error largely

depends on the specific sample at hand. This implies that it is difficult to select the model with lowest estimation error with the sample sizes available in typical psychological applications.

The second question was: how should one choose between the AR and VAR model for a given data set? Bulteel et al. (2018) suggest that, for any VAR model, the  $n$  at which the prediction errors of both models are equal is, in expectation, (approximately) the same  $n$  at which their estimation errors are equal (i.e.,  $n_{\text{gap}} \approx 0$ ). Combining this claim with a preference towards the more parsimonious AR model, they proposed using the “1 Standard Error Rule”, according to which one should select the AR model if its prediction error is not more than one standard error above the prediction error of the VAR model, and choose the model with lowest prediction error otherwise. We showed that the expected  $n_{\text{gap}}$  varies as a function of the parameter matrix of the true VAR model. Using the relationship between estimation and prediction error we were able to explain when the 1SER is expected to perform better than selecting the model with lowest prediction error. In addition, we show via simulation that the 1SER performs better than selecting the model with the lowest prediction error for  $n < 60$ .

The relative performance of the AR and VAR model shown in our simulations can be understood in terms of the bias-variance trade-off. Because the AR model sets all off-diagonal elements to zero, it has a bias that is constant and independent of  $n$ . In contrast, the VAR model has a bias of zero, since the true model is a VAR model. This is why a VAR model will always perform better than (or at least as good as, if the all off-diagonal elements of the true VAR model are zero) an AR model as  $n \rightarrow \infty$ . However, for finite sample sizes the variance of the estimates of the two models are different: while both variances converge to zero as  $n \rightarrow \infty$ , for finite samples the variance of VAR parameters is much larger than the variance of AR parameters, especially for small  $n$ . This allows for the situation that the biased simpler model is showing a smaller error, even though the true model is in the class of the complex model. This trade-off between bias and variance also explains the relative performance of AR and VAR models: In Figure 7.3 we see that for small  $n$ , the variance of the VAR estimates is so large that the error is larger than the error of the AR model, despite the bias of the AR model. However, with increasing  $n$ , the variance of the estimates of both models approaches zero. This means that the larger  $n$ , the more the bias of the AR model contributes to its error. Thus, at some  $n$  the error of the VAR model becomes smaller than the error of the AR model. We agree with Bulteel et al. (2018) that the fact that a simple (and possibly implausible) model can outperform a complex (and more plausible) model, even though the true model is in the class of the more complex model, is underappreciated in the psychological literature.

An interesting question we did not discuss in our paper is: which model should we choose if the AR and VAR models have equal estimation error? Since we defined the quality of a model by its estimation error, we could simply pick one of the two models at random. However, their model parameters are likely to be very different. The estimation error of the AR model comes mostly from setting off-diagonal elements incorrectly to zero, while the estimation error of the VAR model comes mostly from incorrectly estimating off-diagonal elements. In

terms of the types of errors (false positive/negative) produced by the two models, the AR model will almost exclusively produce false negatives, while the VAR model will produce almost exclusively false positives. A specification of the cost of false positives/negatives in a given analysis may allow to choose between models when the estimation errors are the same or very similar. For example, in an exploratory analysis one might accept more false positives in order to avoid false negatives.

Throughout the paper we compared the AR model to the VAR model. However, we believe that it is unnecessarily restricting to choose only between those extremes (all off-diagonal elements zero vs. all off-diagonal elements nonzero). Instead, one could estimate VAR models with a constraint that limits the number of nonzero parameters or penalizes their size (see e.g., Fan & Li, 2001; Hastie et al., 2015). This would allow the recovery of large off-diagonal elements without the high variance of estimates in the standard VAR model. Similarly, one could estimate a VAR model and, instead of comparing it to an AR model and thus testing the nullity of the off-diagonal elements jointly, test the nullity of the off-diagonal elements of the VAR matrix individually.

It is important to keep the following limitations of our simulation study in mind. First, we claimed that the 6000 models we sampled from the mixed model obtained from the “MindMaastricht” data represent typical applications in psychology. One could argue that there are sets of VAR models that are plausible in psychological applications that are not included in our set of models. While this is a theoretical possibility, we consider this extremely unlikely, since we heavily sampled the mixed model stratified for  $R$  and  $D$ . Any VAR model that is not similar to a model in our set of considered VAR models is therefore most likely non-stationary. When presenting our results we weighted all models by their frequency given the estimated mixed model in order to avoid giving too much weight to unusual VAR models. This means that it could be that the weighting obtained from the mixed model does not well represent the frequency of VAR models in psychological applications. While we consider this unlikely, we also used a uniform weighting across VAR models as a robustness check which left all main conclusions unchanged. A second limitation is that we only considered VAR models with  $p = 6$  variables. While this is not a shortcoming compared to Bulteel et al. (2018) who use VAR models with 6, 6, and 8 variables, the results shown in the present paper would likely change when considering more than six variables. Specifically, we expect that the  $n$  at which VAR outperforms AR becomes larger across all settings. Similarly, we would expect that the 1SER outperforms selecting the model with lowest prediction error for sample sizes larger than 60. While the exact values will change for larger  $p$ , we expect that the general relationships between  $n$ ,  $R$ , and  $D$  extend to any number of variables  $p$ .

Although Bulteel et al. (2018) also consider mixed VAR and AR models, in the simulation studies presented above we focus exclusively on single-subject ( $N = 1$ ) time-series for simplicity. Mixed models can be seen as a form of regularization, in which individual parameter estimates are shrunk towards the group-level mean if the number of observations  $n$  is small. One would expect that for small  $n$ , the use of mixed models would improve the estimation and prediction

errors of both models, which is also what Bulteel et al. (2018) report in their results. Indeed, mixed models may improve the performance of VAR methods relative to AR. The reason is that the differential performance of AR and VAR models can be understood in terms of a bias-variance trade-off, where AR models are biased but have lower variance than VAR methods. The use of mixed VAR models should decrease this variance through shrinkage in small  $n$  settings (e.g., Efron & Morris, 1977; Gelman, 2006). The precise effect of using mixed models depends on the variance of parameters across individuals; however, we do not expect the general pattern of results reported here to change when moving from single-subject to mixed settings.

Future research could extend the analysis shown here to VAR models with number of variables larger than  $p = 6$ , which would allow to generalize the simulation results to more situations encountered in psychological applications. Another interesting avenue for future research would be to investigate the link between  $n_{\text{gap}}$  and the VAR parameter matrix  $\Phi$ . Since  $n_{\text{gap}}$  has direct implications for model selection, such a link could possibly be used to construct improved model selection procedures. Finally, it would be useful to extend the simulation study in this paper to constrained estimation such as the LASSO, especially since those methods are already applied in practice (see e.g., Epskamp, Waldorp, et al., 2018).

To sum up, we studied the relative performance of AR and VAR models in typical psychological applications. We were able to make clear statements about the average performance, however, the variance around averages is considerable. Decomposing this variance showed that (i) one cannot expect reliable statements with respect to the relative performance of the AR and VAR models that ignore the characteristics of the generating model, and (ii) that choosing reliably between AR and VAR models is difficult for most sample sizes typically available in psychological research. Finally, we provided a theoretical explanation for when the “1 Standard Error Rule” outperforms simply selecting the model with lowest prediction error, and showed that the 1SER performs better when  $n$  is small.

## Acknowledgements

We would like to thank Don van den Bergh, Riet van Bork, Denny Borsboom, Max Hinne and Lourens Waldorp for their helpful comments on earlier versions of this paper.



# THE INPUT MATTERS: INTERPRETING THE ISING MODEL

---

## Abstract

The Ising model is a model for pairwise interactions between binary variables that has become popular in the psychological sciences. It has been first introduced as a theoretical model for the alignment between positive (1) and negative (-1) atom spins. In many psychological applications, however, the Ising model is defined on the domain  $\{0,1\}$  instead of the classical domain  $\{-1,1\}$ . While it is possible to transform the parameters of the Ising model in one domain to obtain a statistically equivalent model in the other domain, the parameters in the two versions of the Ising model lend themselves to different interpretations and imply different dynamics, when studying the Ising model as a dynamical system. In this tutorial, we provide an accessible discussion of the interpretation of threshold and interaction parameters in the two domains and show how the dynamics of the Ising model depends on the choice of domain. Finally, we provide a transformation that allows one to transform the parameters in an Ising model in one domain into a statistically equivalent Ising model in the other domain.

## 8.1 Introduction

The Ising model is a model for pairwise interactions between binary variables that originated in statistical mechanics (Ising, 1925; Glauber, 1963) but is now used in a large array of applications in the psychological sciences (e.g., Borsboom & Cramer, 2013; Marsman, Maris, Bechger, & Glas, 2015; Boschloo et al., 2015; Boschloo, Schoevers, van Borkulo, Borsboom, & Oldehinkel, 2016; Fried et al., 2015; Cramer et al., 2016; Dalege et al., 2016; Rhemtulla et al., 2016; Van Der Maas, Kan, Marsman, & Stevenson, 2017; Haslbeck & Fried, 2017; Afzali et al., 2017; Deserno et al., 2017; Savi, van der Maas, Maris, et al., 2018; Marsman, Tanis, Bechger, & Waldorp, 2019)

The original Ising model has been introduced as a model for the interactions between atom spins, which can be positive (1) and negative (-1) (Brush, 1967). In this setting, with variables taking values in the domain  $\{-1, 1\}$ , the interaction parameters in the Ising model determine the *alignment* between variables: If an interaction parameter between two variables is positive, the two variables tend to take the same value; on the other hand, if the interaction parameter is negative, the two variables tend to take different values.

In most psychological applications, however, the Ising model is defined with variables taking values in the domain  $\{0, 1\}$ . While it is possible to transform the parameters of a given Ising model in one domain to obtain a statistically equivalent model in the other domain, the parameters in the two versions of the Ising model lend themselves to different interpretations and imply different dynamics, when studying the Ising model as a dynamical system. If unaware of those subtle differences, one might erroneously apply theoretical results from the  $\{-1, 1\}$  domain to an estimated model in the  $\{0, 1\}$  domain, or simply interpret parameters incorrectly. To prevent such confusion in the emerging psychological networks literature which makes heavy use of the Ising model, we provide a detailed discussion of both versions of the Ising model in the present tutorial.

We begin by discussing the different interpretations of the Ising model in the  $\{-1, 1\}$  and  $\{0, 1\}$  domain in Section 8.2, using a simple two variable example which allows the reader to follow all calculations while reading. We explain the differences in the interpretation of the threshold and interaction parameters in the two versions of the Ising model, and discuss in which situation which version might be more appropriate. While most psychological applications of the Ising model use it as a statistical model, it has also been studied as a dynamical system in psychological research (e.g., Cramer et al., 2016; Dalege et al., 2016; Lunansky, van Borkulo, & Borsboom, 2019). In Section 8.3 we discuss how the dynamics of the Ising model depends on the choice of domain, and show that the domain changes the *qualitative* behavior of the model. Finally, in Section 8.4 we provide a transformation that allows one to transform the parameters in an Ising model in one domain into a statistically equivalent Ising model in the other domain.



## 8.2 Different Domain, Different Interpretation

In this section we estimate an Ising model with  $p = 2$  variables in both domains,  $\{-1, 1\}$  and  $\{0, 1\}$ , and show that the resulting threshold and interaction parameters have different values and lend themselves to different interpretations. We choose the  $p = 2$  variable case to make the explanation as accessible as possible. However, all results immediately extend to the general situation with  $p$  variables. The Ising model for two variables is given by

$$P(y_1, y_2) = \frac{1}{Z} \exp\{\alpha_1 y_1 + \alpha_2 y_2 + \beta_{12} y_1 y_2\}, \quad (8.1)$$

where  $y_1, y_2$  are either elements of  $\{-1, 1\}$  or  $\{0, 1\}$ ,  $P(y_1, y_2)$  is the probability of the event  $(y_1, y_2)$ ,  $\alpha_1, \alpha_2, \beta_{12}$  are parameters in  $\mathbb{R}$ , and  $Z$  is a normalization constant which denotes the sum of the exponent across all possible states. There are  $2^p = 4$  states in this example.

To illustrate the differences across models, we generate  $n = 1000$  samples of the labels  $A, B$  with the relative frequencies shown in Table 8.1:

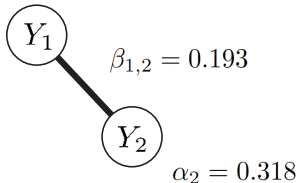
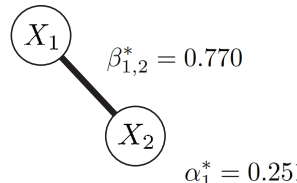
	A	B
A	0.14	0.18
B	0.18	0.50

**Table 8.1:** Relative frequency of states in the example data set.

In applications, the labels  $A, B$  can stand for any pair of categories such as being for or against something, some event having happened or not, or a symptom being present or not. The two domains are two different ways to numerically represent these labels.

We obtain the Maximum Likelihood Estimates (MLE) of the parameters in two different ways: once, by filling in  $\{-1, 1\}$  for  $\{A, B\}$ ; and once by filling in  $\{0, 1\}$  for  $\{A, B\}$ . Figure 8.1 summarizes the two resulting models. The first column in Figure 8.1 shows the parameter estimates  $\alpha_1, \alpha_2$  and  $\beta_{12}$ , and log potentials in domain  $\{-1, 1\}$ . We first focus on the interpretation of the interaction parameter  $\beta_{12}$ . To understand the interpretation of this parameter we take a look at the log potentials for all four states  $\{(-1, -1), (-1, 1), (1, -1), (1, 1)\}$ , which we obtain by plugging the four states into the expression within the exponential in equation (8.1). The resulting log potentials are displayed in the second row in Figure 8.1 and show us the following: if  $\beta_{12}$  becomes larger, the probability of the states  $(-1, -1), (1, 1)$  increases relative to the probability of the states  $(-1, 1), (1, -1)$ . This means that the interaction parameter determines the degree of *alignment* of two variables. That is, if  $\beta_{12} > 0$  the *same* labels align with each other, and if  $\beta_{12} < 0$  *opposite* labels align with each other. In other words,  $\beta_{12}$  models the probability of the states  $(-1, -1), (1, 1)$  relative to the probability of the states  $(-1, 1), (1, -1)$ .

This is not the case in the  $\{0, 1\}$  domain. The second column in Figure 8.1 shows that the parameter estimates  $\alpha_1^*, \alpha_2^*, \beta_{12}^*$  in domain  $\{0, 1\}$ , and we see that

Domain	$y \in \{-1, 1\}$	$x \in \{0, 1\}$
Model Parameters	$\alpha_1 = 0.318$  $\beta_{1,2} = 0.193$ $\alpha_2 = 0.318$	$\alpha_2^* = 0.251$  $\beta_{1,2}^* = 0.770$ $\alpha_1^* = 0.251$
Log Potentials	$(Y_1 = -1, Y_2 = -1) : -\alpha_1 - \alpha_2 + \beta_{1,2}$ $(Y_1 = 1, Y_2 = -1) : \alpha_1 - \alpha_2 - \beta_{1,2}$ $(Y_1 = -1, Y_2 = 1) : -\alpha_1 + \alpha_2 - \beta_{1,2}$ $(Y_1 = 1, Y_2 = 1) : \alpha_1 + \alpha_2 + \beta_{1,2}$	$(X_1 = 0, X_2 = 0) : 0$ $(X_1 = 1, X_2 = 0) : \alpha_1^*$ $(X_1 = 0, X_2 = 1) : \alpha_2^*$ $(X_1 = 1, X_2 = 1) : \alpha_1^* + \alpha_2^* + \beta_{1,2}^*$

**Figure 8.1:** The threshold and interaction parameters estimated from the data generated from Table 8.1, and the log potentials for each combination of states, separately for the two domains  $\{-1, 1\}$  and  $\{0, 1\}$ . The log potentials (also called energy function or Hamiltonian) are obtained by filling each state (e.g.,  $y_1 = -1, y_2 = -1$ ) in the expression within the exponential in equation (8.1).

they have different values than in the  $\{-1, 1\}$  domain. To understand why this is the case, we again look at the interpretation of the interaction parameter  $\beta_{12}^*$  by inspecting the four log potentials. The key observation is that  $\beta_{12}^*$  only appears in the log potential of the state  $(1, 1)$ . What happens if  $\beta_{12}^*$  increases? Then the probability of the state  $(1, 1)$  increases relative to the probability of all other states  $(0, 1), (1, 0), (0, 0)$ . In other words,  $\beta_{12}^*$  models the probability of state  $(1, 1)$  relative to the probability of the states  $(0, 1), (1, 0), (0, 0)$ .

Next, we turn to the interpretation of the threshold parameters. If all interaction parameters are equal to zero, the threshold parameters in both domains indicate the tendency of a variable to be in one state or the other. That is,  $\alpha, \alpha^* > 0$  implies a larger probability for the states  $(1) \in \{0, 1\}, (1) \in \{-1, 1\}$  than for states  $(0) \in \{0, 1\}, (-1) \in \{-1, 1\}$ . If  $\alpha, \alpha^* < 0$  the reverse is true, and if  $\alpha, \alpha^* = 0$ , the corresponding states have both probability 0.5. However, in the general case in which interaction parameters are allowed to be nonzero, the interpretation depends on the domain: in the  $\{-1, 1\}$  domain the threshold parameter indicates the tendency of a variable averaged over all possible states of all other variables. In more formal terms, the threshold parameter of a given variable indicates the marginal mean of that variable. In contrast, the threshold in the  $\{0, 1\}$  domain indicates the tendency of a variable when all other variables are equal to zero. We return to the different interpretations of thresholds in Section 8.3, in which we discuss the dynamics of the Ising model.

In this section we showed that depending on its domain, the parameters of the Ising model have different interpretations. What are the consequences for applied researchers? In terms of reporting, it is important to state which domain has been used such that the reported model can be re-used in the correct way: if someone reports a set of parameters estimated from in the  $\{0, 1\}$  domain, and a reader applies it to the  $\{-1, 1\}$  domain they will obtain the incorrect probabilities. Note that in order to use the model one also has to report the threshold parameters. Not reporting the threshold parameters is a common problem and irrespective of the issue discussed in this chapter. The only situation in which the domain does not matter is if the only goal is to compare the relative size of interaction parameters since the relative size is the same in both domains (see Section 8.4).

The second consequence is that researchers have to choose which version of the Ising model is more appropriate for the phenomenon at hand. The above clarified interpretations of the Ising model in its two different domains allow to take this decision. For example, the  $\{-1, 1\}$  parameterization may be more plausible for labels that are not qualitatively different, but rather opposing each other in some way such as supporting or opposing a certain viewpoint, for example agreeing or disagreeing with a statement like “Elections should be held every two years instead of every four years”. This also reflects the origin of the Ising model as a model for atom spins, which are either positive or negative. The parameterization implied by  $\{0, 1\}$  could be more appropriate if the two labels are qualitatively different, such as the presence or absence of an event or a characteristic. Take psychiatric symptoms as an example: while it seems plausible that *fatigue* leads to *lack of concentration*, it is less clear whether the absence of *fatigue* also leads to the increase of *concentration*. In such a case, we can encode the possible belief that the absence of something cannot have an influence on something else by choosing the  $\{0, 1\}$  domain. Importantly, the decision of which version to pick has to be based on information beyond the data, because the models are statistically equivalent and therefore cannot be distinguished by observational data. In Appendix E.1 we prove this equivalence for the example shown in Figure 8.1.

While Ising models in psychological research are usually fit to cross-sectional data, one is typically interested in within-subjects dynamics. In this context, one is often interested in inferring from an estimated Ising model how to best intervene on the system. In the next section we will show how the dynamics of the Ising model depends on its domain, and that the different versions of the Ising model make different predictions for optimal interventions.

### 8.3 Different Domain, Different Dynamics

The choice of domain also determines the dynamics of the Ising model, when studying it as a dynamical system describing within-person dynamics. The dynamical version of the Ising model is initialized by  $p$  initial values at  $t = 1$ , and then each variable at time  $t$  is a function of all variables it is connected to via a

nonzero interaction parameter at  $t - 1$ <sup>1</sup>. An often studied characteristic in this model is how its behavior changes when the size of the interaction parameters increases. A typical behavior of interest is the number of variables in state (1) (e.g., Dalege et al., 2016; Cramer et al., 2016).

Which behavior would we expect in the two domains  $\{-1, 1\}$  and  $\{0, 1\}$ ? From the previous section we know that in domain  $\{-1, 1\}$ , the interaction parameter  $\beta_{ij}$  models the probability of states  $\{(-1, -1), (1, 1)\}$  relative to the states  $\{(-1, 1), (1, -1)\}$ . Now, when increasing all  $\beta_{ij}$ , connected variables become more synchronized, which means that all (connected) variables tend to be either *all* in state (-1) or (1). In terms of number of variables in state (1), we would therefore predict that the expected number of variables in state (1) remains unchanged, because the states (-1) and (1) occur equally often in the aligned  $((-1, -1)$  and  $(1, 1))$  and not aligned  $((-1, 1)$  and  $(1, -1))$  states. And second, we predict that the probability that at a given time point either all variables are in state (-1), or all variables are in state (1), increases. The reason is that, in the  $(-1, 1)$  domain, the larger the interaction parameter, the stronger the alignment between variables. This second prediction implies that the variance of the number of states in (1) increases.

In the domain  $\{0, 1\}$ , the interaction parameter  $\beta_{ij}^*$  models the probability of the state (1, 1) relative to the remaining three states  $\{(0, 1), (1, 0), (0, 0)\}$ . Now, when increasing  $\beta_{ij}^*$ , connected variables will have a higher probability to be all in state 1. Importantly, the frequency of 1s in the high probability state (1, 1) is higher than in the other three states. We therefore expect that the number of variables in state (1) increases and that the probability that all variables are in state (1) increases. The second prediction implies that the variance of the number of states in (1) decreases.

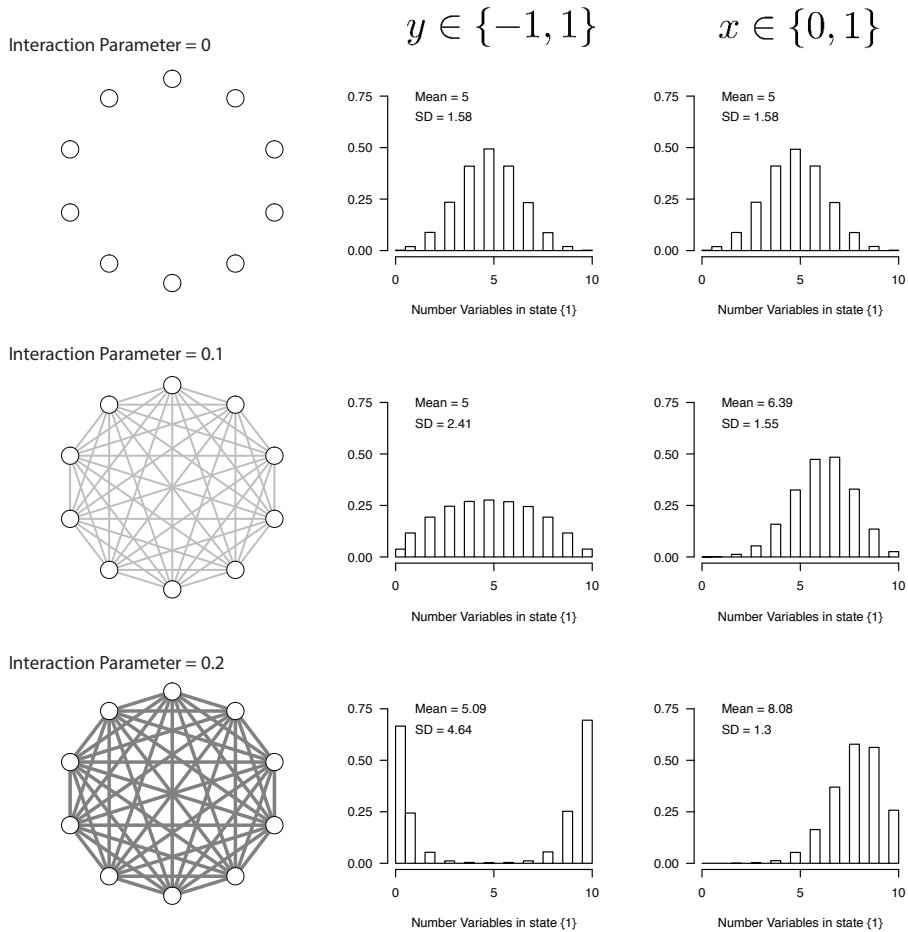
We prove that the expected number of variables in state (1) remains unchanged for  $\{-1, 1\}$  and increases for  $\{0, 1\}$ , if  $\beta_{ij} > 0$  for the case  $p = 2$  variables in Appendix E.2. Here, we show via simulation that our predictions are correct. We sample  $n = 10^6$  observations from a fully connected (i.e., all interaction parameters are nonzero) Ising model with  $p = 10$  variables in which all edge weights (interaction parameters) have the same size and all thresholds are set to zero. We vary both the size of the interaction parameters  $\beta_{ij} \in \{0, 0.1, 0.2\}$  and the domain<sup>2</sup>. Figure 8.2 shows the distribution (over time steps) of the number of variables that are in state {1}.

The first row of Figure 8.2 shows the distribution of the number of variables in state (1) across time when all interaction parameters are equal to zero. We see a symmetric, unimodal distribution with mean 5 for both domains. This is what we would expect since the probability of each variable being in state (1) can be seen as an unbiased (because the thresholds are zero) coin flip that is independent of all other variables. Thus, since we have 10 variables, the means are equal to  $10 \times 0.5 = 5$ .

---

<sup>1</sup>Glauber dynamics (Glauber, 1963) describe a different way to sample from a dynamic Ising model. The qualitative results presented in this section also hold for Glauber dynamics.

<sup>2</sup>The code to reproduce the simulation and Figure 8.2 is available at <http://github.com/jmbh/IsingVersions>.



**Figure 8.2:** The distribution of the number of variables being in state one as a function of the size of the interaction parameter in a fully connected Ising model  $\{0, .1, .2\}$  and used domain  $\{-1, 1\}$  and  $\{0, 1\}$  of the random variable.

However, when increasing the interaction parameter from 0 to 0.1 (second row) the distributions become different: in domain  $\{-1, 1\}$  the mean remains unchanged and the probability mass shifts from around 5 to more extreme values, resulting in increased variance. In contrast, in domain  $\{0, 1\}$  the distribution shifts to the right, which implies that the mean increases and the variance slightly decreases. When further increasing the interaction parameters to 0.2 (third row), in domain  $\{-1, 1\}$  most of the probability mass is concentrated on 0 and 10, while leaving the mean unchanged; in domain  $\{0, 1\}$  the mean further increases and the variance further decreases. From a dynamical perspective, this means that for strongly connected Ising models (with thresholds equal to zero) the domain  $\{-1, 1\}$  implies two stable states (all variables in state  $(-1)$  or  $(1)$ ), while the do-

main  $\{0, 1\}$  implies only a single stable (all variables in state (1)), whose position depends on whether interaction parameters are positive or negative. This means that the dynamic Ising model in the  $\{-1, 1\}$  can switch between stable states, while  $\{0, 1\}$  it always stays in the same stable state<sup>3</sup>.

For the general case of Ising models that are not fully connected and also have negative interaction parameters, the results described above extend local clusters of two or more variables: in the domain  $\{-1, 1\}$ , increasing the interaction parameter will leave the means of all variables in the cluster unchanged, however, the variables become increasingly aligned (if interaction parameters are positive) or disaligned (if interaction parameters are negative). Alignment will lead to an increase in variance, while disalignment will lead to a decrease in variance. In contrast, in the  $\{0, 1\}$  domain the mean of all variables in the cluster will increase in the case of positive interaction parameters, and decrease in the case of negative interaction parameters.

This shows that, depending on which domain is used one can come to entirely different conclusions about the dynamics of the Ising model. For example, (Cramer et al., 2016) model the interactions between psychiatric symptoms with an Ising model in domain  $\{0, 1\}$  and conclude that densely connected Ising models imply a larger number of active (in state (1)) symptoms and therefore represent “pathological” models. The above argument and simulation show that this is only true when using the  $\{0, 1\}$  domain, which encodes the belief that the absence of a symptom cannot influence the absence of another symptom. If one decides that an alignment between variables is a more plausible interaction (as implied by the  $\{-1, 1\}$  domain), then densely connected Ising models do not imply a large number of active symptoms. Instead, high density implies high variance and two stable states. Thus, the characterization of dense networks as pathological networks as in (Cramer et al., 2016) hinges on choosing the  $\{0, 1\}$  domain.

This has important consequences: when choosing the  $\{0, 1\}$  domain, we would conclude that highly connected symptom networks are necessarily “bad”, and interventions on the interactions between symptoms as suggested by (Borsboom, 2017) should always reduce symptom activation. On the other hand, in the  $\{-1, 1\}$  domain highly connected symptom networks are not necessarily bad, but in fact can lead to high resilience, if the threshold parameters are large negative values. In such a situation strong interactions would keep the system in a state in which all symptoms are deactivated.

## 8.4 Transforming from $\{-1, 1\}$ to $\{0, 1\}$ and vice versa

The Ising model is typically estimated by a sequence of  $p$  logistic regressions, which require the domain  $\{0, 1\}$ . However, the previous sections showed that in some situations the domain  $\{-1, 1\}$  may be more appropriate. In Table 8.2 we

---

<sup>3</sup>The result about bistability is true for the considered fully connected Ising model with zero thresholds. It is also possible to construct a bistable Ising model in the  $\{0, 1\}$  domain by choosing large negative thresholds and large positive interaction parameters. The relationship between mean/variance and changing the interaction parameter in the two domains, however, is always true.

present a transformation that allows one to obtain the parameterization based on domain  $\{-1, 1\}$  from the parameterization based on domain  $\{0, 1\}$  and vice versa (see Appendix E.3 for the derivation of the transformations). We define  $\beta_{i+}^* = \sum_{j=1}^p \beta_{ij}^*$  as the sum over the interaction parameters associated with a given variable  $y_i$ .

Transformation	Thresholds	Interactions
$\{0, 1\} \Rightarrow \{-1, 1\}$	$\alpha_i = \frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*$	$\beta_{ij} = \frac{1}{4}\beta_{ij}^*$
$\{-1, 1\} \Rightarrow \{0, 1\}$	$\alpha_i^* = 2\alpha_i - 2\beta_{i+}$	$\beta_{ij}^* = 4\beta_{ij}$

**Table 8.2:** Transformation functions to obtain the threshold and interaction parameters of one parameterization from the threshold and interaction parameters of the other parameterization. Parameters with asterisk indicate parameters in the  $\{0, 1\}$  domain.

Table 8.2 shows that the interaction parameters  $\beta_{ij}$  in the  $\{-1, 1\}$  domain are 4 times smaller than the interaction parameters  $\beta_{ij}^*$  in the  $\{0, 1\}$  domain. We also see that the threshold parameter  $\alpha_i$  is a function of *both* the threshold and the interaction parameters  $\alpha_i^*, \beta_{ij}^*$  in the other parameterization.

We now apply the transformations in Table 8.2 to the  $p = 2$  variable example in Figure 8.1. We choose to transform from  $\{0, 1\}$  to  $\{-1, 1\}$ :

$$\begin{aligned} a_1 &= \frac{1}{2}a_1^* + \frac{1}{4}\beta_{i+}^* = \frac{0.251}{2} + \frac{0.77}{4} = 0.318 \\ \beta_{12} &= \frac{1}{4}\beta_{12}^* = \frac{.77}{4} = 0.1925 \approx 0.193 \end{aligned}$$

And indeed, we obtain the parameters obtained when estimating the Ising model in the  $\{-1, 1\}$  domain (see first column in Figure 8.1).

From the transformation in Table 8.2 follows that the two models are statistically equivalent. This implies that one could also estimate the model in the  $\{-1, 1\}$  domain, transform the parameters, and would obtain the parameters one would have obtained from estimating in the  $\{0, 1\}$  domain. Also, note that the standard errors of estimates are subject to the same transformation, and therefore one always reaches the same conclusion regarding statistical significance in both domains.

However, note that one does not necessarily arrive at statistical equivalent models when estimating in the two different domains using biased estimators. An example of a biased estimation method is the popular  $\ell_1$ -regularized estimator (Van Borkulo et al., 2014). We discuss why statistical equivalence is not guaranteed in this specific example in Appendix E.4. The possibility that different domains lead to models that are not statistically equivalent highlights the importance of choosing the most plausible Ising model on substantive grounds.

## **8.5 Conclusions**

In this chapter we have investigated the subtleties in choosing the domain of the Ising model. We showed that estimating the Ising model in the domains  $\{0, 1\}$  and  $\{-1, 1\}$  results in parameters with different values and different interpretations. We also showed that the qualitative behavior of the dynamical Ising model depends on the chosen domain. Finally, we provided a transformation that explains the relation between the two parameterizations and allows one to obtain one from the other. This is useful in practice, because typically used software packages require the  $\{0, 1\}$  domain. This transformation also implies that the two parameterizations are statistically equivalent, which means that one cannot choose one over the other on empirical grounds. Thus, researchers should carefully reflect on which interactions between variables are plausible and choose the domain accordingly.

## **Acknowledgements**

We would like to thank Joost Kruis, Oisín Ryan, Fabian Dablander, Jonas Dalege and Joris Broere for helpful comments on earlier versions of this chapter.



**Part II**

**Formal Theories**



---

# RECOVERING BISTABLE SYSTEMS FROM TIME SERIES DATA

---

## Abstract

Conceptualizing mental disorders as complex dynamical systems has become a popular framework to study mental disorders. Especially bistable dynamical systems have received much attention, because their properties map well onto many characteristics of mental disorders. While these models were so far mostly used as stylized toy models, the recent surge in psychological time series data promises the ability to recover such models from data. In this chapter we investigate how well popular (e.g., the Vector Autoregressive model) and more advanced (e.g., differential equation estimation) data analytic tools are suited to recover bistable dynamical systems from time series. Using a simulated high-frequency time series (measurement every six seconds) as an ideal case we show that while it is possible to recover global dynamics (e.g., position of fixed points, transition probabilities) it is difficult to recover the microdynamics (i.e., moment to moment interactions) of a bistable system. Repeating all analyses with a sampling frequency typical for Experience Sampling Method studies (measurement every 90 minutes) showed that the recovery of the global dynamics was still successful, but no microdynamics could be recovered. These results raise two fundamental issues involved in studying mental disorders from a complex systems perspective: first, it is generally unclear what to conclude from a statistical model about an underlying complex systems model; and second, if the sampling frequency is too low, it is impossible to recover microdynamics. In response to these results we propose a new modeling strategy based on substantively plausible dynamical systems models.

## 9.1 Introduction

Conceptualizing mental disorders as complex dynamical systems has become a popular framework to study mental disorders (e.g., Wichers, Wigman, & Myin-Germeys, 2015; Cramer et al., 2016; Borsboom, 2017). This framework is attractive because it acknowledges the fact that many mental disorders are massively multifactorial (e.g., Kendler, 2019), and because it allows one to specify powerful within-person dynamical systems models that capture many of the characteristics hypothesized for mental disorders. The central goal of this framework is to obtain such models to further our understanding of mental disorders, and allow us to develop and test more successful interventions.

The class of dynamic systems that has received most attention in this emerging literature is the class of *bistable systems* (e.g., Wichers et al., 2015; Cramer et al., 2016; Borsboom, 2017; Wichers, Schreuder, Goekoop, & Groen, 2019; van de Leemput et al., 2014a; Nelson, McGorry, Wichers, Wigman, & Hartmann, 2017; Kalisch et al., 2019). The reason is that its behavior maps well on many phenomena observed in mental disorders: bistable systems describe variables that have two stable states, which can be interpreted as different psychological states such as “healthy” or “unhealthy” (e.g., depressed). The stability landscape reflecting the dynamics of the system determines how easy it is to transition from one state to the other, and thereby offers a possible formalization of properties of the mental disorder, such as vulnerability or resilience to developing it (Scheffer et al., 2018). Bistable systems can also show sudden transitions from one state to another, thereby mapping well on, for example, bipolar disorder or the phenomenon of sudden gains and losses in psychotherapy (Stiles et al., 2003; Lutz et al., 2013).

In parallel, the realization that inferences from between-subjects data to within-person data are only possible under stringent assumptions (Molenaar, 2004; Hamaker, 2012) together with the increasing availability of psychological time series collected from mobile devices has led to a surge in studies aiming at recovering the within-person dynamics associated with mental disorders (e.g., Bringmann et al., 2013; Pe et al., 2015; Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019). This is an exciting development, because within-person time series potentially allow one to recover bistable systems and other dynamical systems from empirical data. This would be a major step forward for studying mental disorders as complex systems, because so far these models were only used as stylized toy models.

However, so far there has been no systematic investigation into to what extent dynamical systems models can in fact be recovered from psychological time series. To investigate this question for a given dynamical system and data analytic method, it has to be broken down into two parts: The first question is whether the method at hand can recover (some aspect of) a dynamical system *in principle*, that is, with “ideal” data (long time series, extremely high sampling frequency). If this is the case, the second question is whether the method also works with realistic data (shorter time series, much lower sampling frequency). In the present chapter, we investigate both questions for a bistable dynamical system and a se-

lection of the most popular (e.g., the Vector Autoregressive (VAR) model; Hamilton, 1994) and some more advanced (e.g., differential equation estimation; Boker, Deboeck, Edler, & Keel, 2010) methods. Specifically, we use a basic bistable dynamical system for emotion dynamics to simulate both an ideal time series with extremely high sampling frequency (measurement every six seconds) and a more realistic time series with a sampling frequency common for Experience Sampling Method (ESM) studies (measurement every 90 minutes). Using these time series, we evaluate how useful each method is for recovering bistable dynamical systems *in principle*, and how useful it can be *in practice* when analyzing realistic ESM time series.

We will show that the popular VAR model (and the Gaussian Graphical Model fitted on its residuals; Epskamp, Waldorp, et al., 2018) is in principle unable to recover the global dynamics (e.g., location and variance of stable states, frequency of transitions) and succeeds only in recovering some of the microdynamics (moment-to-moment interactions) of the true bistable system. However, descriptive statistics, data visualization and more flexible statistical models are able to capture the global dynamics. The only method that recovered the complete bistable system is an iterative model building procedure that directly estimates the system of differential equations (DEs). Reducing the sampling frequency from every six seconds to every 90 minutes affects the considered methods differently: the VAR model and its extensions no longer recover any microdynamics, and the DE-estimation procedure fails. However, descriptives, data visualization and appropriate statistical models still recover the global dynamics. These results raise two fundamental issues involved in studying mental disorders from a complex systems perspective: first, it is generally unclear what to conclude from a statistical model about an underlying complex systems model; and second, if the sampling frequency is too low, it is impossible to recover microdynamics. In response to these findings, we outline a different research strategy to arrive at dynamical systems models for mental disorders: Proposing initial formal models which can subsequently be scrutinized and developed by deriving data implications that can be tested empirically. We will show that in this process many of the presented methods are instrumental to testing predictions of the formal model and thereby triangulating the formal model that captures the true dynamical system best.

This chapter is structured as follows. In Section 9.2 we introduce a simple bistable dynamical system for emotion dynamics, discuss its dynamics and characteristics, and describe how we generate the ideal and the more realistic time series from it. We use the ideal data (measurement every six seconds) in Section 9.3 to evaluate for each method to which extent it can recover a bistable dynamical system. Next, in Section 9.4 we evaluate the same methods but using the time series with a sampling frequency that matches typical ESM studies (measurement every 90min). Finally, in Section 9.5 we discuss the implications of our results for the framework of empirically studying mental disorders from a complex systems perspectives, and outline a new research strategy based on formal modeling, which avoids shortcomings of a purely data analytic approach.

## 9.2 Bistable Emotion System as Data Generating Model

In this section we present a bistable dynamical system and describe its dynamics (Section 9.2.1), show how we generate data from this system (Section 9.2.2) and discuss its qualitative characteristics (Section 9.2.3).

### 9.2.1 Model Specification

Bistable dynamical systems are typically formalized within the framework of differential equations (e.g., Hirsch, Smale, & Devaney, 2012; Strogatz, 2015) and we therefore also choose this framework. Our goal is to provide an accessible first investigation of how well bistable systems can be recovered from psychological time series and therefore use the one of the simplest multivariate bistable systems. Specifically, we choose a system with four variables that is a generalization of the classic Lotka Volterra model for competing species (e.g., H. I. Freedman, 1980) to four variables; a similar model was used by (van de Leemput et al., 2014a) who interpreted the four variables as positive and negative emotion variables, an interpretation we also adopt here. In an appropriate parameter regime, this model exhibits two stable states: one in which positive emotions are high and negative emotions are low (the “healthy” state); and one in which the positive emotions are low and the negative emotions are high (the “unhealthy” state).

Note that different types of (bistable) dynamical systems will differ in how difficult they are to recover with a given method and type of time series, and much research is needed to map out the space of dynamical systems model classes, data analytic methods and types of time series. However, the intuition we rely on in the present chapter is that if there are fundamental problems in recovering one of the simplest multivariate bistable systems, then these problems will be at least equally severe when recovering a more complicated bistable dynamical system.

The bistable system we use throughout this chapter consists of two emotions with positive valence (Cheerful ( $x_1$ ) and Content ( $x_2$ )) and two emotions with negative valence (Anxious ( $x_3$ ) and Sad ( $x_4$ )). The dynamics of the system is defined by the stochastic differential equations

$$\frac{dx_i}{dt} = r_i x_i + \sum_{j=1}^4 C_{i,j} x_j x_i + a_i + \epsilon_i , \quad (9.1)$$

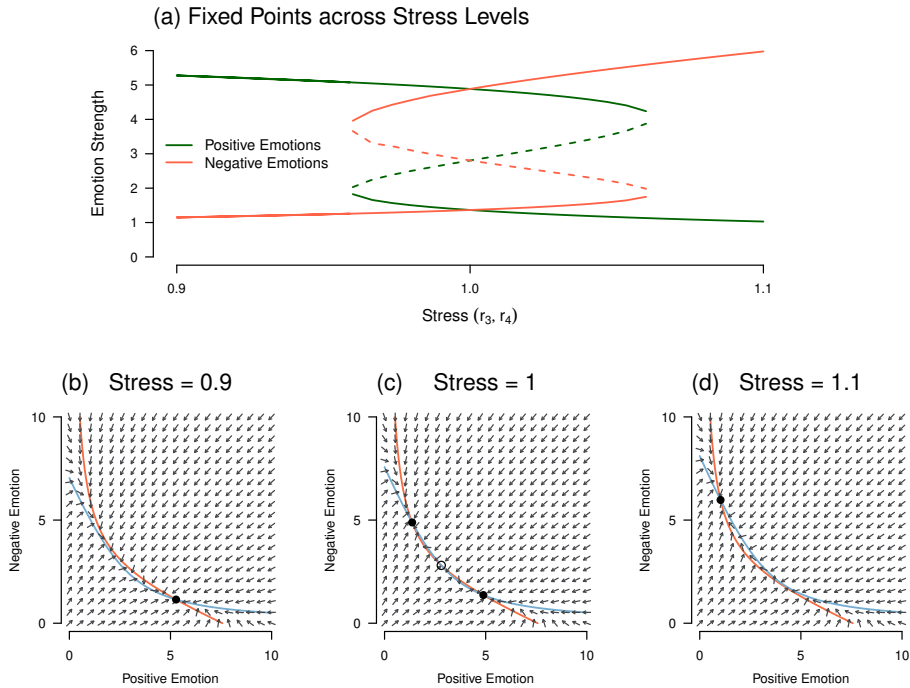
where  $r_i$  can be thought of as the main effect of an emotion on itself over time, that is, the effect of  $x_i$  on its own rate of change. This parameter is set to 1 for positive emotions, and will be varied between  $r_3, r_4 \in [0.9, 1.1]$  for negative emotions. We interpret the variations in  $r_3, r_4$  as being related to stress: higher stress means that the effects of a high degree of negative emotion stays in the system longer. The matrix  $C$  represents the dependencies between emotions in the form of interaction effects

$$C = \begin{bmatrix} -0.2 & 0.04 & -0.2 & -0.2 \\ 0.04 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & 0.04 \\ -0.2 & -0.2 & 0.04 & -0.2 \end{bmatrix}.$$

The interactions in the matrix  $C$  show that emotions of the same valence reinforce each other, while emotions of different valence suppress each other. For example,  $C_{1,2} = 0.04$  indicates that the rate of change of  $x_1$  (Cheerful) depends on the product of  $x_1$  and  $x_2$  (Content) weighted by 0.04. Similarly,  $C_{1,3} = -0.2$  indicates that the rate of change of  $x_1$  depends on the product of  $x_1$  and  $x_3$  (Anxious) weighted by  $-0.2$ . The diagonal elements are quadratic effects: for example,  $C_{1,1} = -0.2$  indicates that the rate of change of  $x_1$  depends on the product  $x_1 x_1 = x_1^2$ . Note that the matrix  $C$  is symmetric because we aim at choosing the simplest possible bistable system. In general, however, it does not have to be symmetric.

We interpret 0 as the absence of positive/negative emotion, and therefore do not allow emotions to become negative. We ensure this with high probability by setting the constant  $a_i = 1.6$  for all  $i$ . The Gaussian noise term  $\epsilon_i$  has a mean of zero and a fixed standard deviation  $\sigma$  and represents short-term fluctuations in emotions due to the environment the system interacts with. Note that we used the same parameterization as (van de Leemput et al., 2014a), except that in our model we use an additive noise term instead of a multiplicative noise term for simplicity and set all  $a_i = 1.6$ .

Due to the symmetries in  $C$ ,  $r$  and  $a$ , emotions with the same valence are exchangeable. We can therefore describe the dynamics of the 4-dimensional system using a 2-dimensional system consisting of one dimension for positive emotions and one dimension for negative emotions (for details see Appendix F.1). Figure 9.1 illustrates the dynamics of the deterministic part (i.e., with  $\epsilon_i = 0$ ) of this model: Panel (a) displays the stable (solid lines) and unstable (dashed lines) fixed points for positive (green) and negative (red) emotions, as a function of stress. For example, for a low stress level of 0.9 there is only a single fixed point: the positive emotions (PE) have the value 5.28 and the negative emotions (NE) have the value 1.15. We therefore also refer to this fixed point as the healthy state. If the stress level remains unchanged, the system will always end up at this fixed point, no matter how one chooses the starting values. This dynamic is illustrated in the corresponding vector field in panel (b). The arrows depict the partial derivatives with respect to the two emotions and therefore describe the linearized dynamics at a given point in the 2-d space. The vector field shows us that whichever initial values we choose, the system will always end up at the fixed point at (PE = 5.28, NE = 1.15). Thus, the system with stress = 0.9 describes a person whose emotions can be changed by external influences, but eventually always returns to the healthy state of having strong positive emotions and weak negative emotions. The solid lines in panel (b) indicate the values of positive and negative emotions for which the two differential equations are zero. At the intersections of those lines both differential equations are equal to zero, which means that the system does not change anymore, which is the definition of a fixed point.



**Figure 9.1:** The dynamics of the bistable system we will use as the data generating model throughout the chapter. Panel (a) shows the fixed points of the deterministic part of the model as a function of stress, operationalized by the rate of change of the negative emotions. Solid lines indicate stable fixed points and dashed lines indicate unstable fixed points. Panels (b), (c) and (d) show the vector fields of the system for the stress values  $r_3, r_4 = 0.9, 1$  and  $1.1$ . Solid points indicate stable fixed points and empty points indicate unstable fixed points. The solid lines indicate the values at which derivative of positive emotion (orange) and negative emotion (light blue) is equal to zero. At the points at which the two lines meet, both derivatives are equal to zero and the system remains in this (stable) state.

Panel (a) of Figure 9.1 shows that when increasing stress from 0.9 until around 0.95, the stable fixed point changes *quantitatively*: the value of positive emotion value decreases, and the value of negative emotion value increases. However, from around 0.95 on the dynamics of the system change *qualitatively*: the system now has three fixed points. For example, at stress = 1, the fixed points are (PE = 4.89, NE = 1.36), (PE = 2.80, NE = 2.80), and (PE = 1.36, NE = 4.89). The first fixed point is the stable healthy fixed point we also observed for values smaller than 0.9. The second fixed point is an unstable fixed point. Specifically, it is a saddle point, because the arrows in the vector field flow towards this fixed point in one direction, but flow away in the other direction (Strogatz, 2015). The third fixed point is again stable, however, now negative emotions have a high value and positive emotions have a low value. We could call this fixed point the unhealthy fixed point.

The presence of these three fixed points means that, if the system is initialized anywhere except on the diagonal, the system will end up at one of the two stable



fixed points. This behavior is illustrated in panel (c), which shows the vector field of the system for stress = 1. We see that eventually all arrows point away from the unstable fixed point at (PE = 2.80, NE = 2.80) and towards one of the two stable fixed points. Thus, the system will never converge to this point except if it is initialized exactly on the diagonal. For all other starting values, the system will converge to one of the two stable fixed points. For the particular case of stress = 1, starting values above the diagonal line will converge to the unhealthy fixed point (PE = 1.36, NE = 4.89), whereas starting values below the diagonal line will converge to the healthy fixed point (PE = 4.89, NE = 1.36). This system describes a person that starts out in the healthy (unhealthy) state, and always returns to the healthy (unhealthy) state after small outside influences. However, a large influence can push the person into the unhealthy (healthy) state, and now the person remains there until a large enough influence pushes her back into the healthy (unhealthy) state.

When increasing stress further until around 1.06, we observe again a quantitative change of the three fixed points: the negative emotions go up, and the positive emotions go down. However, from around 1.06 on the system changes again qualitatively. It now again exhibits only one fixed point, which is now the unhealthy fixed point. Thus, when stress is larger than around 1.06, the system will always converge to the unhealthy fixed point. This behavior is illustrated in panel (e), which depicts the vector field for the system with stress = 1.1. We see that there is only a single fixed point at (PE = 1.03, NE = 5.98) and the arrows show that the system will always converge to this point. This system describes a person that will always return to the unhealthy state, no matter how large of an outside influence is applied.

So far, we only discussed the deterministic part of the model, that is, our model with noise set to zero (i.e., with  $\epsilon_i = 0$ ). Introducing noise changes the dynamics of the system, and how exactly it changes depends on the stress level. For low stress (below 0.95), the system will fluctuate around the healthy fixed point. For high stress (above 1.06), the system will fluctuate around the unhealthy fixed point. The interesting behavior is observed for stress values between 0.95 and 1.06: then, the system will fluctuate around one of the two fixed points, but occasionally the noise will be large enough to push the system to the other fixed point. The frequency of switching is a function of the distance between the two fixed points, the vector field between the two fixed points, and the variance of the Gaussian noise process  $\epsilon_i$ . If the variance is low, the probability of a noise draw that is large enough to “push” the system to the other fixed point is small, and consequently the frequency of switching is low. In contrast, if the variance is high, the probability of a large enough noise draw to switch to the other fixed point is high, and consequently the switching frequency is high.

## 9.2.2 Generating Time Series from Bistable System

In the previous section we have shown that our dynamical system is bistable for stress values  $(r_3, r_4) \in [0.96, 1.06]$ . In the remainder of this manuscript we keep stress constant at stress = 1, and therefore study the bistable dynamical

system with the dynamics displayed in panel (c) of Figure 9.1 and the fixed points described above. Apart from stress we chose all parameters as indicated in the previous section.

To obtain a plausible switching frequency for emotion dynamics we set the standard deviation of the Gaussian noise term  $\sigma = 4.5$ . Note that a system can be bistable, but the outside influences (the noise term) are so weak that the system switches very infrequently or not at all. In such cases the bistable system is more difficult (infrequent) or impossible (no switches) to recover. Thus, our choice of  $\sigma$  represents an ideal situation, and all presented methods will perform worse with a lower switching frequency.

In the remainder of this section we describe how we generated the two time series that we will use throughout the chapter: an “ideal” time series with an extremely high sampling frequency of 1 measurement every six seconds (Section 9.2.2.1); and a more realistic time series with measurements every 90 minutes, a sampling frequency typical for ESM studies (Section 9.2.2.2).

### 9.2.2.1 Ideal Time Series

We generated data by computing the numerical solution to the model in Section 9.2.1 with stress = 1 on the interval  $[0, 20160]$ , using Euler’s method (e.g., Atkinson, 2008). We chose a step size of 0.01 to limit computational cost and disk space, however the system shows qualitatively the same behavior for smaller step sizes. We interpret a time step of 1 as one minute, and therefore the time series spans two weeks ( $60 \times 24 \times 14 = 20160$ ). We obtain a time series by sampling the numerical solution obtained via Euler’s method 10 times per minute (or every six seconds). We therefore obtain the ideal time series with  $20160 \times 10 = 201600$  measurements, which appears to switch between fixed points around 17 times<sup>1</sup>. Figure 9.2 displays this time series.

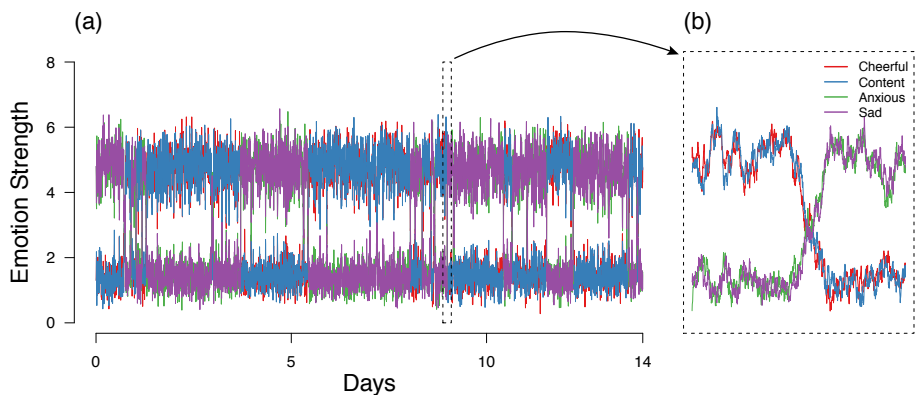
We choose this unrealistically ideal time series (two weeks, one measurement every six seconds, continuous response scale, no measurement error or missing values) with 201600 measurements to be able to study the usefulness of different data analytic methods *in principle*. That is, we study the usefulness of all methods on the *population level*, which is the situation in which we have infinitely many observations and sampling variation does not exist. With 201600 observations, in this setting we approximate “infinitely many” for all practical purposes.

Note that we would not be able to investigate how well different methods perform in principle, if we made the time series more realistic by choosing a shorter time interval or sampling it with a lower sampling frequency: in such a case we would not know whether a method cannot recover (an aspect of) the bistable system for fundamental reasons, or because the time series is too short or the sampling frequency too low. We therefore first study all methods with the ideal time series in order to identify their fundamental limitations. In the second part of the chapter, we make the time series more realistic by taking measurements at a sampling frequency that is typical for ESM studies. This will allow us to inves-

---

<sup>1</sup>The code to generate data and reproduce all analyses and results shown in this chapter can be found at <https://github.com/jmbh/RecoveringBistableSystems/>.

tigate the impact of sampling frequency on all methods. In the following section we describe how we generate this ESM time series.



**Figure 9.2:** Panel (a) shows the ideal time series of the four emotion variables Cheerful, Content, Anxious and Sad. We see that the system switches 17 times between healthy and unhealthy state. Panel (b) displays the twelfth switch, which is a transition from the unhealthy to the healthy state, which occurs on day 9.

### 9.2.2.2 Experience Sampling (ESM) Time Series

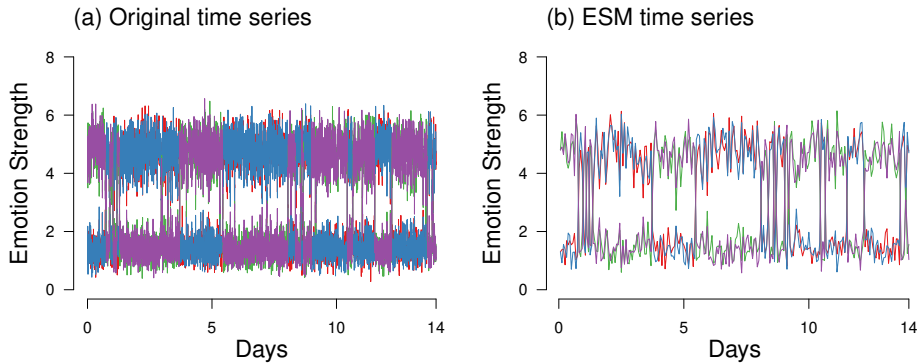
Clearly, the ideal time series is very different from time series data sets obtained from typical ESM studies. The perhaps two most important differences between the ideal time series and realistic time series are the measurement scale and the sampling frequency. With respect to the measurement scale, most ESM studies do not use a continuous response but, for example, a 7-point Likert scale. Regarding the sampling frequency, ESM studies investigating psychological variables typically do not measure more frequently than every 90 minutes (e.g., Bringmann et al., 2013; Pe et al., 2015; Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019). Thus, ESM time series have a much lower sampling frequency (every 90 minutes) than the ideal time series used above (every six seconds). To be able to explain possible drops in performance of certain methods when making the time series more realistic, we are only allowed to change one aspect of the time series. While the measurement scale can possibly be made near-continuous, there are certainly hard limits on how many times one can notify a person each day with an ESM questionnaire. We therefore consider the sampling frequency the more fundamental constraint in realistic data, and thus make it the focus of our investigation in Section 9.4.

Taking a measurement every 90 minutes in the two weeks of the original ideal data leads to 224 measurements. This would mean that we would compare the “ideal” time series with 201600 measurements which essentially implies the absence of sampling variation to an ESM time series with 224 measurement which implies a lot of sampling variation. Thus, any comparison would be confounded by the difference in the number of measurement points (i.e., sample size). To

avoid this confound, we increase the measurement interval of the ESM time series to 1800 weeks, which ensures that the new ESM time series has exactly the same sample size as the ideal data ( $\frac{224}{2} \times 1800 = 201600$ ). Thereby, we provide that any drop in performance is a function of the lowered sampling frequency and cannot be explained by lower sample size (and higher sampling variation). Note that studying the performance of methods as a function of sample size (sampling variation) is of paramount importance to evaluate how useful a given method is in a realistic application. However, here we study the more fundamental question of the impact of reducing the sampling frequency to a level that is typical for ESM studies. We do this because if we find that a method is ill-suited to recover (some aspect of) a bistable system with a realistic sampling frequency on the population level (i.e., with infinite sample size), then it does not make sense to investigate the performance of the method in the less ideal scenario with realistic (small) sample sizes.

So far, we only discussed that we sample every 90 minutes. However, to emulate ESM measurements, we also need to formalize how exactly ESM questions measure the four emotion variables. This is far from trivial: questions in some ESM studies refer to the very moment of measurement and are phrased along the lines of “How cheerful do you feel right now?”. Such measurements could be formalized by defining the measurement as the set of current values of the system (a “snapshot” of the system) at the measurement time. In contrast, other ESM studies refer to the time period since the last measurement. A question of this type could be phrased “How cheerful did you feel in the time since the last notification?”. Such measurements could be formalized by defining the measurement as the average values of the system since the last measurement. However, many other measurement functions are also possible. In this chapter we analyze the first type of ESM question, because its measurement function is the simplest. However, we also performed all analyses with the second kind of ESM question, and all our main conclusions also hold in this situation.

Figure 9.3 displays the two week long original time series (see also Figure 9.2 panel (a)) next to the ESM time series which was obtained by taking “snapshots” of the process at 90 min intervals. The ESM time series in panel (b) appears less dense, which is what we would expect since it contains only 1/900 of the time points of the ideal time series in panel (a). However, we see that the system is still bistable and that the location of and variance around the fixed points is largely the same. In Section 9.4 we will use this emulated “snapshot” ESM time series to try to recover the true bistable system using the same array of methods as in Section 9.3, in which we analyze the ideal time series.



**Figure 9.3:** Panel (a) shows the original time series that was already shown in panel (a) of Figure 9.2. Panel (b) shows the ESM time series which was obtained by taking snapshots every 90 minutes in the series. Note that the ESM time series we analyze in Section 9.4 is much longer (1800 weeks) than the 14 day ESM time series shown here.

### 9.2.3 Qualitative Characteristics of the Model

In this section we discuss the key qualitative characteristics of the bistable system introduced in the previous section. We list these characteristics because most considered methods are models that are misspecified (i.e., they do not contain the true system as a special case). In such a situation one can only hope to recover some characteristics of the true system, and we therefore evaluate how well a method recovers the bistable system based on how well it recovers the following seven characteristics:

#### Global dynamics

1. Bistability (two stable fixed points)
2. Position of fixed points
3. Variability around fixed points
4. Frequency of transitions

#### Microdynamics

5. Suppressing effects between valences, reinforcing effects within valences
6. Relative size of suppressing/reinforcing effects
7. All parameters are independent of time and independent of variables outside the model

The first four characteristics describe the global dynamics of the dynamical system. The first is bistability, which means that the data generating mechanism exhibits two stable fixed points. This is the case for the data generating mechanism with stress set to 1, which we use to generate data from and aim to recover throughout the chapter (see Figure 9.2, panel (a) and (c)). The second characteristic is the position of the fixed points, which are at (PE = 4.89, NE= 1.36) for the healthy fixed point, and (PE = 1.36, NE= 4.89) for the unhealthy fixed point. Third, we consider the variability around the different fixed points. Figure 9.2 shows that, for both fixed points, the variability of the emotions with lower values is smaller than the variability of the emotions with larger values. The fourth characteristic is the frequency of transitions between the area around the healthy fixed point and the area around the unhealthy fixed point. In the time series shown in Figure 9.2 we see that the system switches around 17 times.

The remaining three characteristics describe the microdynamics of the dynamical system. The fifth characteristic is that emotions of the same valence reinforce each other, while emotions of different valence suppress each other. The sixth characteristic is the fact that the size (absolute value) of the reinforcing effects (0.04) are smaller than the suppressing effects (0.2). The last characteristic is that all parameters in the system of differential equations are independent of time and independent of variables outside the model.

### 9.3 Recovering the Bistable System from Ideal Data

In this section we analyze the ideal time series to evaluate how well different methods recover the data generating bistable system. The methods considered here primarily consist of the most popular models used in analyzing time series in clinical psychology and psychiatry, and some extensions thereof. In all but one instance, this entails the estimation of misspecified models, that is, models which do not contain the true bistable system as a special case. Thus we will focus our investigation on whether these models allow one to recover some of the characteristics of the true model, as outlined in Section 9.2.3.

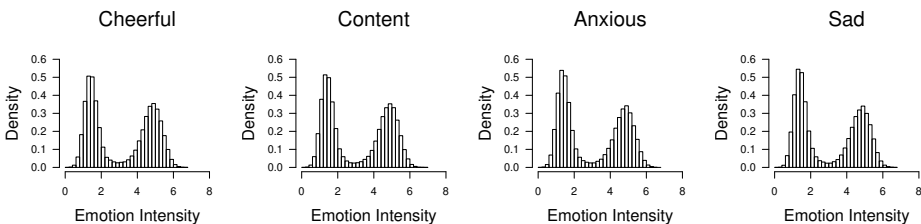
We analyse each method in order of increasing complexity, moving from methods which may be helpful in recovering global characteristics alone to methods which are typically used with the aim of characterising the microdynamic structure. We begin by inspecting the time series using descriptive statistics (Section 9.3.1); in Section 9.3.2 we characterize the switching behaviour in the system using a mean-switching Hidden Markov Model (Hosenfeld et al., 2015; Hamaker, Grasman, & Kamphuis, 2016). Next, in Section 9.3.3 we analyze the multivariate lag-0 (same time point) relationships using correlations and partial correlations with the popular Gaussian Graphical Model (GGM) (Epskamp, Waldorp, et al., 2018). In Section 9.3.4 we use the most popular approach to modeling microdynamics in experience sampling settings, the lag-1 Vector Autoregressive (VAR(1)) model (e.g., Bringmann et al., 2013; Pe et al., 2015; Fisher et al., 2017; Groen et al., 2019). Next, we evaluate the Threshold VAR model, an extension of the VAR model that allows the modeling of state-switching be-

haviour using time-varying parameters (Hamaker, Zhang, & van der Maas, 2009; Hamaker & Grasman, 2012) (Section 9.3.5). While all models so far are misspecified, we include one final method that is capable of recovering the full bistable system: a two-step model building approach based on direct estimation of differential equations from data, following the dynamic systems modeling approach of (Boker et al., 2010) and (Chow, 2019).

### 9.3.1 Descriptive Statistics

To get a rough overview of the behavior of the system, we inspect the time series plot of all four emotion variables shown above in Figure 9.2 panel (a). We see that at almost every time point the two positive emotion variables have high values around 5, and the two negative emotion variables have small values around 1, or the other way around. At the remaining time points, the variables seem to transition between those two states (see panel (b) in Figure 9.2). In addition, we see that the variables switch between states 17 times.

We can extract a considerable amount of information from simply inspecting the time series plot. There seem to be two stable states (fixed points), one in which positive emotions are high and negative emotions are low, and one in which the reverse is true. Further, we see that the variance is higher for the emotion with higher values, and we saw that the system switches around 17 times in the two week window. To get a more direct picture of the distribution of variables and possible fixed points we show the histograms for each variable in Figure 9.4:



**Figure 9.4:** The histograms of the emotion intensity of the four modeled emotions Cheerful, Content, Anxious and Sad, for the ideal data.

We see that at most time points in the time series, each emotion either takes on values around 1 or around 5. This is what we would expect from inspecting the time series plot, however, the histograms give a more precise picture of the distributions and allow one to guess possible fixed points with greater precision. For instance, we could separate the two distributions (using a fixed threshold, or clustering algorithm) and take their means as estimates for the fixed points.

While eyeballing the data should be the first step in any time series analysis, the conclusions are subjective and do not allow us to quantify how certain we are about bistability and the switching frequency. We can quantify the observation that there are two states and that the system is switching between them by fitting a Hidden Markov Model (HMM) (e.g., Rabiner, 1989) to the data, which we will

do in the following section. Such quantification is especially valuable in more realistic situations, in which the two states are probably harder to separate than in our ideal simulated data.

### 9.3.2 Hidden Markov Model

In this section we fit a mean-switching Hidden Markov Model (HMM) in order to scrutinize the intuition that the system switches between two states and to quantify the switching frequency. The HMM models the observed data as consisting of  $K$  latent states or components, characterized by  $K$  multivariate Gaussian distributions, which may differ in their means  $\mu_k$  and variances  $\sigma_k$ .<sup>2</sup> Each observation over time is drawn from one or other of these distributions, and the switching between these states is governed by a matrix of transition probabilities  $A$ . For more details about this model see Appendix F.2.1.

Here we choose  $K = 2$  components, and fit this model to our time series using the R-package *depmixS4* (Visser & Speekenbrink, 2010), obtaining the following parameter estimates

$$\hat{\mu}_1 = \begin{pmatrix} 1.47 \\ 1.46 \\ 4.71 \\ 4.71 \end{pmatrix}, \hat{\sigma}_1 = \begin{pmatrix} 0.41 \\ 0.40 \\ 0.63 \\ 0.62 \end{pmatrix}, \hat{\mu}_2 = \begin{pmatrix} 4.75 \\ 4.76 \\ 1.45 \\ 1.45 \end{pmatrix}, \hat{\sigma}_2 = \begin{pmatrix} 0.63 \\ 0.62 \\ 0.40 \\ 0.40 \end{pmatrix}, \hat{A} = \begin{matrix} & C_1 & C_2 \\ \begin{matrix} C_1 \\ C_2 \end{matrix} & \begin{pmatrix} 0.9996 & 0.0004 \\ 0.0004 & 0.9996 \end{pmatrix} \end{matrix}.$$

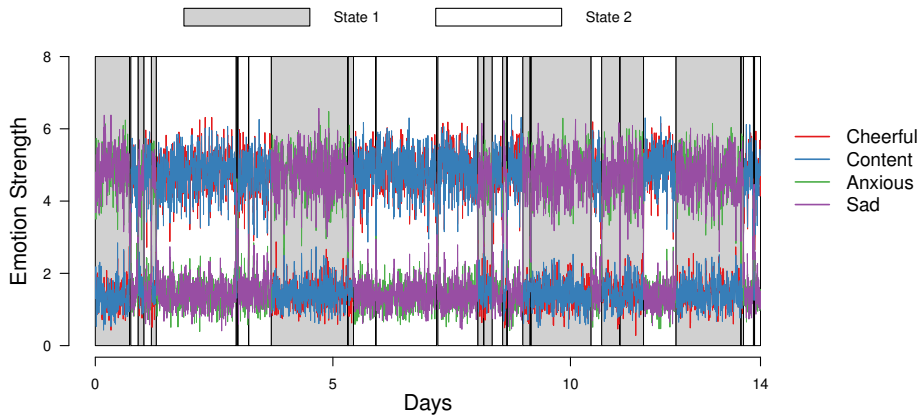
We can see from the estimate  $\hat{\mu}_1$  that in state 1 the means of positive emotions are low, and the means of negative emotions are high. We can therefore identify state 1 as the unhealthy state. We also see that the standard deviations of positive emotions are lower than for negative emotions in the unhealthy state which is what we already observed in the time series plot in Figure 9.2. Similarly, state 2 can be identified as the healthy state, with high means and standard deviations for positive emotions, and low for negative emotions. The transition matrix  $A$  indicates the probabilities of switching between states. We see that there is a very high probability for remaining in the same state ( $A_{11} = A_{22} = 0.9996$ ), and a correspondingly low probability to switch states ( $A_{12} = A_{21} = 0.0004$ ). This is what we would expect, because we take one measurement every six seconds, but the system changes states only a couple of times within the two week window. Multiplying the number of time points of the time series with the switching probability we obtain  $201600 \times 0.0004 \approx 81$  switches, which is in the same order of magnitude of the eyeballed number of switches (17) reported in Section 9.2.2.

In addition to obtaining estimates of means and standard deviations of the two fixed points and the transition matrix, the HMM allows to predict the most likely state for each time point. We show the predicted states for the entire time series in Figure 9.5:

---

<sup>2</sup>In principle, the distributions may also differ with respect to their covariances, but in this analysis, we set all covariances to zero due to limitations of the software package used in estimation.





**Figure 9.5:** Time series of the four emotion variables, also shown in panel (a) of Figure 9.2, with background color indicating whether a given time point is assigned to the first or second component of the mean-switching HMM.

When inspecting the predicted states visually, it seems that the HMM captured the switches well. Next to the larger blocks in which the system stays at the same fixed point, it also identifies switches in which the system switches back and forth within only a few time points. These switches might have been missed when inspecting the time series visually alone.

Taking all results together, which characteristics of the bistable system did we recover with the HMM (see list in Section 9.2.3)? We got an estimate of the location (characteristic 2) and variance (characteristic 3) around two fixed points, which are very close to the healthy and unhealthy fixed points in the true bistable system. We also quantified the frequency of transitions in the transition matrix  $A$ . Since the transition frequency (characteristic 4) is not explicit in the true bistable model, there is no clear way to evaluate this estimate. However, the number of predicted transitions (81) is at least in the same ballpark as the number of transitions eyeballed from the entire time series (at least 17). Note that while a bistable HMM seems to fit the data well, we provided  $K = 2$  as an *input* to the model, and therefore bistability (characteristic 1) cannot be considered a characteristic we recovered with this model. Instead of fixing a particular  $K$ , an optimal  $K$  can be obtained via model selection. However, in Appendix F.2.2 we show that at least the standard approach to selecting  $K$  in mean-switching HMMs performs poorly since the data was not generated from a mean-switching HMM.

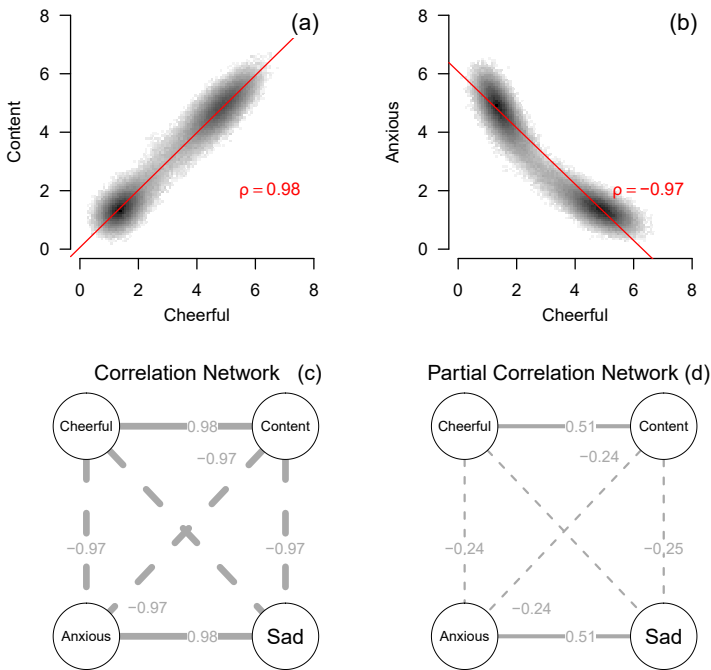
One additional way to visualize or ascertain how much of the true systems behaviour a given model is able to capture is by generating new data from the estimated model parameters. In Figure F.2 in Appendix F.3.1 we generate a two week time series from the estimated mean-switching HMM and compare it to the original time series. We find that the data generated from the HMM is similar to the original data, except for two features: First, the system tends to switch between states somewhat more frequently, and second, there are no observations on the transitions between states.

The remaining three characteristics (5-7) are about the microdynamics of the

true bistable system, that is, about how the components are related to each other. Clearly, the mean-switching HMM we used here cannot elucidate these characteristics since it does not model any dependencies between the four emotion variables. In the following sections we fit models that include such dependencies.

### 9.3.3 Lag-0 Relationships / Gaussian distribution

In this section we analyse the relationships between variables at the same time point. Figure 9.6 panel (a) displays the relationship between Content and Cheerful, two emotions of the same valence. We see that the observations cluster around two points, one close to (1, 1) with smaller variance, and one close to (5, 5) with larger variance. The red line indicates the best fitting regression line (correlation 0.98). Panel (b) displays the relationship between Cheerful and Anxious, two emotions of different valence. We see that that the observations cluster around two points, one close to (1, 5) and the other one close to (5,1). The red line indicates the best fitting regression line (correlation  $\rho = -0.97$ ).



**Figure 9.6:** Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, at the same time point. The red line indicates the best fitting regression model. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence. Panel (c) displays the correlation matrix as a network, and panel (d) displays the partial correlation matrix as a network.

Panel (c) displays the correlation network for all four emotion variables. As we have already seen in panel (a) and (b) there is a positive correlation ( $\rho = 0.98$ ) between Content and Cheerful, and a negative correlation ( $\rho = -0.97$ ) between Cheerful and Anxious. Due to the symmetry in the true bistable system, all correlations between emotions with the same valence are equal to  $\rho = 0.98$  and all correlations between emotions with different valences are equal to  $\rho = -0.97$ . Panel (d) shows the partial correlation network (i.e., GGM). We see that the partial correlations between emotions with the same valence are equal to  $\theta = 0.51$ , and the partial correlations between emotions with different valences is equal to  $\theta = -0.24$  or  $\theta = -0.25$ .

What can we learn from these results about the underlying bistable system? From inspecting the pairwise relationships of emotions with same and different valence in panels (a) and (b) one could guess the location and variance of possible fixed points, similarly to inspecting the histograms in Section 9.3.1. However, the 2-dimensional representation offers additional information about the stability landscape, for example the shape around the fixed points and the most likely paths to transition between them. When interpreting the correlations in panel (b) as “contemporaneous” relationships, we would conclude that there are strong positive linear relationships between emotions with the same valence, and similarly strong negative linear relationships between variables with different valences at a relatively short time scale. The partial correlations in (d) are smaller than the correlations, which is what one would expect since all correlations are high.

Using our knowledge about the true bistable system, which characteristics did we correctly recover? From inspecting the scatter plots in panels (a) and (b) one sees that most observations fall in one of two clusters indicating bistability (characteristic 1). Also, one can obtain rough estimates of the position of the fixed points (characteristic 2) and sees that the variances around the fixed points is different (characteristic 3). Note that the shape of the scatter plot in panel (b) is determined by the vector field in Figure 9.1 (c). The two clusters are exactly at the location of the two fixed points, and the observations between the clusters are both due to variance around the fixed points and switches between fixed points.

From the correlation and partial correlation network, we correctly find that there are reinforcing effects within valences, and suppressing effects between valences (characteristic 5). However, the correlation network suggests that their relative size is equal, and the partial correlation network suggests that the reinforcing effects are stronger. In the true bistable system, however, the suppressing effects between valences are larger than the reinforcing effects within valences. Thus, judging the relative size of suppressing/reinforcing effects within/between valences from (partial) correlation would lead to incorrect conclusions.

In sum, inspecting scatter plots of pairwise relationships indicated bistability, and allowed us to obtain a rough estimate of the location of and variances around the fixed points and. The scatter plots also allowed one to get a projection of the stability landscape on two dimensions and thereby provide more information than histograms. While inspecting the scatter plots allows one to recover global dynamics of the true bistable system, one cannot infer the coupling between the

emotion variables in the true bistable system from (partial) correlations. This is not too surprising since the Gaussian distribution is very restrictive in that it only models pairwise linear relationships (opposed to e.g., 3-way, 4-way, etc. interactions). In addition, it does not model any dependencies across time, which are the types of dependencies that constitute the microdynamics (characteristics 5-7) of the true model. In the next section, we inspect those dependencies across time and fit a Vector Autoregressive (VAR) model to the data, which captures temporal linear dependencies.

### 9.3.4 Lag-1 Relationships / VAR Model

In this section we aim to characterize the microdynamics between the four emotion variables by modeling the lagged linear relationships between them. Panels (a) and (b) of Figure 9.7 show the *marginal* relationship of Content at time  $t$  with Cheerful at the previous time point  $t - 1$ , and Anxious at time  $t$  with Content at time  $t - 1$ , respectively. Although the data is generated from a dynamic model, the marginal lagged relationships look very similar to the contemporaneous relationships shown in Figure 9.6: when averaging over all other variables, the lagged relationships within-valence are positive, and between-valence are negative. The reason is that the system largely stays around the two stable fixed points, and relative to the length of the time series, switches are infrequent. As such, these marginal relationships are largely driven by the relative location of the two fixed points.

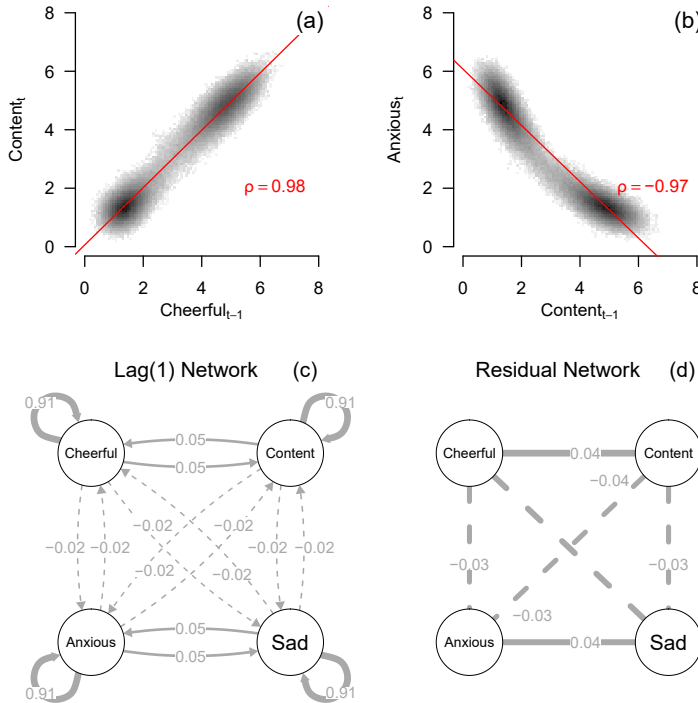
We can gain further insight into the dependencies between variables in our model by examining the *conditional* lagged relationships between pairs of emotions, that is, when keeping the other emotion variables at the previous time point(s) fixed. A popular model for such conditional lagged relationships is the first-order vector auto-regressive (VAR(1)) model. The VAR(1) model is one of the simplest multivariate dynamic models which can be fit to repeated measurement data, allowing linear relationships between all pairs variables observed at consecutive measurement occasions  $t$  and  $t - 1$ :

$$X_t = b + \Phi X_{t-1} + e_t . \quad (9.2)$$

where  $b$  is a vector of intercepts,  $\Phi$  is a matrix containing the auto-regressive ( $\Phi_{ii}$ ) and cross-lagged ( $\Phi_{ij}, i \neq j$ ) effects, that is, conditional linear dependencies, and  $e_t$  is a vector of normally distributed residuals  $e_t \sim \mathcal{N}(0, \Psi)$ , which are independent across time, with residual variance-covariance matrix  $\Psi$ .

The VAR(1) has been used widely to analyze experience sampling data in psychopathology research, particularly in the form of dynamic network analysis, wherein the  $\Phi$  and  $\Psi$  matrices are used to construct directed and undirected network structures, respectively (e.g., Bringmann et al., 2013; Pe et al., 2015; Epskamp, Waldorp, et al., 2018). The VAR(1) model describes a system which fluctuates around a *single* stable fixed point: Stochastic input in the form of a residual term pushes the system away from this fixed point, and the system returns to the fixed point with an exponential decay (Hamilton, 1994). The location

of the stable fixed point is given by the mean vector  $\mu$ , a function of the intercepts and the lagged relationships  $\mu = (I - \Phi)^{-1}b$ , where  $I$  is the identity matrix.



**Figure 9.7:** Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, spaced one time point apart (at a lag of one). The red line indicates the best fitting regression model. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence, at a lag of one. Panel (c) displays the matrix of lagged regression parameters, estimated from a VAR(1) model, as a network, and panel (d) displays the partial correlation matrix of the residuals of the VAR(1) model as a network. This latter network is often referred to as the contemporaneous network.

Panel (c) of Figure 9.7 displays the network of estimated lagged regression coefficients ( $\hat{\Phi}$ ) between Cheerful, Content, Anxious and Sad. We can see that the auto-regressive parameters are large and positive for all four variables ( $\hat{\Phi}_{ii} = .91$ ). Furthermore, there are positive cross-lagged relationships between variables of the same valence ( $\hat{\Phi}_{12} = \hat{\Phi}_{21} = \hat{\Phi}_{43} = \hat{\Phi}_{34} = .05$ ) and weaker, negative cross-lagged effects between variables of opposite valence ( $\hat{\Phi}_{13} = \hat{\Phi}_{31} = \dots = -.02$ ). All within-valence effects, and all between-valence effects, are of roughly equal magnitude, respectively. In panel (d) we show the partial correlations of the residuals (i.e., standardized  $\hat{\Psi}^{-1}$ ), sometimes referred to as the “contemporaneous” network or the residual GGM (Epskamp, Waldorp, et al., 2018). Here we see a similar pattern as above: the residuals have negative conditional relationships between-valence, and slightly greater in magnitude positive conditional

relationships within-valence. Note that the residual variance in this case is quite low for each variable ( $\hat{\Psi}_{ii} \approx 0.0185$  with in-sample explained variance of approximately 99 percent). The estimated fixed point (that is, the mean) is approximately  $\hat{\mu}_1 = \hat{\mu}_2 = 3.16$  for Cheerful and Content, and  $\hat{\mu}_3 = \hat{\mu}_4 = 3.04$  for Anxious and Sad.

Which characteristics of the bistable system can we recover based on the VAR(1) estimates? The strong auto-regressive effects correctly capture the strong linear auto-effects present in the true system, defined by the  $r$  parameters in Equation 9.2. The lagged regression parameters suggest that there are suppressing effects between valences, and reinforcing effects with valences, capturing characteristic number five of the data generating mechanism (Section 9.2.3). However, the relative size of the suppressing and reinforcing effects is flipped in the VAR(1): The suppressing effects are in fact larger in absolute value than the reinforcing ones (see Section 9.2.1).

It is unclear what conclusions we can draw from the weak relationships present in the residual network — as there are no such additional instantaneous relationships present in the data generating system. We assume a-priori when fitting the VAR(1) model that these parameters are independent of (i.e., constant across) time. Finally, the VAR(1) model describes a uni-stable system, precluding us from capturing any characteristics related to bistability. The dynamics implied by the VAR(1) are illustrated by generating new data from the estimated parameters, displayed in Appendix F.3.2. The estimated location of the single fixed point is not equivalent to either of the two stable fixed points or the unstable fixed point in the true system.

Importantly, we can use our knowledge of the true bistable system to determine how we arrived at these observed parameter estimates. First, the estimated position of the fixed point (given by  $\hat{\mu}$ ) is roughly halfway between the positions of the two stable fixed points, and is approximately equal to the sample mean for each variable, reflecting that the system spends roughly the same amount of time around each of the two stable fixed points. The main counter-intuitive result from the VAR(1) model is that the order of magnitude of the between- and within-valence relationships is different than in the true bistable system. In the true system, these pairwise relationships are non-linear, taking the form of interaction effects, and that the VAR(1) model captures the best linear approximation of these non-linear relationships. As we can see from panel (a) of Figure 9.7, the linear approximation of the within-valence relationship is largely driven by the strong positive relationship present when both variables take on a high value, e.g., when Cheerful and Content are both near the healthy fixed point. From panel (b) we can see that, in contrast, the linear between-valence relationship of Content on *Anxious* is in fact a mixture of the strong negative effect near the unhealthy fixed point (Content is low, Anxious is high) and the weaker effect near the healthy fixed point (Content is high, Anxious is low). Combined, this results in higher linear within-valence relationships and lower linear between-valence relationships.

Finally, the residual covariances displayed in panel (d) of Figure 9.7 are produced by a combination of model-misspecification (linear approximation of non-linear relationships) and paths between each process at a shorter time scale than

observed, due to the Euler steps used in data generation. We stress here that, even in the current idealized situation, it is not trivial to derive an exact explanation for the residual covariance structure, and so its utility in drawing conclusions about the underlying system should be approached with great caution.

In summary, the VAR(1) model gives us rather limited information regarding the core characteristics of the bistable model. In principle, the VAR(1) model is unable to capture any features which relate to bistability (characteristics 1-4), as one would expect from a model that exhibits only a single fixed point. What is perhaps more surprising is that, while the sign of the lagged relationships (characteristic 5), and their symmetries are captured, their relative ordering (characteristic 6) is not. This observation is critical: while we could expect that the VAR(1) model would not reproduce the global dynamics of the system, even when we have ideal data, the linear relationships in the VAR(1) model also fail to appropriately capture the local microdynamics in this instance. Fundamentally, this is due to the non-linear relationships which must be present in the underlying system in order to induce bi-stability: In general we would not expect that linear approximations of non-linear effects would preserve the same rank ordering. This observation has potentially major implications for the analysis of dynamic network structures, because many network derivatives such as centrality metrics are strongly dependent on the relative ordering of effects. In the following we will examine if extending the VAR(1) model to allow for bi-stability brings us closer to recovering a more accurate characterization of the data generating bistable system.

### 9.3.5 Threshold VAR Model

*Regime-switching* VAR(1) models extend the VAR(1) model to allow for observations to be drawn from two different conditional distributions for  $X_t$  given  $X_{t-1}$ , that is, two different *regimes*, described by two different sets of model parameters. These extensions in principle allow us to directly capture a notion of multi-stability, by interpreting the mean vector of each conditional distribution as a separate fixed point. Different extensions allow for different mechanisms by which to model the switch between these regimes.

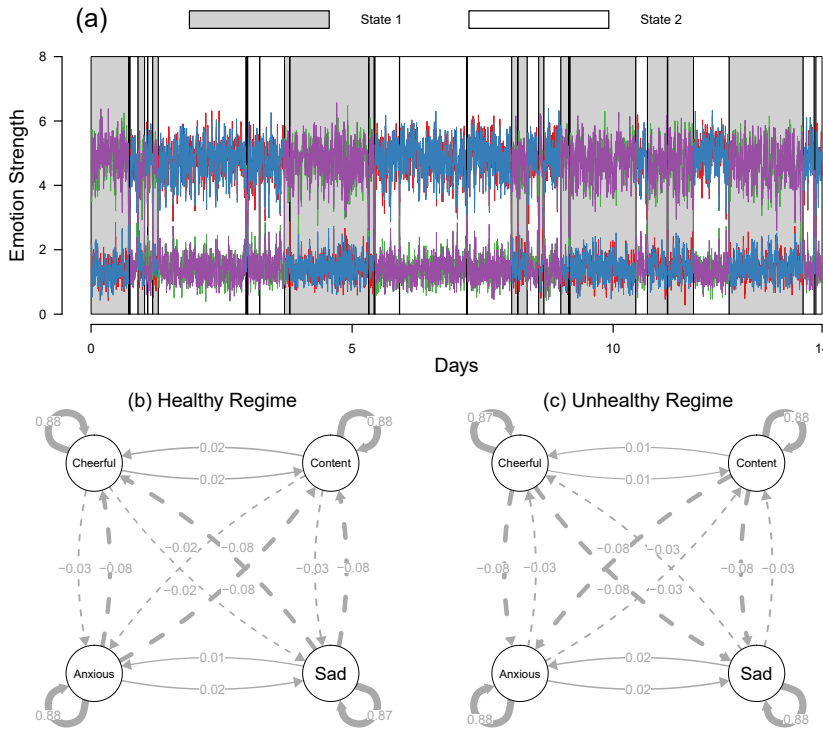
One popular regime-switching VAR(1) model is the Threshold TVAR(1) model, where the system enters a different regime whenever a threshold value or values  $\tau$  of an a-priori specified threshold variable  $z_t$  is crossed, written

$$\begin{aligned} X_t &= b^{(1)} + \Phi^{(1)}X_{t-1} + e_t^{(1)} && \text{if } z_t \leq \tau \\ X_t &= b^{(2)} + \Phi^{(2)}X_{t-1} + e_t^{(2)} && \text{if } z_t > \tau \end{aligned}$$

for a two-regime model with a single threshold, where the VAR(1) parameters are indexed by regime, with  $e_t^{(r)} \sim \mathcal{N}(0, \Psi^{(r)})$ , and mean vectors  $\mu^{(r)} = (I - \Phi^{(r)})^{-1}b^{(r)}$  (Tong & Lim, 1980; Hamaker et al., 2010). The threshold variable  $z_t$  may be an exogenous variable, or one of the variables in the VAR model. Here we choose to use Cheerful ( $z_t = x_{1,t-1}$ ) as the thresholding variable. The threshold value  $\tau$  is a

hyper-parameter that is estimated. Here, we estimate the TVAR(1) model using the R-package *tsDyn* (Fabio Di Narzo et al., 2009), which estimates  $\tau$  using a grid search which selects the model with minimum summed squared residuals.

Figure 9.8 displays the main results from the estimated TVAR(1) model, in which the threshold is estimated as  $\hat{\tau} = 2.811$ . In panel (a) of Figure 9.8 we show the time-series with shading indicating which observations are below (grey) or above (white) the threshold. We can see that the estimated threshold nicely separates the time series into periods in which the system is in an unhealthy state (based on Cheerful values below the threshold) and a healthy state (Cheerful values above the threshold).



**Figure 9.8:** Panel (a) shows the two weeks of the time series, with observations shaded in either grey or white as a function of whether  $x_{1,t-1}$  is above or below the threshold  $\hat{\tau} = 2.811$ . Panels (b) and (c) show the estimated VAR(1) parameters as lagged networks in the healthy (white) and unhealthy (grey) regimes respectively.

Inspecting the lagged networks for each regime in panels (b) and (c) of Figure 9.8 we can see that the auto-regressive effects, and the within-valence cross-lagged effects are pretty similar across both regimes. However, the cross-lagged effects between variables of opposite valence are different. In the healthy regime, negative valence emotions have much stronger cross-lagged effects on positive emotions ( $\hat{\Phi}_{13}^{(2)} = \hat{\Phi}_{14}^{(2)} = \hat{\Phi}_{23}^{(2)} = \hat{\Phi}_{24}^{(2)} = -0.08$ ), and vice versa for the unhealthy



regime ( $\hat{\Phi}_{31}^{(1)} = \hat{\Phi}_{41}^{(1)} = \hat{\Phi}_{32}^{(1)} = \hat{\Phi}_{42}^{(1)} = -.08$ ). Residual partial correlation networks for both regimes are shown in Appendix F.4, which display a similar pattern to the regular VAR(1) model of weak positive residual partial correlation within-valence and weak negative residual partial correlation between-valence. For the TVAR, however, the residual covariance matrix is not symmetric across regimes: In the healthy regime there is a slightly higher covariance between positive emotions than negative emotions, and vice versa. The estimated means are given as  $\hat{\mu}_2 = \{4.74, 4.75, 1.45, 1.46\}$  for the healthy state and  $\hat{\mu}_1 = \{1.49, 1.48, 4.69, 4.69\}$  for the unhealthy state. Data generated by the TVAR(1) model estimates is shown in Figure F.4 in Appendix F.3.3. From this figure we can see that most of the global dynamics are well reproduced, although the system contains fewer switches between regimes than we would expect and there are fewer observations on the switches between states compared to the original time series.

Which characteristics of the bistable system do we recover on the basis of the TVAR(1) parameter estimates? First, the model picks up a number of characteristics related to the bistability of the system: The estimated mean vectors capture approximately the position of the two stable fixed points (characteristic 2), and the estimated threshold correctly captures the position of the unstable fixed point in the Cheerful dimension. However, note that bistability (characteristic 1) has been specified a-priori and therefore cannot be considered to be recovered by the model. Second, although the simulated data in Figure F.4 (Appendix F.3.3) exhibits less frequent switches between states than we would expect, we can see that the combination of state-dependent lagged parameters and residual variances does reproduce higher variability of positive emotion in the healthy state in comparison to the unhealthy state, and vice versa for negative emotions (characteristic 3). Finally, the lagged regression parameters in each regime correctly capture that there are reinforcing effects within valence, and suppressing effects between valence (characteristic 5).

The result that stands out in this analysis is the asymmetry in lagged regression coefficients across both regimes. This asymmetry would appear to indicate that the parameters relating processes either change over time or are all explicitly a (step) function of the Cheerful variable. This last result is striking because this intuitive interpretation does not correctly characterize the relationship between variables of different valences in the true bistable system. This is because we know that the dependencies in  $C$  are invariant over time and fully symmetric. However, the dependencies in  $C$  relate to pairwise interaction effects rather than linear dependencies in the VAR(1) model. For example, the relationship between Anxious, denoted  $x_3$ , and the rate of change of Content,  $\frac{dx_2}{dt}$ , depends both on the value of  $C_{23}$  and on the current value of  $x_2$

$$\frac{dx_2}{dt} = r_2x_2 + (C_{23} \times x_2)x_3 + \dots \quad (9.3)$$

If we view  $x_2$  as a moderator, we can see that, when  $x_2$  is high, the effect of  $x_3$  on the rate of change, given by  $C_{23} \times x_2$ , is relatively greater than when  $x_2$  is low. In our system, separating the time-series into two regimes based on a threshold of 2.811 for the Cheerful emotion essentially means we condition on high val-

ues of  $x_1$  and  $x_2$  in the healthy regime, and low values in the unhealthy regime. This leads to the relatively stronger linear relationship from negative emotions to positive emotions in the healthy regime, and vice versa in the unhealthy regime. As such, we can see that the asymmetry in lagged relationships over time picked up by the TVAR(1) model is a characteristic of the true bistable system. Notably, however, the mechanism by which this asymmetry occurs is entirely due to non-linear relationships between the observed variables and the similarity of variables that share the same valence, while the TVAR(1) modeller might be tempted to ascribe this entirely to the effect of the level of Cheerful.

To summarize, the TVAR(1) model allows us to recover global dynamics, and it recovers some aspects of the microdynamics. However, we saw that a naive interpretation of the TVAR parameter estimates may easily lead to the incorrect conclusion that there is one time-varying variable which moderates the relationships between all variables. In addition, we provided bistability as an input to the model, and therefore cannot be considered a characteristic recovered from data. In principle one could perform model selection between TVAR(1) models with different numbers of components, however compared to the Mean switching HMM in Section 9.3.2, the run time for such a model comparison was unfeasible for the large data set used in our chapter.

Furthermore, note that the threshold VAR(1) model does remarkably well for this specific system for the following reason: While TVAR(1) models have frequently been discussed in the literature (e.g., Warren, 2002; Hamaker et al., 2009, 2010; De Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2016) a major limitation of this method is the difficulty in choosing a thresholding variable. In our data generating mechanism, we know there to be an unstable fixed point defined in multivariate space,  $x_1 = x_2 = x_3 = x_4 = 2.8$ . It just so happens that in this case, almost always when we pass this position in one dimension (e.g.,  $x_1 > 2.8$ ) we also do so in all other dimensions (e.g.,  $x_2 > 2.8, x_3 < 2.8, x_4 < 2.8$ ). This means that the true mechanism of state-switching behaviour is very well approximated by the univariate mechanism in the TVAR(1) model, for this choice of parameter values. In more general situations, the choice of thresholding variable(s), and number of thresholds, is likely to be less trivial. While the TVAR(1) model does capture that there are suppressing and reinforcing effects between and within valences, it does not capture the relative size of these effects, and it may easily lead to the incorrect conclusion that there is a single time-varying variable which moderates all of the relationships between other variables in the system.

Finally, the TVAR(1) is only one of a variety of different regime-switching dynamic models which could be fitted to the data at hand. Another alternative would be the Markov-Switching (MS-)VAR model (Hamilton, 1989; Hamaker et al., 2010; Hamaker & Grasman, 2012; Chow et al., 2018), a combination of the HMM and VAR models, in which the regime-switching behaviour is determined by a random Markov process operating between latent categorical variables. While this model is more flexible than the Threshold VAR model, we show here the TVAR results for two reasons. First, in this instance the switching behaviour will be less well approximated by the MS-VAR model, leading to even less straightforward conclusions about the data generating process, but other-

wise likely highly similar lagged parameter estimates. Second, while recent advances such as the *dynR* package (Ou et al., 2019) have made this model easier to estimate, it is still prohibitively difficult and time consuming to fit to data.<sup>3</sup>

Now that we have shown the capabilities and limitations of the TVAR model in recovering the bistable system, there are a few different avenues we could pursue to further increase our model complexity in the hope of recovering more and more of the features of underlying system. For example, both the TVAR and MS-VAR can be considered special cases of *time-varying parameter* models, that assume the true time-varying model is a partition between a finite set of components. Other types of time-varying VAR models assume the parameters are a smooth function of time (e.g., Haslbeck et al., 2020). However, we would not expect these models to outperform the threshold VAR in this instance for two reasons: First, the threshold VAR model is already able to capture the major source of variation in parameters over time, that is, the step-like switches between stable states. Second, since these models are still based on fitting locally stationary VAR models, the fundamental limitations of approximating the dynamics with linear relationships remain. As such, in the next section we examine an approach which aims to recover the exact system of differential equations (DEs) from data, by allowing non-linear terms to enter into a step-wise model building procedure.

### 9.3.6 Differential Equation Model Building

In the previous sections we have shown that some models have been able to recover some characteristics of the true model, but that it is generally difficult to make inferences about the characteristics of the true system from these models. Also, since all of these models were misspecified they were fundamentally unable to recover the exact true bistable system. In this section, we aim to recover the exact system of differential equations (DEs) directly from the ideal time series.

#### 9.3.6.1 Model Building Procedure

The structure of the true model is typically unknown in practice, and therefore has to be learned from the data. (Chow, 2019) describes a general methodology for building dynamic systems models which consists of two steps: In the first step, we approximate the first-order derivatives by taking difference scores between consecutive measurement occasions, divided by the length of the time-interval between those occasions

$$\frac{dx_{i,t}}{dt} \approx \frac{x_{i,t+1} - x_{i,t}}{\Delta t} \quad (9.4)$$

where in each case,  $\Delta t = .1$ , as described in Section 9.2.2 (cf., Boker et al., 2010).

In the second step, we use this approximate derivative as the outcome variable, and try to find a regression model that predicts this outcome variable as well as possible with as few parameters as possible. Here, we use results obtained

---

<sup>3</sup>Despite numerous attempts and correspondence with the authors of the package, we were unable to get the model estimates for the dataset described here to converge.

from the statistical models in the previous sections as a starting point, and then follow the standard model-building approach of fitting models with increasing complexity and evaluating the improvement in out-of-sample fit.

From the descriptive statistics (Section 9.3.1) the marginal lag-0 (Section 9.3.3) and lag-1 (Section 9.3.4) relationships and the mean-switching Hidden Markov Model (Section 9.3.2), we saw that the system is bistable. From dynamical systems theory we know that bistability is only possible in the presence of non-linear terms (Strogatz, 2015). Similarly, we saw from the TVAR(1) model (Section 9.3.5) that the linear relationships between pairs of variables is dependent on where in the state-space other variables are located (i.e., below or above a given threshold). Both of these pieces of information suggest the presence of interaction effects between variables: However, we have no information about what specific interaction terms, or what other linear or non-linear dependencies should be in the model. Here, we start out with a main effects-only model, and then add more and more non-linear terms (interactions, quadratic effects, etc.). We evaluate the fit using the mean out-of-bag proportion of explained variance ( $R^2$ ) obtained from a 10-fold cross-validation scheme (see Appendix F.5 for details), and we choose the model that maximizes this value. If two models result in the same fit, we choose the model with less parameters.

The fit of the all models considered here is shown in Table 9.1. First, we test a baseline model (Model A) where each derivative is a linear function of all other variables. As we described above, the absence of interaction effects makes it unlikely that this is a suitable candidate model, but it gives us a baseline explained variance value of  $R^2 = 0.04664$ . In Model B, we add to the baseline model all pairwise interactions between the outcome process (e.g.,  $x_1$  when the DV is  $dx_1/dt$ ) and the other variables in the model  $x_j$  (i.e.,  $x_1 \times x_j, \forall j \in p$ ). Adding these pairwise interactions increases the variance explained to  $R^2 = 0.06874$ . In Model C, we further extend this model by adding all possible pairwise interactions between all variables  $x_i \times x_j, \forall (i, j) \in p$ . However, we see that adding these parameters in fact leads to a slight decrease in explained variance,  $R^2 = 0.06870$ , indicating overfitting. For brevity, we display only these three models, but adding further complexity to model in terms of additional interaction terms, quadratic or cubic terms also fails to increase the out-of-bag  $R^2$  (see Appendix F.5). As such, we can take Model B to be our final model.

Model	$\frac{dx_{it}}{dt} \sim a + r_i x_i + \dots$	$q$	$R^2$
A	$\sum_{j \neq i} r_j x_j$	5	0.04464
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	0.06874
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	0.06870

**Table 9.1:** Model fit results for each of the four models described in text. The second column gives the model equation for each variable,  $q$  denotes the number of parameters estimated per univariate regression model, and the final column indicates the mean proportion of explained variance  $R^2$ , calculated on the hold-out sets of a 10-fold cross-validation scheme (for details see Appendix F.5)

### 9.3.6.2 Dynamics and Data Generated by Final Model

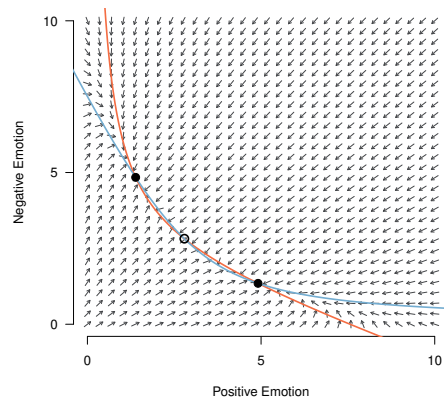
We can see that the structure of Model B is highly similar to the structure of our data generating model, with additional main effects between variables, that is, the linear effects denoted by the  $p \times 1$  vector  $r$  in the true model is replaced by a  $p \times p$  matrix  $R$  in our chosen model. Furthermore, we can see from the left panel of Figure 9.9 that the parameter estimates are highly similar, but not exactly equal to the data generating parameters:

$$\hat{a} = \begin{bmatrix} 1.40 & 1.37 & 1.25 & 1.27 \end{bmatrix}^T$$

$$\hat{\sigma} = \begin{bmatrix} 1.35 & 1.34 & 1.34 & 1.34 \end{bmatrix}^T$$

$$\hat{R} = \begin{bmatrix} 0.88 & 0.02 & -0.01 & 0.01 \\ 0.03 & 0.95 & 0.01 & 0.00 \\ -0.02 & 0.05 & 0.96 & 0.04 \\ 0.04 & -0.01 & 0.08 & 0.91 \end{bmatrix}$$

$$\hat{C} = \begin{bmatrix} -0.18 & 0.04 & -0.17 & -0.18 \\ 0.03 & -0.19 & -0.19 & -0.19 \\ -0.18 & -0.19 & -0.19 & 0.03 \\ -0.19 & -0.18 & 0.02 & -0.18 \end{bmatrix}$$



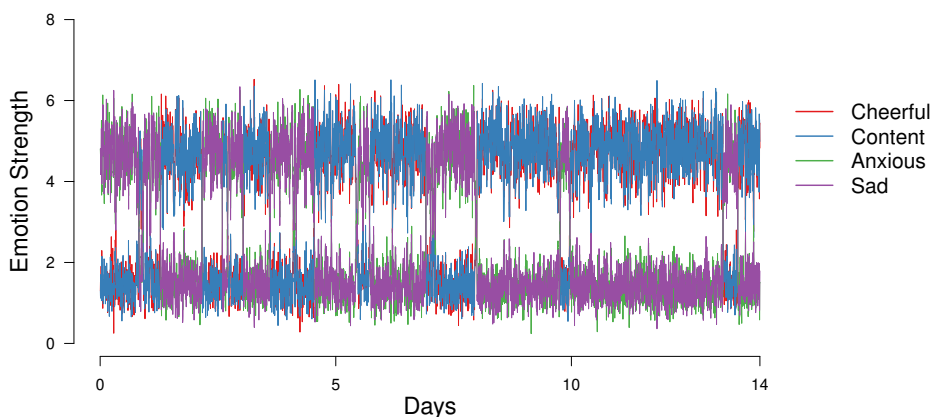
**Figure 9.9:** Left panel: the parameters estimated from the ideal data. Right panel: the vector field defined by the estimated parameters in the left panel. Solid points indicate stable fixed points and empty points indicate unstable fixed points. The solid lines indicate the values at which derivative of positive emotion (orange) and negative emotion (light blue) is equal to zero. At the points at which the two lines meet, both derivatives are equal to zero and the system remains in this (stable) state.

While we would not expect to recover the exact parameters of the true model with a different functional form, we see that the signs, size and relative orderings of parameters in the estimated  $\hat{C}$  matrix are quite accurate. Based on these parameter, we recover that there are suppressing effects between valences and reinforcing effects within valences (characteristic 5), that the reinforcing effects are smaller in absolute value than the suppressing effects (characteristic 6), and by capturing approximately the correct functional form, we capture that the microdynamic parameters are dependent only on variables inside the model (characteristic 7). Furthermore, we can see that false positive (i.e., off-diagonal) elements of  $\hat{R}$  are of a much smaller order of magnitude than the true positive diagonal elements. The full parameter estimates, with standard errors and p-values are shown in Appendix F.5.

Beyond inspecting the estimated parameters, we can judge how good of an approximation of the true bistable system our estimated model represents by comparing the dynamics implied by that model to that of the true system. The dynamics of a differential equation model are described by its vector field, which we

depict for Model B in the right panel of Figure 9.9. To construct this vector field we use the same two-dimensional approximation (positive and negative emotion) as we did in Section 9.2.1 (see Appendix F.1 for details). The orange and light blue lines are solution lines which indicate the locations where the rate of change in one dimension (orange for no change in positive emotion, light blue for no change in negative) is zero. The points at which these solutions line cross determine the fixed points. We can see that our model correctly identifies three fixed points in this range of values: one stable healthy ( $x_1 = x_2 = 4.91, x_3 = x_4 = 1.34$ ), one stable unhealthy ( $x_1 = x_2 = 1.39, x_3 = x_4 = 4.84$ ), and one unstable fixed point approximately halfway between those two ( $x_1 = x_2 = 2.79, x_3 = x_4 = 2.82$ ). If we compare these global dynamics to the global dynamics of the true bistable system depicted in Figure 9.1 in Section 9.2.1, we see that Model B very accurately reproduces these dynamics, approximating the position of the fixed points in the true system closely. From this we can say that the estimated DE model captures characteristics 1 (bistability) and 2 (location of the fixed points) in the true system.

An additional way to evaluate whether the dynamics of the estimated model are similar to the dynamics of the true model is to generate data from the estimated model and compare this data to the original data. Figure 9.10 shows a time series generated from Model B using a step size of .1:



**Figure 9.10:** Data generated from the estimated DE model, with the same initial values as the observed data

We can see that the data looks very similar to the original data generated from the true bistable system: the fixed points are at roughly the same location, there is a difference in variance across the high and low emotion value fixed points (characteristic 3), and there is a similar number transitions (around 14) between the healthy and unhealthy state (characteristic 4). Thus, even though we did not exactly recover the set of true parameters exactly, we seem to have recovered a model that is equal to the true model in all relevant aspects, capturing all of the seven characteristics listed in Section 9.2.3.

### 9.3.6.3 Exact Recovery of Model Parameters

While this model building procedure performed extremely well in this scenario, the findings here should be approached with a note of caution. Observe that, despite negligible sampling error, we do not succeed in recovering the exact parameter estimates. The reason for this is that, while data is generated using an Euler step of  $\Delta t = .01$ , our ideal time series is created by sub-sampling with  $\Delta t = .1$ . While this is an unrealistically high sampling frequency, it still means that we cannot estimate the derivative perfectly: As the sampling frequency becomes lower, we would expect the quality of this approximation to degrade.

In theory, to recover the data generating parameters, we would need to fit the *integral solution* form of the differential equation (Strogatz, 2015), as this describes the relationships between observed variables spaced  $\Delta t$  apart, as implied by the differential equation. It is well known that this integral solution may contain a seemingly different set of dependency relationships than the differential equation: variables which are independent in the DE form may be dependent in the integral form, and the signs and relative orderings of these dependencies may change depending on the value of  $\Delta t$  (Ryan, 2018; Kuiper & Ryan, 2018; Aalen, Røysland, Gran, & Ledergerber, 2012). Because methods based on approximating integral solutions are expected to suffer from similar problems as the two-step DE estimation procedure, and because these methods are difficult to apply in practice, we limit ourselves to the two-step approach in this chapter (see Discussion section 9.5.3 for further details).

### 9.3.7 Summary: Analysis of Ideal Time Series

In this section we aimed to recover characteristics of the true bistable system from the ideal time series with 10 measurements each minute using a number of time series analysis tools. Table 9.2 provides a rough summary of which method recovered which characteristics of the true bistable system. We showed that data visualization (Histograms and the pairwise marginal relationships in Sections 9.3.3 and 9.3.4) revealed bistability and provided a rough estimate of the position and showed that the variances around the fixed points differ. However, when comparing the eye-balled number of switches with the estimates of the HMM, we saw that we missed instances in which the system quickly switched back and forth. The Mean switching HMM recovered all global dynamics, however, we provided bistability as a model assumption, which is why we mark the check mark at the first characteristic with an asterisk.

We showed that data visualization (Histograms and the pairwise marginal relationships in Sections 9.3.3 and 9.3.4) revealed bistability and provided a rough estimate of the position and showed that the variances around the fixed points differ. However, when comparing the eye-balled number of switches with the estimates of the HMM, we saw that we missed instances in which the system quickly switched back and forth. The Mean switching HMM recovered all global dynamics, however, we provided bistability as a model assumption, which is why we mark the check mark at the first characteristic with an asterisk.

	Bistability (1)	Position (2)	Variance (3)	Transitions (4)	Suppr./Reinf. (5)	Relative Size (6)	Time-constant (7)
Data Visualization	✓	✓	✓	×	×	×	×
HMM	✓*	✓	✓	✓	×	×	×
Lag-0 / GGM	×	×	×	×	✓	×	✓*
Lag-1 / VAR(1)	×	×	×	×	✓	×	✓*
TVAR(1)	✓*	✓	✓	✓	✓	×	×
DE-Estimation	✓	✓	✓	✓	✓	✓	✓*

**Table 9.2:** Summary of which method recovered which of the seven qualitative characteristics listed in Section 9.2.3 from the ideal time series. The first four characteristics are global dynamics, the last three are microdynamics. The check marks with asterisk indicate that the method includes the characteristic as a model assumption, and can therefore not be considered recovered from the time series.

Turning to methods that capture dependencies between variables, the analysis of lag-0 relationships with the GGM and the analysis of lag-1 relationships with the VAR(1) model (and a GGM on its residuals) fundamentally cannot recover any global dynamics of the bistable system, but they recovered some microdynamics: the characteristic that within valence effect are reinforcing, and between valence effects are suppressing; and that the parameters are constant across time, however this is again an assumption of the model and therefore cannot be considered recovered from the data. The TVAR(1) model was able to recover all global dynamics with the same caveat as in the HMM, that bistability is a model assumption and not recovered from data. Similarly to the VAR model, it recovered the reinforcing/suppressing characteristic. However, a naive interpretation of the model parameters would lead one to conclude that the parameters are time-varying. Finally, the DE-estimation method was able to recover all microdynamics reasonably well, which implies that it also recovered all global dynamics.

The purpose of this section was to establish whether or not each method can recover, in principle, some aspect of the bistable system. To do this we used a highly idealized dataset, with an unrealistically high sampling frequency. As such, the performance of each method described above can be considered an upper bound on its performance in any more realistic scenario. It remains to be seen exactly how the performance of each method, and in general our ability to recover global and microdynamic characteristics of the system, changes when a more realistic sampling frequency is used.

## 9.4 Recovering the Bistable Systems from ESM Data

In this section we analyze a time series that is similar to the ideal time series in all aspects, except that the system is sampled every 90 minutes instead of every six seconds (see Section 9.2.2). This allows us to investigate how the ability of



each method to recover (some characteristic of) the bistable system is affected by having only a low sampling frequency time series, as it is typical for ESM studies.

### 9.4.1 Descriptive Statistics, HMM and Lag-0 Relationships

The descriptive statistics, such as histograms, and the lag-0 relationships obtained from the ESM times series (Figures F.6 and F.7 respectively in Appendix F.6) are essentially identical to those obtained from the ideal time series, depicted in Figures 9.4 and 9.6 in Section 9.3.1. This makes sense: we have exactly the same amount of data points, sampled from the same system as in the ideal time series case. The only difference is that in the ESM data set 900 time points are “missing” between each measurement of the ESM time series. However, because lag-0 relations do not pick up on any temporal dependence, the lower sampling frequency does not affect the lag-0 relations. While this suggests that lag-0 relations are robust against low sampling frequency, it also puts their utility to infer the dynamics of an underlying dynamical system into question.

The parameter estimates obtained by fitting a two-component mean-switching Hidden Markov Model on the ESM dataset were

$$\mu_1 = \begin{pmatrix} 1.47 \\ 1.46 \\ 4.71 \\ 4.71 \end{pmatrix}, \sigma_1 = \begin{pmatrix} 0.41 \\ 0.41 \\ 0.63 \\ 0.63 \end{pmatrix}, \mu_2 = \begin{pmatrix} 4.71 \\ 4.71 \\ 1.47 \\ 1.47 \end{pmatrix}, \sigma_2 = \begin{pmatrix} 0.64 \\ 0.64 \\ 0.41 \\ 0.41 \end{pmatrix}, A = \begin{matrix} & C_1 & C_2 \\ \begin{matrix} C_1 \\ C_2 \end{matrix} & \begin{pmatrix} 0.915 & 0.085 \\ 0.090 & 0.910 \end{pmatrix} \end{matrix}.$$

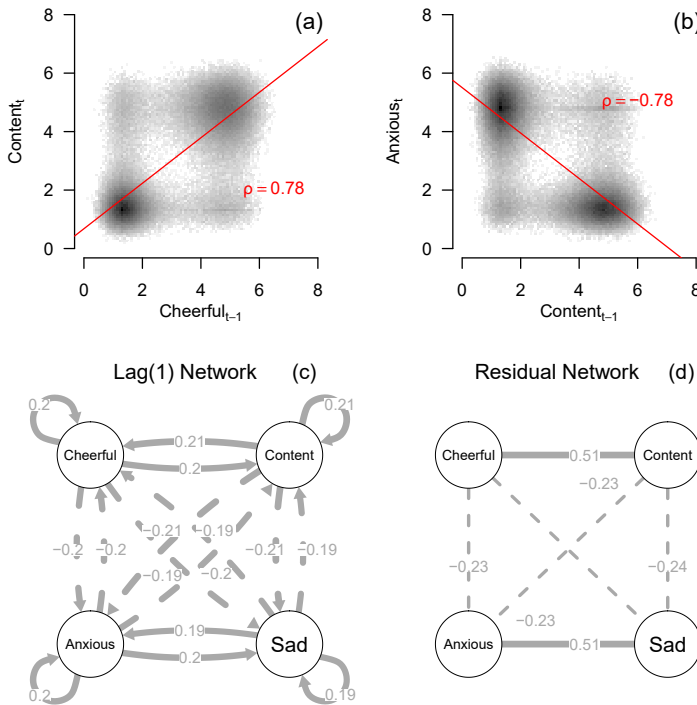
and the predicted states for two weeks of the time series are shown in Appendix F.6 Figure F.8.

We see a very similar pattern of results as obtained from the HMM fit to the ideal time series in Section 9.3.2, with the means and standard deviations of state 1 and state 2 reflecting the unhealthy and healthy states respectively. However, the parameters of the estimated transition matrix  $A$  for this time series show substantially higher switching probabilities,  $a_{12} = .085$  and  $a_{21} = .090$ . As we can see from Figure 9.3, although the sub sampled ESM time series contains only 224 observations for a two-week period, rather than 201600, the sampling frequency is still high enough to capture each of the 17 switches between states in this period. That means that, although the amount of transitions that occur over a period of time remains the same, the number of measurement occasions between any two transitions is lower, which results in a higher transition probability. We can see that this higher transition probability captures the number of transitions over two weeks quite accurately — the model predicts between  $224 \times .090 \approx 20$  and  $224 \times .085 \approx 18$  switches on average over a two week period. As such, the HMM fitted on the ESM time series still allows us to estimate the location of and variance around the two fixed points (characteristics 2 and 3), and approximate the frequency of transitions between these two fixed points (characteristic 4). In fact, the transition probabilities appear to be even more accurate than the ideal case — most likely this numerical imprecision in the ideal case is because the number of transitions relative to total time series was so low that slight changes in the

transition probability value lead to very different prediction about the number of transitions over 201600 time points.

### 9.4.2 Lag-1 Relationships and VAR model

When analysing the lagged relationships in the ESM time series, we begin to see some striking differences from the analysis of the ideal time series: panels (a) and (b) of Figure 9.11 show the *marginal* relationship of Content observed at time  $t$  with Cheerful at the previously observed time point  $t - 1$ , and Anxious at time  $t$  with Content at time  $t - 1$ , respectively. Focusing on panel (a), we see that the density of the lagged variables takes on a square-like shape, and each quadrant seems to be filled with a roughly circular density. This is in contrast to the density of the lagged relationships in the ideal data displayed in Figure 9.6 in Section 9.3.4, which was described by two elliptical shapes at the two fixed points.



**Figure 9.11:** Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, spaced one measurement occasion apart (i.e., at a lag of one but with 90 minutes between measurements) for the ESM dataset. The red line indicates the best fitting regression model. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence, at a lag of one. Panel (c) displays the matrix of lagged regression parameters, estimated from a VAR(1) model, as a network, and panel (d) displays the partial correlation matrix of the residuals of the VAR(1) model as a network. This latter network is often referred to as the “contemporaneous” network.

How can we explain this pattern? In the ideal data, the two elliptical shapes indicate that  $\text{Content}_{t-1}$  and  $\text{Content}_t$  tend to be near the same fixed point (two shapes), and that the two variables are positively correlated (elliptical shape). Now, in the ESM time series, we still have most of the density in the upper-right and the bottom-left quadrant, indicating that if  $\text{Content}_{t-1}$  is at the healthy (unhealthy) fixed point, it is very likely that  $\text{Content}_t$  is also at the healthy (unhealthy) fixed point (noting that  $t - 1$  now reflects a 90 min instead of 6 second time interval). However, we now also observe density at the top-left and bottom-right quadrant. These densities represent the situation in which  $\text{Content}_{t-1}$  is in the healthy (unhealthy) state, but  $\text{Content}_t$  is in the unhealthy (healthy) state. This situation is created when a switch between states falls within the 90min interval between two ESM measurements. Next, we focus on the shape of the density *within* each of the quadrants: we see that each of the densities have roughly a circular shape, which indicates that  $\text{Content}_{t-1}$  and  $\text{Content}_t$  are uncorrelated at each fixed point. This makes sense: in the ESM time series 900 time points are missing between each pair of measurement, which means that there is essentially no temporal dependence anymore between the variables. The relationship between  $\text{Anxious}_{t-1}$  and  $\text{Content}_t$  can be explained in an analogous way. Before fitting the VAR model below, this already shows us that it is futile to recover the microdynamics of the bistable system from these data.

Panel (c) of Figure 9.11 displays the network of estimated lagged regression coefficients from a VAR(1) model fit to the data. If we were to use these to infer the microdynamic characteristics of the system, we would manage to recover the signs of effects between variables: Negative, suppressing effects between-valence, and positive, reinforcing effects within-valence (characteristic 5, see Section 9.2.3). However, we again fail to recover the relative size of the suppressing and reinforcing effects. In fact, in this case, all of the estimated auto-regressive and cross lagged effects have approximately equal absolute value  $|\hat{\phi}_{ij}| \approx 0.2$ . This means that we also fail to recover the strong auto-regressive relationships encoded by the  $r$  parameters in the data generating model, and reflected by the strong auto-regressive effects estimated by the VAR(1) model in the ideal setting. In panel (d) we can see that, as was the case for the ideal time series, we obtain positive residual partial correlations within-valence and negative residual correlations between-valence, although the magnitude of these correlations is now quite high,  $\hat{\theta} = 0.51$  and  $\hat{\theta} = -0.23$  respectively. In addition, note that the residual variances of each variable in the model is considerably higher than the ideal case, and approximately equal for all variables ( $\hat{\Psi}_{ii} \approx 1.1$ , explained variance  $\approx 0.62$ ). As we would expect, taking only every 900th measurement from the ideal time series means that the predictive power of the VAR model decreases.

How can we reconcile these parameter estimates with what we know of the underlying bistable system? Although we may be tempted to interpret the VAR(1) parameters as reflecting the microdynamic structure, we have already seen in the analysis of the marginal relationships above that the large time-interval between observed measurements means that such an interpretation would be incorrect. Instead, the VAR(1) parameter values in the present situation are fully determined by the global characteristics of the system. Essentially,

the estimated lagged relationships reflect that, at a time-scale of 90 minutes, the dynamics of the system from one observation to the next can be boiled down to just two possibilities: Either the entire system stays in the same state or it moves to the other state. Since 1) the most likely behaviour is that the system stays near the same fixed point, and 2) those two fixed points are defined as high-positive and low-negative emotions, or low-positive and high-negative emotions, we end up with positive within-valence relationships (e.g. if Cheerful now is near the high fixed point, it's likely that Content later will be too ) and negative between-valence relationships (e.g. if Anxiety now is high, it's likely that Content later will be low). All of the auto-regressive and cross-lagged relationships are of equal value, as essentially all variables have the same value in predicting what fixed point each other variable will be near at the next measurement occasion: enough time elapses between measurements occasions that even the auto-regressive effect is only as predictive as the cross-lagged effects. As noted above, this interpretation is also reflected in the joint densities in panels (a) and (b) of Figure 9.11: Each density takes the appearance of four quadrants of uncorrelated Gaussian distributions, indicating that the microdynamic dependencies present in the ideal times series are totally absent from the ESM time series.

In summary, having a realistic sampling frequency results in the VAR(1) model providing even less information about the characteristics of the bistable model than in the ideal scenario. The longer time-interval between observations implies that interpreting VAR(1) parameters as reflecting truly microdynamic behaviour would be incorrect: parameters interpreted as reflecting microdynamics in fact have to be interpreted as reflecting the global characteristics of the system. Although the sign of the microdynamic relationships (characteristic 5) is recovered, in this instance it happens that the pattern of microdynamic relationships has the same valence as the pattern of relationships at a longer time-scale, that is, the movement of the process between fixed points. Thus, while in the ideal case the VAR parameters were a mixture of the microdynamics (around each fixed point), and global characteristics (i.e., position of the two fixed points), in the ESM time series these parameters are only reflective of the global characteristics.

### 9.4.3 Threshold VAR Model

We saw in the previous sections that inferring global characteristics using ESM data was somewhat successful, but that inferring microdynamic characteristics using a VAR(1) model was impossible. For the ideal time series, we saw that the threshold VAR(1) model was in principle able to capture some microdynamic and some global characteristics, and so in this section we examine how well that performance generalises to our emulated ESM data. We use the same thresholding variable (Cheerful,  $X_1$ ) and model specification as described in Section 9.3.5.

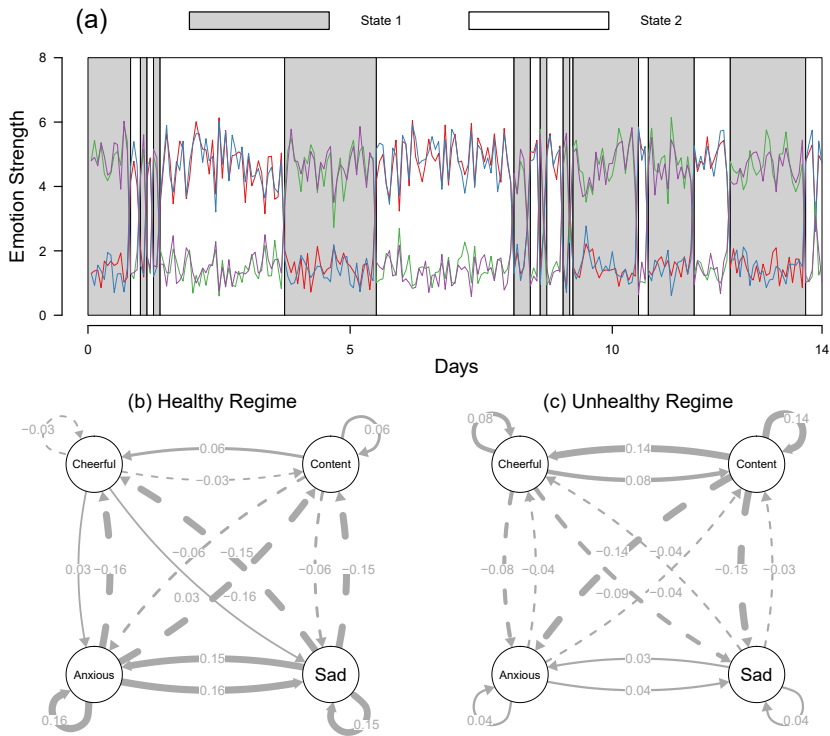
Figure 9.12 displays the main results from the TVAR(1) model estimated on the ESM data. The estimated threshold is  $\hat{\tau} = 2.796$ , very close to the estimated threshold in the ideal case, and from panel (a) of Figure 9.12 we can see that this threshold value does well in separating the time-series into the healthy and unhealthy states. Inspecting the lagged networks for each regime in panels (b)

and (c) of Figure 9.12 we see a similar general pattern of results as the lagged networks for the ideal time series in Figure 9.8 in section 9.3.5: In the healthy regime, the negative variables have much stronger cross-lagged effects on the positive variables, and vice versa for the unhealthy regime. However, we see even more differences between regimes in this case than we did for the ideal time series. For instance, in the healthy regime, the within-valence and auto-regressive relationships for the negative variables is much stronger than for the positive variables, a pattern which is flipped for the unhealthy regime. In both regimes, the within-valence cross-lagged parameters are roughly equal to the auto-regressive effects of the variables involved. The estimated means of each regime are given as  $\hat{\mu}_2 = \{4.31, 4.31, 1.87, 1.87\}$  for the healthy state and  $\hat{\mu}_1 = \{1.74, 1.71, 4.44, 4.62\}$  for the unhealthy state.

We can see from this that the TVAR(1) model for the ESM data succeeds in recovering some global characteristics of the system. Specifically, the estimated mean vectors capture approximately the position of the two stable fixed points (characteristic 2), and the estimated threshold correctly captures the position of the unstable fixed point in the Cheerful dimension. However, the recovery of this characteristic comes with the same caveats as described in Section 9.3.5: The use of a univariate threshold for this particular configuration of the true system happens to be a good approximation of the unstable fixed point in multi-dimensional space.

Regarding the microdynamics, the lagged parameters in each regime approximately capture that there are reinforcing effects within valence and suppressing effects between valence (characteristic 5). Otherwise, however, the recovery of microdynamic relationships performs worse than for the ideal time series, as expected. As was the case for the VAR(1) model, the regime-specific lagged parameters here again reflect global characteristics at the 90 minute time-scale rather than microdynamics: Partitioning the joint densities in panels (a) and (b) of Figure 9.11 using a threshold does not aid us in any way to reproduce microdynamic dependencies which are absent due to the low sampling frequency. Thus, the asymmetry in parameter values across regimes has to be a function of the global characteristics, influenced by both the different variances around the fixed points in each state (i.e., high variance for positive emotions, low variance for negative emotions in the healthy state, and vice versa) and those observations which jump from one fixed point to the other across consecutive measurement occasions, as discussed in the previous section.

As we did throughout Section 9.3, we could evaluate how well this model describes the bistable system by generating data from it. Notably, the dynamics defined by  $\Phi^{(1)}$  and  $\Phi^{(2)}$  reflect an *unstable* system in both regimes: The eigenvalues of both contain a value outside the unit circle (i.e., with absolute value greater than one) (Hamilton, 1994). This means that, if we were to generate data using these parameters, the time series would always diverge (i.e., values of variables go to infinity). This instability also precludes us from making any statement regarding the variance of positive and negative emotions in each regime (characteristic 3), as the long run variances implied by the model are infinite. As such, we can say that overall, the set of estimated parameters for the TVAR(1) based on the



**Figure 9.12:** Panel (a) shows the first two weeks of the time series, with observations shaded in either grey or white as a function of whether  $x_{1,t-1}$  is above or below the threshold  $\hat{\tau} = 2.796$ . Panels (b) and (c) show the estimated VAR(1) parameters as lagged networks in the healthy (white) and unhealthy (grey) regimes respectively.

ESM time series are a poor characterisation of the microdynamics of the model at any time-scale.

In summary, the TVAR(1) model fitted on the ESM time series still picks up a global characteristic of the system, but the recovery of microdynamic characteristics fails. In fact, the relationship between the estimated lagged parameters and the characteristics of the system was much more opaque than in the ideal data case, and our ability to generalize from the estimated parameters to the behaviour of the system at any time scale was considerably worse than in the ideal case. Again here, we should note that the only difference between the ideal and ESM time series is the sampling frequency. Fundamentally, the results here indicate that, if we do not have a sufficiently high sampling frequency, then fitting increasingly complex models, or extensions to simpler models such as the TVAR(1), does not aid us in recovering the characteristics we are interested in: Even when we have an arbitrarily large *number* of observations, we fail to recover basic characteristics of the microdynamics due to the spacing between measurements.

## 9.4.4 Differential Equation Model Building

In this section we will examine whether the DE model-building procedure described in Section 9.3.6 also succeeds in recovering the bistable system when applied to the emulated ESM dataset. Recall from Section 9.3.6 that for the ideal time series, this method succeeded in recovering the microdynamics of the system: The global characteristics were also considered to be recovered as the global characteristics implied by the estimated model (bistability, position of fixed points) matched up with the actual global characteristics of the underlying system.

### 9.4.4.1 Model Building Procedure

Similarly to Section 9.3.6 we first estimate the derivatives directly from the data by differencing the time series, and then search for the best fitting model by fitting a series of regression models with increasing complexity. Table 9.3 displays the fit of seven increasingly complex regression models, evaluated using the mean out-of-bag proportion of explained variance  $R^2$ .

Model A ( $R^2 = 0.13991$ ), Model B ( $R^2 = .16827$ ) and Model C ( $R^2 = 0.16928$ ) are the same models as introduced in Section 9.3.6. However, since we did not observe a clear drop in  $R^2$  as we increased model complexity from Model A to Model C, we also assess the fit of four additional models. Model D adds cubic main effects  $x_1^3, \dots, x_4^3$  as predictors, increasing the model fit to  $R^2 = 0.19455$ . Model E adds four three-way interactions ( $x_i \times x_j \times x_k, i \neq j \neq k$ ) to this, further increasing the model fit to  $R^2 = 0.19940$ . Adding yet more three-way interactions ( $x_i \times x_j \times x_k, \forall (i, j, k) \in p$ ) in Model F still increases model fit ( $R^2 = 0.19940$ ), as does adding all possible four-way interactions in Model G ( $R^2 = 0.20420$ ). As it is not possible to specify more unique product interaction terms, we consider Model G to be our final model.

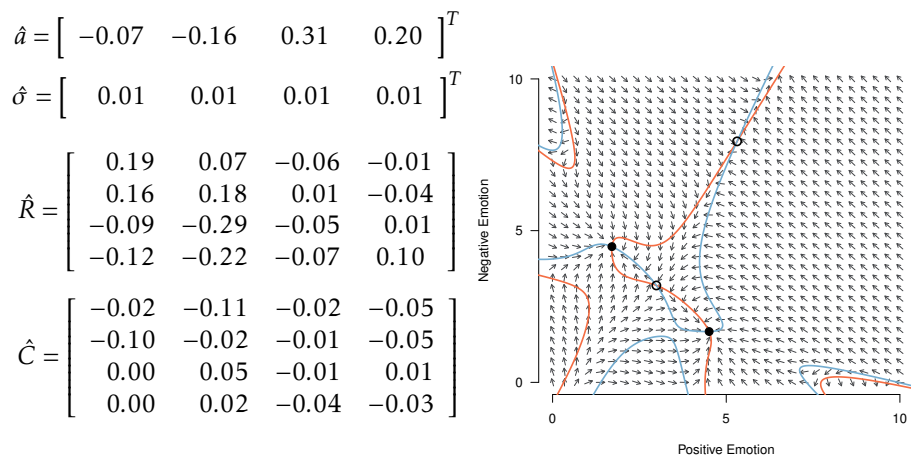
Model	$\frac{dx_{i,t}}{dt} \sim a + r_i x_i + \dots$	$q$	$R^2$
A	$\sum_{j \neq i} r_j x_j$	5	0.13991
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	0.16827
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	0.16928
D	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3$	19	0.19455
E	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{j \neq k \neq l}^p \zeta_j (x_j x_k x_l)$	23	0.19801
F	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l)$	35	0.19940
G	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l) + \sum_{(j,k,l,m)}^p \eta_j (x_j x_k x_l x_m)$	70	0.20420

**Table 9.3:** Model fit results for each of the seven models described in text, for the emulated snapshot ESM data. The second column gives the model equation for each variable,  $q$  denotes the number of parameters estimated per univariate regression model, and the final column indicates the mean proportion of explained variance  $R^2$ , calculated on the hold-out sets of a 10-fold cross-validation scheme (for details see Appendix F.5)

### 9.4.4.2 Dynamics and Data Generated by Final Model

Clearly, the model-building procedure for the emulated ESM data failed to recover the functional form of the true bistable system. Furthermore, we have arrived at a final model which is so complex ( $4 \times 70 = 280$  vs.  $4 \times 6 = 24$  parameters in the true model) that it is close to uninterpretable. In theory we could continue adding complexity to the model in the form of non-linear transformations or spline functions, which we know to be absent from the data generating mechanism, but which may increase fit. However, this would make the model even more difficult to interpret.

In the left panel of Figure 9.13 we present the estimated parameters that are also contained in the true model, with full parameter estimates and standard errors shown in Appendix F.5.3. We can see that the estimates deviate widely from the parameters in the true bistable system. In addition, the estimated parameters in the  $C$  matrix fail to capture the sign and relative ordering of all parameters in the true  $C$  matrix, though a full evaluation of whether suppressing and reinforcing effects of different sizes are present (i.e., characteristics 5 and 6) is infeasible due to the large number of parameters present in the model. Thus, we can say that this approach fails to recover the microdynamics of the system at least to the degree that they can be interpreted.



**Figure 9.13:** Left panel: the parameters estimated from the snapshot ESM time series. Right panel: the vector field defined by the estimated parameters. Solid points indicate stable fixed points and empty points indicate unstable fixed points. The solid lines indicate the values at which derivative of positive emotion (orange) and negative emotion (light blue) is equal to zero. At the points at which the two lines meet, both derivatives are equal to zero and the system remains in this (stable) state.

While the system did not recover the microdynamics in the sense that it captures the qualitative characteristics of the true bistable system, it could still be the case that this more complex system exhibits global characteristics that are similar to the true bistable system. Similarly to Section 9.3.6, we can evaluate



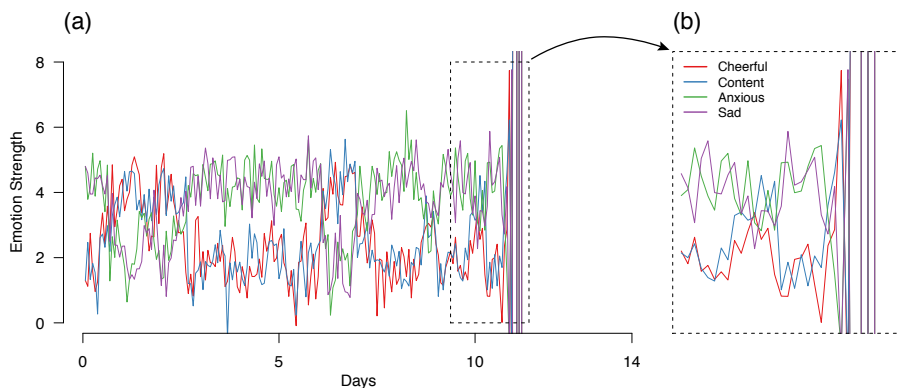
these dynamics by inspecting its vector field, shown on the right-hand side of Figure 9.13. As in the vector field obtained from the ideal data (Figure 9.9), the intersections of the two solution lines (blue and orange) indicate the position of the different fixed points in the shown range of the state space. These fixed points are further denoted by dots, with filled dots indicating a stable fixed point, and empty dots indicating an unstable fixed point.

We can immediately see from Figure 9.13 that the stability landscape is much more complex than the one of the true bistable system, with high-degree polynomial solution lines, and with four rather than three fixed points. Interestingly, the system correctly identifies (1) that there are two stable fixed points relating to healthy state ( $x_1 = x_2 = 4.51, x_3 = x_4 = 1.67$ ) and unhealthy state ( $x_1 = x_2 = 1.71, x_3 = x_4 = 4.47$ ) and (2) that there is an unstable fixed point approximately half-way between those two ( $x_1 = x_2 = 2.99, x_3 = x_4 = 3.19$ ). Despite having an entirely different functional form, the estimated model *does* capture two stable fixed points (characteristic 1) and the approximate position of those fixed points (characteristic 2). This shows that Model G performs well in capturing the characteristics of the system for emotion values that were observed in the time series, that is, near the two stable fixed points.

Crucially, however, we cannot say that this system recovers the global dynamics of the true system, not least because the system contains an additional unstable fixed point at ( $x_1 = x_2 = 5.31, x_3 = x_4 = 7.94$ ), which is not present in the true bistable system. The presence of this unstable fixed point means that if, for instance, both negative and positive emotions take on a high value simultaneously, then the system diverges (i.e., the values of variables go to infinity). If we examine the behaviour of the system even further outside the range of observed values ( $-\infty > X > 0$  and  $10 < X < \infty$ ) even more fixed points and regions of stability and instability can be found. We can further demonstrate these dynamics by generating data from Model G. Figure 9.14 shows a time series generated from the difference-form of Model G (i.e., with a step size equal to that of the observed data).<sup>4</sup> We see that the process moves between the healthy and unhealthy fixed point for the first ten days, exhibiting the bistable behaviour we see in the true system. However around the eleventh day, the stochastic input is large enough to move the system to an unstable region in the vector field and which leads the system to diverge.

---

<sup>4</sup>This is obtained by re-fitting the differential equation using the unscaled difference  $x_{i,t+1} - x_{i,t}$  as the outcome variable, leading to equivalent results with parameters approximately scaled by  $dt = 90$ . The residual variance used is the estimated residual variance scaled down to .65 the magnitude, to account for the non-normal residual distribution. Using the estimated residual standard deviation results in shocks which immediately move the system into an unstable region.



**Figure 9.14:** Data generated from the estimated DE model, with the same initial values as the “ideal” data

Note that the complexity of the final model here is not a result of over-fitting the data because we performed model selection based on the out-of-bag  $R^2$ , an approximation of the out-of-sample  $R^2$ . Rather, the complexity of this model can be attributed to two factors. First, due to the low sampling frequency, our approximation of the derivative at each point in time is poor. The second, as we discussed in Section 9.3.6.3, is that given the spacing between observations, the best one can hope for is to approximate the integral solution to the data generating equation, which is likely of a highly complex functional form. The ability of the misspecified Model G to reproduce some characteristics in regions where we have observed data can be attributed to the high flexibility afforded by the many non-linear terms. In that sense, this behaviour is highly comparable to the problem of using a high-degree polynomial regression model to make predictions outside of the range of observed values. The vector field in Figure 9.14 is constructed by obtaining predicted values for the derivatives across a grid of input values: As such, it is unsurprising that the vector field is accurate where the input values are close to the observed data, and inaccurate elsewhere.

In summary, we do not at all recover the functional form or parameters of the system; we do recover some of the global characteristics and behaviour of the system in the region where we have observations, capturing that there are two stable fixed points and one unstable fixed point, and their locations. However, the estimated model also implies the presence of at least one extra unstable fixed point, which has major implications for the dynamics of the model, implying divergent behaviour. Thus, the estimated model implies fundamentally different microdynamic and global characteristics. Based on the simulated data in Figure 9.14, it does not seem that we correctly capture the variability around these fixed points, or the frequency of transitions, as any reasonable simulation of data from this model eventually leads the system to diverge. Crucially, we fail in recovering an interpretable approximation of the data generating model. As such, it is not feasible to assess whether there are truly suppressing effects between va-

lences and reinforcing effects within valences, or the relative size of these effects (characteristics 5 and 6).

### 9.4.5 Summary: Analysis of ESM Time Series

In this section, we aimed to investigate to which extent lowering the sampling frequency affects the the ability of our considered methods to recover the bistable system. Our findings are summarized in Table 9.4. Our main findings are that, in general, we remain able to recover global characteristics of the system using simple methods, but that we are completely unable to recover any of the microdynamics. Since the dependencies close in time are not present in the data anymore, the VAR model, the TVAR model and the DE-Estimation procedure have to fail to recover these dependencies. This is despite the fact that the time series we used in this section can be considered a highly idealized approximation to ESM time series, in terms of the number of observations and the quality of measurements, suggesting that sampling frequency is a fundamental barrier to inference which needs further investigation.

	Bistability (1)	Position (2)	Variance (3)	Transitions (4)	Suppr./Reinf. (5)	Relative Size (6)	Time-constant (7)
Data Visualization	✓	✓	✓	×	×	×	×
HMM	✓*	✓	✓	✓	×	×	×
Lag-0 / GGM	×	×	×	×	×	×	✓*
Lag-1 / VAR(1)	×	×	×	×	×	×	✓*
TVAR(1)	✓*	✓	✓	✓	×	×	×
DE-Estimation	×	×	×	×	×	×	×

**Table 9.4:** Summary of which method recovered which of the seven qualitative characteristics listed in Section 9.2.3 from the ESM time series. The first four characteristics are global dynamics, the last three are microdynamics. The check marks with asterisk indicate that the method includes the characteristic as a model assumption, and can therefore not be considered recovered from the time series.

The recovery of global characteristics was more successful: Using data visualization and the Hidden Markov Model it was still possible in principle to learn about the position, variance around and frequency of transitions between fixed points, and the threshold estimate from the TVAR model succeeded in capturing the unstable fixed point. Finally, the predictions made by the best-fitting differential equation model did allow us to get some tentative indication of bistable behaviour, and the possible location of stable fixed points. However, the resulting model suffered from a high degree of complexity, limiting both substantive interpretation and our ability to extrapolate the model parameters to predict the behaviour of the system under different conditions.

In summary, the results in this section call into question to what extent it is possible to investigate moment-to-moment microdynamics using data sampled at a rate typical of ESM studies. We have showed that interpreting model estimates from ESM time series as reflecting the microdynamics can be highly misleading, when the process of interest is varying at a higher frequency than the sampling frequency. Although the recovery of global characteristics is more promising, we remind the reader that the time series considered here is still highly idealized, with essentially infinite sampling size, and so the performance of these methods should be considered an upper bound on performance in any realistic situation.

## 9.5 Discussion

In this chapter we explored to what extent dynamical systems models can be recovered from psychological time series by investigating two successive questions: first, how well does a set of popular and more advanced methods recover (characteristics of) a basic bistable system with an ideal data set sampled at extremely high sampling frequency (every six seconds)? And second, how is the performance of each method affected when reducing the sampling frequency to one measurement every 90min, which is typical for ESM studies.

When analyzing the ideal time series we found that the popular VAR model (and the GGM fitted on its residuals) can in principle not recover the global dynamics of the true bistable system, and only recovers some of its microdynamics. However, we showed that descriptive statistics, data visualization and statistical models which are based on mixtures (the HMM and threshold VAR) were able to capture the global dynamics of the bistable system. The only method that recovered the full bistable system was a differential equation (DE) model building procedure. Reducing the sampling frequency from every six seconds to every 90 minutes affected the considered methods differently: The VAR model and its extensions no longer recover any microdynamics, and the DE-estimation procedure fails. However, descriptives, data visualization and appropriate statistical models still recover the global dynamics. Overall, our analysis therefore suggests that it is neither possible to estimate dynamical systems directly from realistic time series, nor is it possible to reliably infer its microdynamics from the parameter estimates of statistical models.

### 9.5.1 Implications for Studying Mental Disorders as Complex Systems

Our results raise fundamental questions about how to study mental disorders from a complex systems perspective. First, they show that it is unclear what exactly one can in principle conclude from statistical models estimated from psychological time series about an underlying dynamical system. Clearly, these models are always misspecified (i.e., do not include the true system as a special case), so one cannot hope to directly recover the underlying dynamical system. More surprisingly, however, recovering the qualitative characteristics of the true

system also turned out to be difficult: while it was possible to recover the global dynamics, no statistical model correctly recovered the microdynamics. For example, the VAR model fundamentally cannot capture the global characteristics (e.g. location of fixed points) of the true bistable system and only recovered some of its microdynamics (e.g., reinforcing vs. suppressing effect between two variables). This is a problem for the emerging framework of studying mental disorders as complex systems, because one is typically interested in the microdynamics (the “mechanics”) of a disorder because one hopes to intervene on them. In contrast, it is usually less clear how interventions can target global dynamics, since they can be seen as the aggregate behavior implied by the microdynamics. Especially the failure of the popular VAR model to correctly recover the qualitative nature of the microdynamics in the true model is concerning, because it calls into question whether it allows any reliable conclusions about an underlying dynamical system. It therefore seems to be an open question how useful VAR models and other statistical models are to studying mental disorders from a complex systems perspective.

Second, the analysis of the ESM time series raises the question of which process can be recovered with which sampling frequency. While we were still able to recover the global dynamics of the system, each method that provides some approximation of the microdynamics was strongly affected by sampling only every 90min instead of every six seconds: The qualitative characteristics of the VAR and TVAR models were even less in agreement with those of the true model, and the DE-estimation method, which was the only fully successful method in the ideal data case, returned a model with uninterpretable parameters and incorrect global- and microdynamics. Thus, our results suggest what also seems intuitive: it is impossible — or at least extremely difficult — to recover microdynamics at a time scale that is much smaller than the sampling frequency. This intuition is also in line with sampling theorems from the field of signal processing: for example, the Nyquist-Shannon sampling theorem states that a sine wave (a process much simpler than our bistable system) that completes one cycle within, say, 2 minutes, has to be sampled at least every minute to be recovered (e.g., Marks, 2012; Papoulis & Pillai, 2002). This suggests that it is futile to try use a time series sampled every 90 minutes to directly recover dynamics of emotions that operate on a time scale of seconds or minutes (Houben, Van Den Noortgate, & Kuppens, 2015) or even from moment to moment (Wichers et al., 2015). However, this also means that ESM time series can certainly be used to recover processes that unfold at a time scale of several hours or days.

To summarize, we identified two fundamental problems to studying mental disorders from a complex systems perspective: first, even with extremely high sampling frequency it is generally unclear how to make inferences from a statistical model to an unspecified dynamical systems model. Second, the sampling frequency of the data collection constrains the type of processes one can recover. Specifically, a process can only be recovered if the sampling frequency is sufficiently high. Clearly, these are profound problems every empirical discipline struggles with and no simple answers can be expected. Indeed, they might imply that studying some phenomena of mental disorders will always remain out of of

reach. That said, we believe that that much progress *can* be made by studying mental disorders as complex systems and that acknowledging and studying the above issues allows one to do so more efficiently. As a way forward, in the following section we suggest a new research strategy based on proposing substantively plausible dynamical systems, which opens up avenues to creatively tackle the two problems identified in this section.

### 9.5.2 Moving Forward: Proposing Plausible Dynamical Systems Models

A more abstract perspective on the first problem identified in the previous section is the following: we have parameters of a statistical model which we estimated from a time series sampled from some system, and we hope to infer some characteristics (e.g., global or microdynamics) of the data generating system from them. The problem, however, is that the mapping from parameters of statistical model to the parameters and structure (and the implied dynamics) of the true model is unknown. Thus, this inference cannot be made. The main reason this mapping is unknown is the trivial reason that no true dynamical system model is specified.

We propose that, in order to overcome this fundamental problem, researchers must begin the research process by proposing a “first guess” model of the dynamical system. While this is clearly difficult and the validity of this model should certainly be questioned, this approach has one major advantage: it is much clearer how to draw conclusions from descriptive statistics, data visualizations or statistical models about the underlying dynamical systems model. This is because one can generate time series from the “first guess” model and fit a statistical model of choice; that way, one always knows which statistical model is implied by the dynamical systems model. This implied model can then be compared to the model fitted to corresponding empirical data. If the implied model and the empirical model are in agreement, we have tentative evidence that the dynamical system model is correct; if not, we can use the nature of the disagreement to improve the dynamical system model. Clearly, this modeling approach, which is typical to more quantitative disciplines such as physics, chemistry and biology, is different to the statistical modeling framework most psychological researchers are familiar with. On the one hand these formal dynamical systems models are harder to build, since they cannot be estimated directly from the data. On the other hand, they are powerful enough to be plausible for complex phenomena such as mental disorders, and have additional benefits such as synthesizing knowledge, revealing unknowns, laying open hidden assumptions and the checking the internal consistency of a model (Epstein, 2008; Lewandowsky & Farrell, 2010; Smaldino, 2017).

This modeling approach also allows to tackle the problem of sampling frequencies that are too low to recover the process of interest directly, as one can generate a time series from the specified dynamical systems model and reduce the sampling frequency to a level that is also available in empirical data. Then, similarly to above, one can again compute the statistical model of choice that

is implied by the dynamical systems model with a given sampling frequency, compare it to the corresponding model fit on empirical data, and in the case of disagreement adapt the dynamical systems model accordingly. Of course, this approach is not a panacea: less information is available when the sampling frequency is low, which makes model identification more difficult. However, specifying an initial dynamical systems model allows one to gauge how difficult it is to recover a given type of process on a given time scale with a given sampling frequency.

In addition, starting out with a dynamical systems model also allows to study the *measurement function* that defines the mapping from the variables in the dynamical systems to the obtained measurements, a topic we only touched on briefly in this chapter. In our emulated ESM time series we took the measurement function to return the exact values of variables at the time point of measurement. However, different questions imply different measurement functions. For example, if the phrasing of a particular question refers to the entire period since the last measurement, one could instead define the measurement as a function of the variable values since the last measurement, such as the average. Next to formalizing which experiences an ESM question refers to exactly, defining a measurement function also allows to formalize known response and memory biases, such as the recency effect (Ebbinghaus, 1913/2013).

Finally, having a plausible dynamical systems model allows one to explicitly address a behavior that has been largely ignored in the psychological time series modeling literature: the fact that humans sleep. Sleep interacts with essentially everything physiological and psychological, is part of the definition of several mental disorders (e.g., Major Depression) and related to many more (e.g., Walker, 2017). Thus, for many mental disorders, it seems necessary for a plausible model to include sleep. This may also allow using existing data in new ways, because data around the “day-night shift” does not have to be excluded anymore, but instead can be used to test hypotheses about the sleep-related assumptions of the dynamical systems model.

Clearly, this brief outline of the proposed modeling approach leaves many important questions unanswered: Where should the initial “first guess” dynamical system come from? How to formalize different substantive aspects in a dynamical systems model? Which statistical models should one choose to test which implications of the dynamical systems model? Given some disagreement between predicted and observed statistics, how should one adapt the existing dynamical systems model? These and other questions are difficult ones and answering them requires the combined creativity of a large research community. Nonetheless, a more detailed account of our proposed new modeling approach would be desirable. However, since such a detailed account is beyond the scope of the present work, we address it in Chapter 11.

### 9.5.3 Limitations

Several limitations of our work require discussion. First, our goal was to explore to which extent one can recover (bistable) dynamical systems for mental disorder.

ders from psychological time series. However, we only studied a single bistable system. Therefore, it could be that the fundamental problems identified in this chapter and summarized in Section 9.5.1 are in fact a particularity of the chosen bistable system. This, however, seems extremely unlikely: First, because we identify the problems in this chapter as examples of well-known issues such as model misspecification and sampling systems with a sampling frequency that is sufficient for recovery. Second, the bistable system we chose is arguably the simplest bistable system for four variables one can find. Choosing a different model therefore results most likely in choosing a more complex model, and our intuition is that the methodological difficulties discussed in this chapter become more and not less relevant in such models.

Second, a more specific criticism of our bistable system could be that the time scale of the process is unrealistically small, and we therefore exaggerated the problem of recovering dynamics of psychological processes from ESM time series. We agree that it is possible that some psychological processes are easier to recover from ESM data than the dynamical system used in this chapter. Thus, strictly speaking, we only showed that it is impossible to recover a system if the sampling frequency does not appropriately match the time scale of the system. In principle, it is therefore an open question whether there is a mismatch between the time scale of the system of interest and the available sampling frequency. However, intuition — and the sampling theorems such as the one mentioned in Section 9.5.1 — strongly suggest that it is impossible (or at least very difficult) to recover a process that operates at a time scale of seconds or minutes from an ESM time series that is measured every 1.5 hours. Clearly, however, our investigation is only a first treatment of the important topic of sampling frequency, and much work on it is required to establish a tight connection between psychological time series and dynamical systems models.

Third, one could reverse the argument in the previous two paragraphs and argue that our model is so ideal that many analyses perform better than in most realistic applications. This is certainly the case for the Threshold VAR model, which performs well only because of the simple dynamics of the bistable system as we discussed in Section 9.3.5. Other examples are the descriptive statistics and data visualization which may not be as insightful if fixed points are closer to each other and if there is more noise in the system. Also, the two-step approach to estimating the differential equations in Section 9.3.6 may work less well for a more complicated model. Thus, we would agree with this assessment, however chose to use a simple bistable system in order to make the chapter more accessible to applied researchers.

Fourth, we analyzed a bistable system whose structural parameters do not change over time. However, much of the framework of considering mental disorders as complex systems is based on the idea that pathology is defined with respect to a structural change in the underlying system, and therefore structural change is of central interest. We expect that structural change renders the recovery of a system more difficult, and we therefore did not include this feature in order to keep the chapter at a reasonable length. However, we believe that future methodological research into how to recover such structural changes both in



principle and with realistic time series would be extremely helpful to better understand phenomena such as early warning signals (Scheffer et al., 2009; van de Leemput et al., 2014a) and more generally structural change in mental disorders.

Fifth, in order to estimate a differential equation from data, we took a rather simple two-step approach based on local linear approximation of the derivative (cf., Boker et al., 2010). This approach involves first estimating the derivative itself using scaled difference scores, and then using this derivative as an outcome variable in a regression model. While this method benefits from being extremely simple to implement, we could expect that it would perform poorly in the presence of low sampling frequency as the quality of the derivative approximation degrades (as we noted in Section 9.3.6.3 and observed in Section 9.4.4). There are multiple alternative approaches to estimating DE equations which we did not consider here. For example, approaches based on numerical integration of the DE equation during estimation, such as implemented in *dynR* (Ou et al., 2019) and *stan* (Carpenter et al., 2017) (with additional functionality in the *ctsem* package; Driver, Oud, & Voelkle, 2017) may in general perform better than the two-step procedure when the sampling frequency is low. However, for the analysis shown in the present chapter, neither the *ctsem* nor *dynr* package performed better than the two-step approach. In general, however, more research is needed to map out which method deals best with the problem of low sampling frequencies.

Lastly, throughout our chapter we studied how well certain analysis methods can recover the true bistable system *in principle*. We did this by studying the population properties of these methods, that is, the situation in which one has essentially infinite sample size, which we approximated with a huge number (201600) of measurements. This was necessary in order to study the more fundamental questions of (1) whether a given method can recover our bistable system in principle and (2) whether a given method can recover our bistable system based on a time series with realistic sampling frequency. We did this because it would be meaningless to study the performance of a method as a function of sample size, if the method already fails with infinite sample size. Clearly, however, to apply any of the methods we studied in practice, one has to know how reliable they are with which sample size, and much more research is necessary to map out these sample size requirements (e.g., Dablander, Ryan, & Haslbeck, 2019).

### 9.5.4 Summary

In this chapter we identified two fundamental problems involved in studying mental disorders from a complex systems perspective: first, it is generally unclear what to conclude from a statistical model about an unspecified underlying complex systems model. Second, if the sampling frequency of a time series is not high enough, it is futile to attempt to recover the microdynamics of the underlying complex system. In response to these problems, we proposed a new modeling strategy that takes an initial substantively plausible dynamical systems model as a starting point, and develops the dynamical systems model by testing its predictions. In this approach it is much clearer what we can learn from data and statistical models about an underlying dynamical system, and in addition it pro-

vides avenues to move the field forward by formalizing the sampling process, measurement, response and memory biases, measurement reactivity and the influence of sleep.

## **Acknowledgements**

We would like to thank Denny Borsboom, Fabian Dablander, Charles Driver, Peter Edelsbrunner, Jens Lange, Don Robinaugh, Noémi Schuurman, Arnout Smit and Lourens Waldorp for their feedback on an earlier version of this chapter. Especially, we would like to thank Don Robinaugh for many great discussions about a related project, which influenced the present chapter.

# A FORMAL THEORY OF PANIC DISORDER

---

## Abstract

The network theory of psychopathology posits that mental disorders are complex systems of mutually reinforcing symptoms. This overarching framework has proven highly generative but does not specify precisely how any specific mental disorder operates as such a system. We address this gap in the literature by developing a network theory of Panic Disorder and formalizing that theory as a computational model. We first review prior psychological theory and research on Panic Disorder in order to identify its core components as well as the plausible causal relations among those components. We then construct and evaluate a formal theory of Panic Disorder as a non-linear dynamical system. We show that this formal theory can explain a great deal, including individual differences in the propensity to experience panic attacks, key phenomenological characteristics of those attacks, the onset of Panic Disorder, and the efficacy of cognitive behavioral therapy. We also show that the theory identifies significant gaps in our understanding of Panic Disorder and propose a theory-driven research agenda for Panic Disorder that follows from our evaluation of the theory. We conclude by discussing the implications of the model for how we understand and investigate mental disorders as complex systems.

---

This chapter has been adapted from: Robinaugh, D., Haslbeck, J. M. B., Waldorp, L., Kosakowski, J. J., Fried, E. I., Millner, A., McNally, R. J., van Nes, E. H., Scheffer, M., Kendler, K. S. & Borsboom, D. (submitted). Advancing the Network Theory of Mental Disorders: A Computational Model of Panic Disorder. Preprint: <https://psyarxiv.com/km37w/>

## 10.1 Introduction

The network theory of mental disorders posits that symptoms cohere, in part, because of causal relations among the symptoms themselves (Borsboom, 2017). From this perspective, mental disorders are analogous to an ecosystem. They do not appear as a coherent whole because of a shared underlying essence, but because of the web of causal interactions among the features of the disorder (Kendler et al., 2011).

The notion that there are etiologically important causal interactions among symptoms has prompted the development of new methods for assessing the structure of relationships among symptoms (Marsman et al., 2015; Epskamp, Maris, et al., 2016; Epskamp, Rhemtulla, & Borsboom, 2017) and a host of empirical studies applying those methods across numerous psychiatric disorders (for an overview, see Fried & Cramer, 2017). Moreover, the core idea that there are causal relations among symptoms has expanded into an overarching theory of mental disorders, how they develop, and how they remit (Borsboom, 2017). However, network theory remains abstract. It provides a conceptual framework for thinking about mental disorders but does not posit specific relationships among symptoms. Empirical network studies provide information about these relationships, but are not rich enough on their own to fully inform a network theory, as this requires a substantively interpreted model: a model that does not merely statistically associate variables, but rather specifies the mechanisms through which variables influence one another. Consequently, the network approach has produced statistical models that suggest putative network structures, but no theories that posit precisely how any given mental disorder operates as a complex system of interacting symptoms.

We aim to address this gap in the literature by developing such a theory for Panic Disorder. Panic Disorder is a suitable starting point for several reasons. First, theories are about phenomena (Bogen & Woodward, 1988; Haig, 2005) and panic attacks are a robust phenomenon: a “stable, recurrent, and general feature of the world” (Haig, 2005, p. 374). Experiences resembling panic appear in medical consultation reports as far back as the mid-18th century (Coste & Granger, 2014) and have been described consistently in the medical literature since the late 19th century (Berrios, 1996; Dechambre, 1864)). Accordingly, these attacks are a suitable phenomenon about which to develop a theory. Second, Panic Disorder symptoms are structurally inter-connected in the network of symptoms from the Diagnostic and Statistical Manual (DSM; Boschloo et al., 2015), suggesting that these symptoms commonly co-occur and, thus, represent precisely the type of phenomenon that network theory seeks to explain. Third, theorists have posited causal relations among Panic Disorder symptoms (e.g., a mutually reinforcing relationship between panic attacks and avoidance behavior; Goldstein & Chambless, 1978). Indeed, some of these relationships are embedded in the disorder’s diagnostic criteria (e.g., to meet the diagnostic criterion, avoidance behavior must be related to panic attacks; Borsboom, 2008). Fourth, there is strong body of research on the etiology, phenomenology, and epidemiology of panic attacks (Barlow & Craske, 1988; McNally, 1994). Moreover, there are well-established

interventions that treat Panic Disorder (e.g., cognitive behavioral therapy; Barlow, 1997) and reliable ways to induce panic attacks (e.g., biological challenges; Gorman, Liebowitz, Fyer, & Klein, 1987). This work provides criteria by which to evaluate our theory, as any theory failing to account for these empirical findings will be found wanting. The prior literature is thus sufficiently rich to both inform theory development and to provide a basis for theory evaluation.

Our development of Panic Disorder theory will proceed as follows. In Section 10.2, we review theory and research on Panic Disorder, identifying its essential components and the posited functional relations among them. In Section 10.3, we integrate prior work and propose a formal theory of Panic Disorder as a complex system, formalizing the relationships among individual symptoms of Panic Disorder in a mathematical model. In Section 10.4, we implement this model in R, a freely available software environment for statistical computing (R Core Team, 2014). Computational modeling is an effective tool for theory development because it allows us to simulate the model's behavior and assess what the theory can and cannot explain (Epstein, 2008). We will show that the theory can explain a great deal, including core phenomenological qualities of panic attacks, individual differences in the vulnerability to panic attacks, the onset of Panic Disorder following an initial panic attack, and the efficacy of cognitive behavioral therapy for Panic Disorder. We will also show that explicating Panic Disorder theory in this way reveals significant gaps in our understanding. Moreover, the model fails to explain some key features of Panic Disorder. These shortcomings suggest further theory development is needed. In Section 4, we propose a theory-driven research agenda for Panic Disorder that follows from our evaluation of the model. We give particular focus to the need for further theory development and illustrate how such development could proceed using the model proposed here as well as a previously proposed mathematical model of panic attacks (Fukano & Gunji, 2012). Finally, in Section 5, we discuss the implications of the model for our understanding of mental disorders.

## 10.2 A Survey of Panic Disorder Theory and Phenomenology

The symptoms identified in diagnostic manuals provide a tractable starting point for identifying the components of a mental disorder's causal system (Borsboom, 2017). Current diagnostic criteria for Panic Disorder are rooted in work by Sigmund Freud from the late 19th century (Frances et al., 1993). In his work on anxiety neurosis, Freud described *angstfallen*: sudden attacks of anxiety characterized by “ideas of the extinction of life... or of a threat of madness” accompanied by intense somatic symptoms (Freud, 1962, pp. 93-94). The key features of his description can be traced from an early precursor to the DSM (Association & for Mental Hygiene. Bureau of statistics, 1918), through the DSM-II (American Psychiatric Association, 1968), to the first diagnostic criteria for anxiety neurosis (Feighner et al., 1972), where these attacks were required to include apprehension, fearfulness, or a sense of impending doom accompanied by at least four

of the following six somatic symptoms: dyspnea, heart palpitations, chest pain, choking or smothering sensations, dizziness, and paresthesias. In the DSM-III, these attacks were separated into the newly created Panic Disorder diagnosis (Kendler, 2017; McNally, 1994), but the criteria used by Feighner and colleagues were largely retained, defining panic attacks by the sudden and unexpected onset of two symptoms: somatic symptoms and fear (American Psychiatric Association, 1980).

The DSM symptoms of a panic attack provide a fuzzy delineation of panic phenomenology. The somatic symptoms include the experience of fear (e.g., the fear of dying) and fear is itself characterized by many of the somatic symptoms (American Psychiatric Association, 1980; Cacioppo et al., 2000; Cowen & Keltner, 2017). Consequently, these symptoms do not identify fully distinct components. In the surge of research that followed Panic Disorder's inclusion in the DSM-III, researchers instead began to use a closely related but somewhat cleaner delineation, identifying three components of panic phenomenology: physiology, cognition, and behavior (Barlow & Craske, 1988; Salkovskis, 1988; Griez et al., 1983). We will consider these to be the core building blocks of a panic attack.

## **10.2.1 Building Blocks of a Panic Attack**

### **10.2.1.1 Physiological Arousal**

Individuals experiencing panic attacks report the sudden onset of intense bodily sensations, most commonly heart palpitations, difficulty breathing, sweating, trembling, dizziness, and faintness (Barlow & Craske, 1988; Brown & Cash, 1990; De Beurs et al., 1994; Hibbert, 1984), sensations long recognized as products of the autonomic nervous system (Berrios, 1999). Consistent with patient reports, panic attacks that occur in the laboratory in response to placebo injections or relaxation procedures are associated with increases in autonomic nervous system activity, including heart rate, skin conductance, body temperature, and ventilation (Cohen, Barlow, & Blanchard, 1985; Goetz et al., 1993; Lader & Mathews, 1970). Ambulatory assessments have produced more equivocal findings, with elevations in heart rate, temperature, and respiration observed during many, but not all, patient-identified panic attacks (R. R. Freedman, Ianni, Ettedgui, & Puthezhath, 1985; Cameron, Lee, Curtis, & McCann, 1987; Hoehn-Saric, McLeod, Funderburk, & Kowalski, 2004; Margraf, Taylor, Ehlers, Roth, & Agras, 1987; C. B. Taylor et al., 1986). However, bodily sensations more consistently accompany attacks that patients rate as especially severe in these studies. Together, patient self-report and studies of physiology suggest that autonomic arousal is an important building block of prototypical panic attacks.

### **10.2.1.2 Perceived Threat**

The second building block concerns perceived threat. Individuals experiencing panic perceive themselves to be under threat. Early accounts of panic attacks describe uncontrollable worry and “ideas of the extinction of life” (Freud, 1962; Dechambre, 1864, pp. 93) and fear of dying or going crazy are among the DSM

diagnostic criteria for panic (American Psychiatric Association, 2013). In early cognitive studies, participants reported a fairly narrow range of thoughts that focused on perceived physical (e.g., heart attack), psychological (e.g., going crazy), or social (e.g., making a fool of oneself) consequences of the bodily sensations associated with panic (Hibbert, 1984; Ottaviani & Beck, 1987). Subsequent studies have shown that panic patients in non-Western cultures often fear additional bodily sensations in keeping with local (“folk”) understandings of physiology (Lewis-Fernández et al., 2011). For example, Cambodian refugees with Panic Disorder commonly dread sensations of orthostatic dizziness, interpreting such symptoms as signaling a potentially lethal episode of *kyol goeu* (“wind overload”; Hinton, So, Pollack, Pitman, & Orr, 2004; Hinton, Um, & Ba, 2001). Although “wind overload” and heart attacks are superficially distinct, they both concern seemingly uncontrollable bodily sensations associated with increased autonomic arousal (S. Taylor, 1994). Accordingly, the unifying theme of cognitions observed during panic attacks is the appraisal of arousal-related bodily sensations as a source or indicator of threat.

### 10.2.1.3 Escape Behavior

The third building block is behavioral. An “irresistible need to run” was noted in the earliest medical literature on panic attacks (e.g., Millet, 1884; as cited in Berrios, 1996) and the urge to escape is among the more strongly endorsed features of these attacks (Norton, Zvolensky, Bonn-Miller, Cox, & Norton, 2008). Indeed, some researchers have defined panic attacks in part by “the intense desire to escape or flee the situation” (Chambless, Caputo, Jasin, Gracely, & Williams, 1985, p. 42). However, behavioral attempts to prevent feared consequences are not confined to overt efforts to escape. “Patients sit down, hold onto walls. . . and generally engage in behaviors which they believe may abort imminent disaster”, (Salkovskis, 1988, p. 130). These behaviors can be subtle and idiosyncratic but share the intent of mitigating the anticipated consequences of arousal-related bodily sensations (Salkovskis, 1991).

## 10.2.2 The Building Blocks of Panic Disorder

Beginning with DSM-IV, recurrent panic attacks alone were no longer sufficient for the diagnosis of Panic Disorder (American Psychiatric Association, 1994), one also had to endorse persistent concern about additional attacks, worry about the implications of an attack, or change in behavior related to the attacks (American Psychiatric Association, 1994). These expanded criteria, retained in the DSM-5 (American Psychiatric Association, 2013), suggest two additional building blocks of Panic Disorder.

### 10.2.2.1 Persistent Concern

“Concern” and “worry” about panic attacks suggests a cognitive component distinct from that observed in the panic attack itself. Some have argued for the importance of “persistent concern”, noting that individuals with recurrent panic

attacks tend to believe that panic-related bodily sensations are dangerous (Craske et al. (2010, p. 104); McNally (1994, p. 8)). Episodic thoughts about panic attacks and their consequences outside the context of panic attacks themselves (cf. “worry”) are similarly endorsed by a large portion of those who experience panic attacks (Craske et al., 2010). In the remainder of this article, we will focus on the more enduring beliefs associated with panic and panic-related arousal.

#### 10.2.2.2 Avoidance Behavior

The avoidance symptom of Panic Disorder suggests a behavioral component distinct from that observed in the panic attack itself. The “urge to escape” during panic attacks concerns a threat that is already present. In contrast, avoidance is preemptive. Individuals with reoccurring panic attacks refrain from activities that increase physiological arousal (e.g., drinking coffee) and avoid situations where a panic attack may occur or where its consequences may be especially severe (e.g., the middle of a crowded theater). This situational avoidance is closely related to the concept of agoraphobia: the fear of situations where escape would be difficult or help unavailable in the event of a panic attack or panic-like bodily sensations (American Psychiatric Association, 2013). Although agoraphobia can occur in the absence of panic attacks, agoraphobic avoidance remains an important component of the Panic Disorder syndrome (Asmundson, Taylor, & Smits, 2014; Wittchen, Gloster, Beesdo-Baum, Fava, & Craske, 2010).

#### 10.2.3 Functional Relations among Building Blocks

The components of a panic attack are intimately linked. Indeed, they are largely defined in relation to one another. Perceived threat during panic is about the bodily sensations tied to physiological arousal. Physiological arousal is the body’s evolved response to perceived threat, preparing one to escape, the very behavior observed during panic (Barlow & Craske, 1988; Cannon, 1916). Perhaps unsurprisingly then, numerous psychological theorists have postulated causal relations among the components of a panic attack. Moreover, they have argued that those relations figure prominently in Panic Disorder’s etiology. In the remainder of this section, we provide a brief overview of these psychological theories.

##### 10.2.3.1 The Vicious Cycle of Panic Attacks

The most well-known theory of Panic Disorder is Clark’s cognitive model (Clark, 1986; Roth, Wilhelm, & Pettit, 2005). In this model, panic attacks occur when a person misinterprets benign arousal-related bodily sensations as portents of danger (e.g., interpreting increased heart rate as a sign of an impending heart attack; Clark, 1986). This “catastrophic misinterpretation” leads to a perception of threat which, in turn, increases arousal, producing a self-amplifying feedback loop that culminates in a panic attack (Clark, 1986, p. 463).

Clark’s cognitive model is neither the only nor the first theory positing this “vicious cycle” (Roth et al., 2005). As far back as 1937, the physician T.A. Ross



offered a remarkably similar formulation (Ross, 1937, p. 24) and in the years before and after Clark's model was published, numerous theorists offered variations on the vicious cycle theory of panic attacks (Barlow & Craske, 1988; Beck, 1988; Ehlers & Margraf, 1989; Margraf, Ehlers, & Roth, 1986; Rapee, 1987; van den Hout & Griez, 1983). In an early example of such a theory, Goldstein and Chambless (1978) argued that individuals with recurrent panic attacks and agoraphobia have learned to fear the bodily sensations associated with arousal, rendering them vulnerable to the vicious cycle between arousal and perceived threat. A similar argument was advanced by Reiss, McNally, and colleagues, who used the term anxiety sensitivity to denote the belief that the bodily sensations associated with anxiety may have harmful consequences (McNally, 1990; Reiss & McNally, 1985). Anxiety sensitivity was characterized as a dispositional variable that can exist prior to and independent of Panic Disorder and, thus, can help explain individual differences in the propensity to experience panic attacks and Panic Disorder. Individuals who believe anxiety-related bodily sensations to be harmful, they posited, are more likely to interpret such sensations as portents of danger and, consequently, are more likely to become entrapped in the vicious cycle that culminates in a panic attack. There is now considerable evidence to support this position (McNally, 2002). Individual differences in anxiety sensitivity predict both the likelihood of panic in response to the induction of arousal-related bodily sensations as well as the development of unexpected, "spontaneous" panic attacks over time (Schmidt, Lerew, & Jackson, 1997, 1999; Schmidt, Zvolensky, & Maner, 2006). These findings suggest that anxiety sensitivity acts as a moderating variable, strengthening the effect of arousal-related bodily sensations on perceived threat and, thereby, amplifying the vicious cycle that drives panic attacks.

### 10.2.3.2 The Vicious Cycle of Panic Disorder

Reiss and McNally (1985) argued that a history of panic is not necessary for the development of elevated anxiety sensitivity. However, they did identify panic attacks as one path by which an individual may develop the belief that anxiety-related bodily sensations have harmful consequences (Reiss, Peterson, Gursky, & McNally, 1986). In doing so, like Goldstein and Chambless (1978), they posit a second vicious cycle: a cycle that operates on the components of Panic Disorder, rather than on the components of a panic attack. Here, panic attacks strengthen the belief that bodily sensations are dangerous, thereby increasing vulnerability to panic attacks.

Bouton, Mineka, and Barlow (2001) expanded on this vicious cycle, positing that conditioning episodes play a fundamental role in the development of Panic Disorder. The terrifying nature of an initial panic attack conditions the bodily sensations associated with the early stages of the attack. As a result, these bodily sensations themselves become signals of a possible impending attack. As Bouton and colleagues note, "individuals who go on to develop Panic Disorder would learn anxiety sensitivity or, more specifically, that somatic symptoms are potentially dangerous" (Bouton et al., 2001, p. 22).

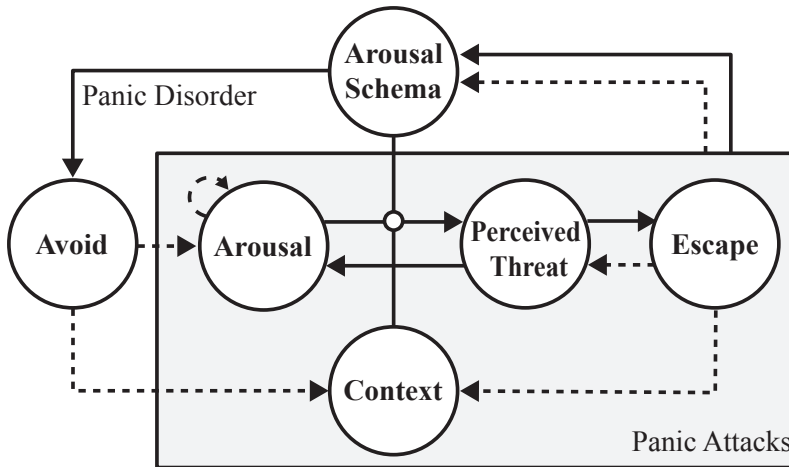
A causal effect of panic attacks on beliefs about panic attacks can account for how an initial attack may lead to the development of Panic Disorder. However, it raises a critical question: why do those who develop Panic Disorder learn that the bodily sensations associated with a panic attack are dangerous rather than harmless? The repeated failure of a catastrophe to materialize should lead to more accurate beliefs regarding arousal, yet beliefs about their danger arise and persist. Clark and colleagues have posited several processes that maintain such beliefs (Clark, 1999), including the possibility that escape behaviors engaged before or during an attack may shield catastrophic beliefs from refutation (Salkovskis, 1991; Salkovskis, Clark, Hackmann, Wells, & Gelder, 1999). The absence of the feared consequence is attributed to the escape behavior, leaving intact the belief that bodily sensations are dangerous. Learning theorists have similarly argued that these behaviors act as inhibitors, predicting an absence of the feared consequence. As such, they serve to eliminate the discrepancy between prediction and observation, and thereby prevent the individual from learning a more accurate and benign prediction about the consequences of arousal-related bodily sensations (Bouton et al., 2001). Hence, across theories, cognitive and behavioral theorists have argued that escape behavior plays a critical role in determining whether substantially elevated arousal strengthens or disconfirms the belief that arousal is threatening.

#### 10.2.4 Summary

Our survey of panic phenomenology identified five core building blocks of Panic Disorder. Panic attacks include *physiological arousal*, *perceived threat*, and *escape behavior*. Panic Disorder additionally includes persistent concern about panic-related arousal (cf., *anxiety sensitivity*) and *avoidance behavior*. Our brief survey of psychological theories suggests key functional relations among the components of panic attacks and Panic Disorder that, together, form three interlocking feedback loops: (1) a positive-feedback loop between physiological arousal and perceived threat (i.e., the “vicious cycle” that gives rise to panic), (2) a negative feedback loop between perceived threat and escape behavior, and (3) a learning feedback loop in which panic attacks either strengthens the belief that arousal is dangerous (a positive feedback loop) or disconfirms such beliefs (a negative feedback loop), with the type of learning that occurs hinging on the presence of escape behavior. In the next section, we integrate these feedback loops and propose a model of Panic Disorder.

### 10.3 A Model of Panic Disorder as a Non-Linear Dynamical System

An integrated model of Panic Disorder derived from our survey of the literature appears in Figure 10.1:



**Figure 10.1:** A Causal Diagram of Panic Disorder. Circles represent individual components of the causal system: the five components identified in our survey of the literature plus a “situational” context variable (in our model we use the broader term “arousal schema” rather than anxiety sensitivity as focused on in our survey of the literature, see text for further details). Panic attack components appear in the gray box. The remaining components, along with panic attacks, compose Panic Disorder. Arrows indicate causal effects. Lines ending in an open circle indicate moderation of the causal effects on which they terminate. Solid arrows indicate positive effects. Dashed arrows indicate negative effects. The dashed “self-loop” that initiates and terminates in arousal represents the regulating effects of homeostatic feedback. Paired positive and negative arrows (i.e., the arrows terminating on Arousal Schema) indicate an effect that can be either positive or negative. Notably, the arrows terminating on Arousal Schema initiate from the panic attack as a whole rather than any individual components of the attack, signifying that this effect is dependent on the aggregate behavior of these variables. Conversely, lines initiating from arousal schema and avoidance terminate on individual components or the relationships among them, signifying the role of these components as parameters in the equations that define panic attacks.

This causal diagram encodes how the symptoms of Panic Disorder directly interact. Causal diagrams such as this are commonly and fruitfully used in the Panic Disorder literature (Clark (1986, p. 463); Ehlers and Margraf (1989, p. 4); Fava and Morton (2009, p. 630); Lader (1991, p. 157); Pauli et al. (1991, p. 138); Pilecki, Arentoft, and McKay (2011, p. 385); Rapee (1995, p. 430); Sandin, Sánchez-Arribas, Chorot, and Valiente (2015, p. 38)). In this section, we aim to take a step beyond causal diagrams by proposing a mathematical model: a model that specifies not just which components are related, but also the functional form of those relationships. That is, we aim to faithfully represent the theoretical framework we have abstracted from the literature as a set of Ordinary Differential Equations that define how each variable changes over time as a function of itself and the other variables in the model (e.g., Strogatz, 2015). Together, these equations and their substantive interpretation constitute the theory of Panic Disorder that will be the focus of the remainder of this article.

In this section, we will gradually build up the model equations one variable at a time, thereby allowing us to discuss how we incorporated each building block of Panic Disorder into the model as well as our substantive interpretation of the

equations. This description is intended to be as accessible and self-contained as possible, though familiarity with Ordinary Differential Equations and dynamical systems will likely facilitate understanding (for good introductions, see Feldman, 2012; Sayama, 2015; Strogatz, 2015). We will first introduce and analyze reduced equations that define only the vicious cycle between arousal and perceived threat. To help illustrate the behavior produced by these equations, we will also introduce the concepts of alternative stable states and tipping points from the dynamical systems and ecology literatures (Scheffer, 2009). We will then expand on these initial equations and incorporate two model components that play the critical role of constraining the positive feedback loop between arousal and perceived threat: escape behavior and homeostatic feedback. Finally, we will define and explain the equations that incorporate the ability to learn from the experience of panic attacks, either reinforcing or disconfirming the belief that arousal is dangerous, and, thereby, altering the vulnerability to panic attacks.

### 10.3.1 The Vicious Cycle of Panic Attacks

The central feature of the model is the positive feedback loop between arousal and perceived threat, denoted  $A \rightleftarrows T$ . It comprises four components: arousal (A), perceived threat (T), the effect of perceived threat on arousal ( $T \rightarrow A$ ), and the effect of arousal on perceived threat ( $A \rightarrow T$ ). We review each in detail below.

#### 10.3.1.1 Arousal (A) & the Effect of Perceived Threat on Arousal ( $T \rightarrow A$ )

The first building block modeled in these equations is Arousal, denoted A, which is taken to represent all arousal-related bodily activity. We will assume arousal to be continuous, with higher values indicating more arousal. A customary way of expressing the behavior of A when modeling with differential equations is to characterize the derivative of A with respect to time. This derivative, denoted  $\frac{dA}{dt}$ , represents the rate of change in arousal: that is, how arousal will change from its current state as time progresses. If  $\frac{dA}{dt}$  is positive, arousal will increase. If  $\frac{dA}{dt}$  is negative, arousal will decrease.

We will define  $\frac{dA}{dt}$  as a function of the current level of arousal and the current level of perceived threat (T). Our objective in doing so is to represent the influence of perceived threat on arousal. We use a *rate parameter*, denoted  $\alpha$ , to define the intrinsic rate at which arousal can change, and a *slope parameter*, denoted  $\nu$ , to define the strength of the effect of perceived threat on arousal. Together, this yields the equation:

$$\frac{dA}{dt} = \alpha(\nu T - A) \tag{10.1}$$

The product of  $\nu$  and T can be thought of as the level of arousal that would correspond to the current level of perceived threat. If  $\nu T$  is greater than A (i.e., perceived threat suggests arousal should be higher than the current level of arousal), then the rate of change will be positive and arousal will increase. If  $\nu T$  is less than A, the rate of change will be negative and arousal will decrease.

This equation defines the effect of perceived threat on arousal ( $T \rightarrow A$ ), which we will take to represent the body's fight-or-flight response to perceived danger (Cannon, 1916). Importantly, this simple equation is not intended to represent a fully-developed quantitative theory of the body's fight-or-flight system. Rather, like each of the equations we will present, it is a minimal model: a simplified representation that qualitatively characterizes the phenomenon of interest. Because we assume small or distant threats will elicit low levels of arousal and severe or proximal threats will elicit high levels of arousal, we defined the  $T \rightarrow A$  effect to be linear and increasing, with the slope given by the parameter  $\nu$ . The assumption of linearity is made for the sake of simplicity but is not essential to the model (a non-linear monotonic curve would produce similar qualitative behavior).

### 10.3.1.2 Perceived Threat (T) & the Effect of Arousal on Perceived Threat ( $A \rightarrow T$ )

Perceived threat incorporates the perceived severity, proximity, and probability of a perceived threat. In this model, we will focus only on perceived threat arising from arousal. Like arousal, we will assume perceived threat to be continuous. Thus, it may entail extreme predictions of impending catastrophe, such as those emphasized by Clark (1986), but may also include appraisals of arousal as a low severity, distal, or low probability threat.

We will define the rate of change of perceived threat  $\frac{dT}{dt}$  as a function of the current level of perceived threat (T) and the current level of arousal (A). Like the equation for arousal, we include a rate parameter,  $\gamma$ , which specifies the intrinsic rate at which perceived threat can change. In contrast to the linear effect of perceived threat on arousal, we define the effect of arousal on perceived threat ( $A \rightarrow T$ ) as being sigmoidal (s-shaped; see next paragraph for further detail). Two parameters determine the shape of this sigmoidal effect:  $\lambda$  and  $\mu$ . Together, this yields the equation:

$$\frac{dT}{dt} = \gamma \left( \frac{A^\mu}{A^\mu + \lambda^\mu} - T \right) \quad (10.2)$$

Similar to Equation 10.1, the s-shaped  $\frac{A^\mu}{A^\mu + \lambda^\mu}$  effect can be thought of as the level of perceived threat that is elicited by the current level of arousal. If  $\frac{A^\mu}{A^\mu + \lambda^\mu}$  is greater than T, then the rate of change will be positive and T will increase. If  $\frac{A^\mu}{A^\mu + \lambda^\mu}$  is less than T, the rate of change will be negative and T will decrease. Thus, perceived threat will move toward the level dictated by the current level of arousal, with the precise rate of change determined by the magnitude of the discrepancy between  $\frac{A^\mu}{A^\mu + \lambda^\mu}$  and T as well as the intrinsic rate parameter  $\lambda$ .

We will assume the  $A \rightarrow T$  effect to incorporate both interoceptive awareness of one's arousal-related bodily sensations and the interpretation of those sensations. That is, detection and interpretation of arousal are the processes by which arousal triggers a perception of threat, a signal detection process akin to a smoke alarm (cf., Barlow & Craske, 1988). We chose an s-shaped function to represent this process because, for such an "alarm system" to work effectively, it is imperative that low level fluctuations in arousal arising from ordinary activities have

negligible effect on perceived threat. However, beyond a given threshold of normal variation, arousal should begin to elicit perceived threat, with the strength of that effect growing with increasing arousal until tapering as it approaches a maximum level of perceived threat.

The s-shaped  $\frac{A^\mu}{A^\mu + \lambda^\mu}$  function allows us to capture these aspects of the A→T effect. The parameter  $\lambda$  determines the sigmoid curve's half-saturation point and the parameter  $\mu$  determines its steepness. Together, these parameters determine the threshold at which A begins to affect T and the strength of that effect. In doing so, these parameters play a critical role in the model, as cognitive behavioral theories posit that the strength of the A→T effect is central to the development of panic attacks (see Section 10.2). In this model, we will use variation in the  $\lambda$  parameter to model variation in the strength of the A→T effect and, thus, variation in sensitivity of this “alarm system”.

### 10.3.2 Moderating the Strength of the Vicious Cycle of Panic Attacks

#### 10.3.2.1 Arousal Schema (S)

Psychological theories posit that “alarm system” sensitivity is dependent on one’s arousal schema (i.e., one’s beliefs about and learned associations with autonomic arousal). Schemata are cognitive structures: “organized elements of past reactions and experience that form a relatively cohesive and persistent body of knowledge capable of guiding subsequent perception and appraisals” (Segal, 1988, p. 147). Arousal schema includes beliefs that arousal-related bodily sensations are dangerous, as reflected in the concept of anxiety sensitivity, as well as beliefs about the likelihood of panic attacks and the perceived ability to cope with arousal or its consequences (i.e., panic self-efficacy; T, 1985; Casey, Oei, & Newcombe, 2004). Arousal schema thus guides the perception and appraisal of arousal, moderating the relationship between arousal and perceived threat.

#### 10.3.2.2 Context (C)

In this model, we will also consider one’s current context to moderate the sensitivity of the A→T “alarm system.” Individuals with Panic Disorder are more likely to experience panic attacks in some situations than others, particularly those where the perceived negative consequences of a panic attack are especially heightened (e.g., a crowded theater; Klein & Klein, 1989). Importantly, for individuals with Panic Disorder, the perception of threat is from arousal in the situational context, not the context itself. To account for this moderating effect of situational context, we added a binary context (C) variable to the model, representing the presence or absence of any context that predisposes an individual to panic. The value of C is chosen probabilistically and remains fixed for a specified period of time (see Supplementary Materials G.1 for further details).

### 10.3.2.3 The Moderating Effects of Arousal Schema and Context

Arousal schema and context have their moderating effect on the  $A \rightarrow T$  relation through their effect on the parameter  $\lambda$  (see Equation 10.2):

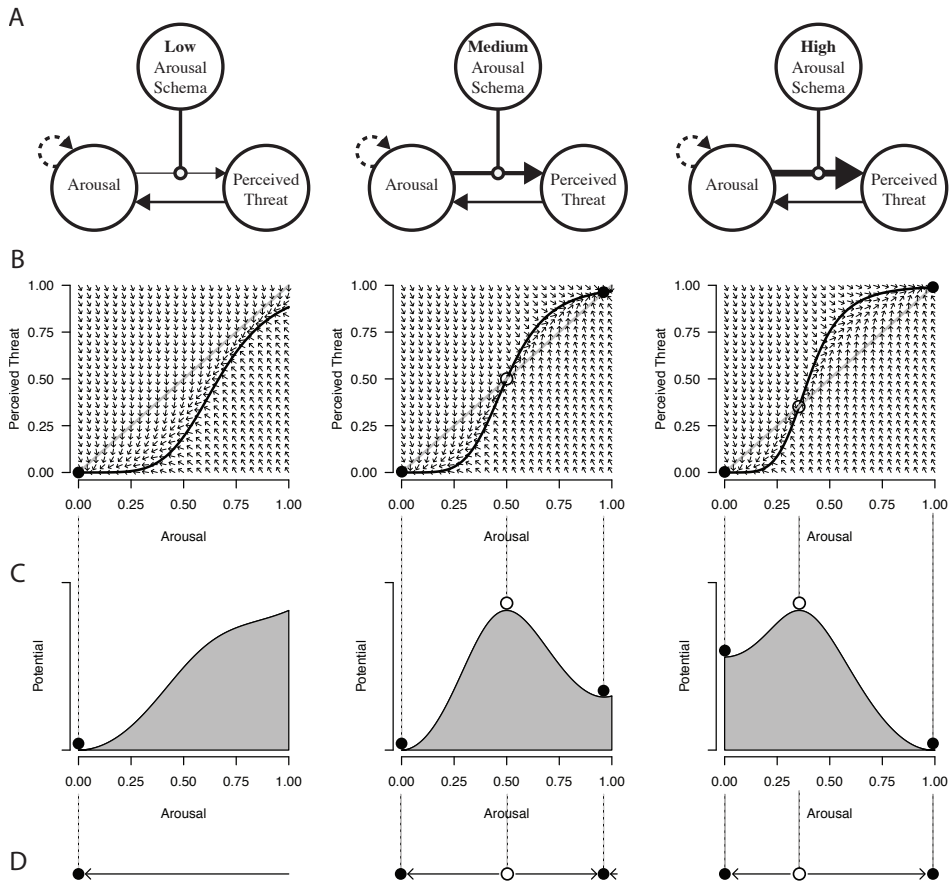
$$\lambda(S,C) = 1 - \frac{S}{S + \xi} - \pi C \quad (10.3)$$

As previously noted,  $\lambda$  is one of two parameters that determines the strength of the effect of arousal on perceived threat. To signify the dependence of this parameter on arousal schema ( $S$ ) and situational context ( $C$ ), we will denote the parameter as  $\lambda(S,C)$ . The parameters  $\xi$  and  $\pi$  constrain the strength of the moderating effects of  $S$  and  $C$ , respectively. The dependence of  $\lambda(S,C)$  on arousal schema means that when arousal schema is high, then  $\lambda(S,C)$  is low. When  $\lambda(S,C)$  is low, the  $A \rightarrow T$  effect is strong. In other words, when arousal is believed to be dangerous, it elicits greater perceived threat. Similarly, if a panic-predisposing situation is present ( $C=1$ ), then  $\lambda(S,C)$  is lowered and the  $A \rightarrow T$  effect is strengthened. The dependence of  $\lambda(S,C)$  on arousal schema and context could be represented mathematically in any number of ways and the precise equation used here should not be overinterpreted. The importance of Equation 10.3 is simply that it allows arousal schema and context to moderate the strength of the  $A \rightarrow T$  effect and, thus, the strength of the  $A \rightleftharpoons T$  feedback loop.

### 10.3.2.4 Illustrating the Moderating Effect of Arousal Schema

In Figure 10.2, we have depicted the  $A \rightleftharpoons T$  feedback loop under different levels of arousal schema to illustrate the effect of arousal schema on the behavior of the  $A \rightleftharpoons T$  feedback loop. As seen in Panel B, when arousal schema is low ( $S = 0.25$ ), the steepness of the s-shaped  $A \rightarrow T$  effect is low. As a result, arousal has relatively little effect on perceived threat and the  $A \rightleftharpoons T$  feedback loop is weak. As arousal schema increases to being moderate ( $S = 0.50$ ) or high ( $S = 0.75$ ), the s-shaped  $A \rightarrow T$  effect becomes steeper and the  $A \rightleftharpoons T$  feedback loop is strengthened. In Panels B-D of Figure 10.2, we illustrate how increasing arousal schema affects the behavior of this simple dynamical system by using three complementary diagrams: vector fields, stability landscapes, and phase lines. Each diagram depicts where the system will go next based on where it is now.

When arousal schema is low, the system will always move toward a state with no arousal and no perceived threat. All paths traced through the vector field will lead to the same point: no arousal and no perceived threat  $(0, 0)$ . In the stability landscape, any starting level of arousal will lead the system to move “downhill” into the basin where  $A = 0$ . In the phase line, any initial level of arousal will fall in the left-facing arrow pointing toward  $A = 0$ . The low-arousal schema system thus has a single *fixed point* or *stable state* (denoted by the filled circle). The point is a *fixed* because its rate of change is 0 (as indicated by the slope of 0 in the stability landscape). It is stable because neighboring system states move toward this point.



**Figure 10.2:** Examining the Positive Feedback Loop Between Arousal and Perceived Threat. The  $A \rightleftharpoons T$  feedback loop is presented under three conditions: low, medium, and high arousal schema ( $S = 0.25, 0.50,$  and  $0.75,$  respectively). Three diagrams appear for each condition: a vector field (Panel B), a stability landscape (Panel C), and a phase line (Panel D). Each diagram describes how the system evolves over time and can be used to determine where the system will go next based on where it is now. In the vector fields, each vector (arrow) indicates the direction the system will move from that point in the state space (the space defined by the values of A and T). Panel B also depicts the effect of perceived threat on arousal ( $T \rightarrow A$ , light grey) and of arousal on perceived threat ( $A \rightarrow T$ , black), illustrating how these effects affect the behavior of the  $A \rightleftharpoons T$  system. The stability landscapes provides a less precise but perhaps more intuitive depiction (for further detail, see Meyer (2016); Scheffer (2009, pp. 98-101)). The landscapes depict the behavior of just one of the system's state variables: arousal. The lateral position of an imaginary ball on the landscape represents the system's current state of arousal. The topography of the landscape describes the rate of change in arousal, with steeper slopes signifying larger rate of change. Movement along the landscape occurs under negative gradient flow: a ball placed on a slope will roll downhill (Meyer, 2016). The phase lines (Panel D) similarly depict the behavior of arousal, with the arrows indicating the direction arousal will move from that point on the phase line.

As arousal schema increases, the  $A \rightarrow T$  effect becomes steeper, the  $A \rightleftharpoons T$  interaction is strengthened, and the behavior of the system is altered. A bifurcation occurs, and an alternative stable state emerges. In the stability landscapes, this



bifurcation is marked by presence of two new features: a second stable fixed point (marked by the filled circle in the newly formed basin) and an unstable fixed point (marked by the open circle at the peak of the landscape). Like the stable fixed point, the unstable fixed point has a rate of change of 0. However, rather than attracting neighboring states, the unstable fixed point repels them. Any deviation from the unstable fixed point will lead the ball to roll downhill to one of the two stable fixed points: a state of no arousal or a state of extreme arousal. Similarly, in the vector field and phase lines, arrows point away from the unstable fixed point (open circle) and toward the stable fixed points (filled circles).

The formation of an alternative stable state strongly affects the behavior of this simple system. At medium arousal schema ( $S = 0.50$ ), if the system is pushed into a state of modestly elevated arousal, it will still return to the original stable state of no arousal. However, if arousal is pushed beyond the unstable fixed point, the system will enter a state of *runaway positive feedback* between arousal and perceived threat and will fall into the newly formed alternative stable state of high arousal. In other words, the unstable fixed point is a tipping point. Once crossed, the system quickly flips into a state of high arousal and high perceived threat. At high arousal schema ( $S = 0.75$ ), this tipping point shifts toward lower values of arousal and perceived threat and the system becomes increasingly vulnerable to runaway positive feedback. Even modest elevations in arousal flip the system into an alternative stable state of extreme arousal and perceived threat. Accordingly, the diagrams illustrate how changes in arousal schema (and, thus, the equations defining the  $A \rightleftharpoons T$  feedback loop) create the conditions for a vicious cycle that culminates in a panic attack.

### 10.3.3 Regulating Vicious Cycle of Panic Attacks

#### 10.3.3.1 The Regulating Effect of Escape Behavior on Perceived Threat ( $T \rightleftharpoons E$ )

Psychological theories not only posit a positive feedback loop between arousal and perceived threat, but also a negative feedback loop between perceived threat and escape behavior ( $T \rightleftharpoons E$ ), in which increases in perceived threat lead to *more* escape behavior and increases in escape behavior leads to less perceived threat. Thus, escape is a behavioral intervention for regulating one's inner state: if perceived threat becomes sufficiently elevated, the person will engage in escape behavior until the perception of threat has been removed.

To model this process, we will incorporate escape behavior (E) into the model. E includes any behavior aimed at coping with a perceived threat that is currently present. We will treat E as continuous, comprising a range of behaviors from subtle safety behavior (e.g., sitting down in case of fainting) to outright flight from the situation (e.g., running out of a crowded theater). Its rate of change ( $\frac{dE}{dt}$ ) is a function of escape behavior and perceived threat (T). The parameter  $\epsilon$  gives the intrinsic rate at which escape behavior can change. The parameters  $\rho$  and  $\sigma$  determine the threshold and rate at which perceived threat leads to escape

behavior. Together, this yields the equation

$$\frac{dE}{dt} = \varepsilon \left( \frac{T^\sigma}{T^\sigma + \rho^\sigma} \right). \quad (10.4)$$

The addition of an effect of escape behavior on perceived threat also requires that we update the equation that defines perceived threat, adding a negative effect of E that defines the rate of change in T, as well as a parameter  $\tau$  that regulates the strength of that effect. This yields an updated equation for T that replaces the prior equation (Equation 10.2) in our model

$$\frac{dT}{dt} = \gamma \left( \left( \frac{A^\mu}{A^\mu + \lambda^\mu} - T \right) - \tau E \right). \quad (10.5)$$

Together, Equations 10.4 and 10.5 allow escape behavior to act like a thermostat (Dretske, 1997): as perceived threat rises, escape behavior is engaged ( $T \rightarrow E$ ). As escape behavior increases, it lowers perceived threat ( $E \rightarrow T$ ) in an effort to prevent panic.

### 10.3.3.2 The Regulating Effect of Homeostatic Feedback on Arousal ( $A \rightleftharpoons H$ )

When panic attacks do occur, they typically peak within three to four minutes of onset (Cohen et al., 1985; R. R. Freedman et al., 1985) and subside within 5 - 20 minutes (R. R. Freedman et al., 1985; Goetz et al., 1993). The termination of panic attacks is surprisingly understudied (Radomsky, Rachman, Teachman, & Freeman, 1998), but presumably homeostatic processes counteract unsustainably elevated arousal and return it to baseline (cf., Ehlers & Margraf, 1989). To represent this effect, we added “homeostatic feedback” (H) to our model. The equation defining homeostatic feedback takes the same form as the equation defining escape behavior and is presented in Supplementary Materials G.1. The addition of homeostatic feedback and its effect on arousal requires that we update the equation that defines arousal, adding the negative effect of H and a parameter ( $\kappa$ ) that regulates the strength of that effect, yielding the updated equation for A:

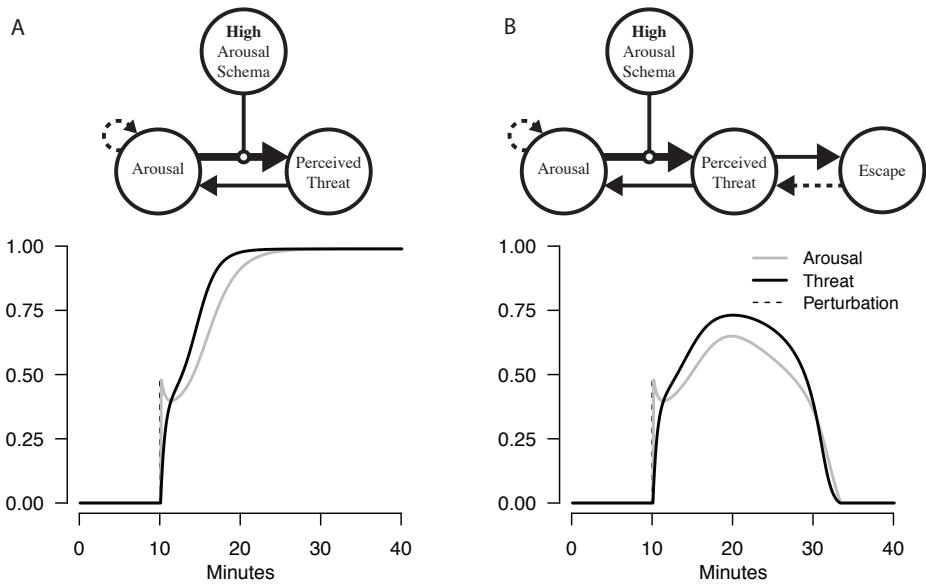
$$\frac{dA}{dt} = \alpha \left( (\nu T - A) - \kappa H \right) \quad (10.6)$$

Together, this equation and the equation defining H allow homeostatic feedback to behave like a thermostat (cf.  $E \rightleftharpoons T$ ). As arousal increases, homeostatic feedback is engaged ( $A \rightarrow H$ ) and, in turn, homeostatic feedback lowers arousal ( $H \rightarrow A$ ). Thus, H lowers arousal, fostering the termination of panic.

### 10.3.3.3 Illustrating the Regulating Effects of Escape Behavior ( $E \rightarrow T$ ) and Homeostatic Feedback ( $H \rightarrow A$ )

Together, escape behavior and homeostatic feedback introduce negative feedback loops that constrain the positive feedback between arousal and perceived threat. To illustrate this effect, we implemented the model with and without these negative feedback loops, as a series of difference equations in the software environment R (R Core Team, 2014).

We then simulated the effect of assigning a moderate level of arousal ( $A = 0.50$ ) to a system with high arousal schema ( $S = 0.75$ ). The results of this simulation appear in Figure 3. As seen in Panel A, in the absence of regulating feedback, the system flips into an alternative stable state of extreme arousal and perceived threat from which it does not recover. In contrast, with regulating feedback (see Panel B), runaway feedback is initiated and the system moves toward the alternative stable state, but the effects of escape behavior and homeostatic feedback pull the system back to a state of no arousal or perceived threat.



**Figure 10.3:** Examining the Effects of Regulatory Feedback. To illustrate the effects of homeostatic feedback (represented as a negative self-loop on arousal) and escape behavior, we simulated a moderate perturbation to arousal ( $A = 0.5$ ) at minute ten. In the left panel, this perturbation is applied to model that includes only the  $A \rightleftharpoons T$  feedback loop and  $S$ . In the right panel, this same perturbation is applied to the model, but with the incorporation of homeostatic feedback and escape behavior. In both simulations, the perturbation to arousal is sufficient to initiate the vicious cycle of panic attacks and move the system toward an alternative stable state. With the incorporation of homeostatic feedback and escape behavior, the system is ultimately pulled back from that alternative stable state and returns to its original state of low arousal and perceived threat.

### 10.3.4 Natural Variation in Arousal

Arousal is determined by factors beyond the effects of perceived threat and homeostatic feedback. To account for fluctuations in arousal arising from either physiological processes or the environment (e.g., running to catch a bus or drinking coffee), we incorporated a final component to the equation that defines the rate of change in arousal ( $dA/dt$ ): a noise function ( $N$ ) that induces stochastic variation in arousal (Hasselmann (1976); van Nes and Scheffer (2004, p. 257)). A complete description of how  $N$  is calculated appears in Supplementary Materials G.1. The

incorporation of N gives the updated and final equation for arousal

$$\frac{dA}{dt} = \alpha((\nu T - A) - \kappa H + N). \quad (10.7)$$

### 10.3.5 The Vicious (or Virtuous) Cycle of Panic Disorder

At this stage, our effort to model Panic Disorder hits an important obstacle: the variables that constitute Panic Disorder change on a significantly different time scale than those that constitute a panic attack. For example, beliefs about the danger of arousal-related bodily sensations (i.e., arousal schema) do not fluctuate on the same seconds-to-minutes time scale as arousal and perceived threat. They change over the course of days or weeks. These different time scales introduce a challenge regarding how to connect the rapidly changing components of a panic attack with the slower changing components of Panic Disorder.

In the ecology literature, researchers have addressed this problem by partitioning variables into fast changing components (e.g., spruce budworms whose population fluctuates on a timescale of months) and slow-moving components (e.g., the spruce-fir trees on which the budworms reside, whose foliage changes on a timescale of years; Ludwig, Jones, Holling, et al., 1978; Rinaldi & Scheffer, 2000). The behavior of the fast variables is then assessed with the slow variables held constant and the behavior of the slow variables is assessed with the fast variables held at their equilibrium. Inspired by this approach, we divide the components of the model into fast changing components (i.e., arousal, perceived threat, escape, and context) and slow changing components (i.e., arousal schema and avoidance), a division that corresponds to the components of a panic attack and Panic Disorder, respectively (see Figure 10.1). We consider the “fast” changing panic attack variables to change on a time scale of minutes and the “slow” Panic Disorder variables to change on a time scale of days (i.e., 1/1,440th the rate at which the panic attack variables change). We then treat the slow-moving elements as constants with respect to the fast-changing elements. That is, the slow-moving elements of Panic Disorder act as parameters in the equations determining the fast-changing elements, thereby “controlling” their behavior. Conversely, as we will show, panic attack variables are also capable of “revolting” against this control and affecting the slower Panic Disorder variables that constrain them (cf. Schulze et al., 1996, p. 32).

#### 10.3.5.1 The Effect of Panic Disorder Variables on Panic Attack Variables

The slow arousal schema (S) variable controls the behavior of the fast-changing elements of a panic attack by being part of the equation that shapes the A→T effect, specifically the  $\lambda(S,C)$  parameter (see Equation 10.3). As illustrated in Figure 10.2, arousal schema has a substantial impact on the vicious cycle of panic attacks through its effect on  $\lambda(S,C)$ , with high arousal schema creating a system that is vulnerable to runaway positive feedback between arousal and perceived threat.

Arousal schema also indirectly controls these variables through its effect on Avoidance (V), the final variable in our model. We will consider avoidance to represent the disposition to preemptively prevent elevated arousal (e.g., refraining from drinking coffee) or exposure to panic predisposing situational contexts (e.g., avoiding public transportation). We will assume that avoidance arises as a function of one's beliefs about the stimuli or situations that may elicit elevated arousal or exacerbate the consequences of a panic attack. That is, we will consider avoidance behavior to be a function of arousal schema. To represent this effect, we will define the rate of change in avoidance ( $\frac{dV}{dt}$ ) to be a function of the current level of avoidance (V) and of arousal schema (S). The parameter  $\eta$  gives the intrinsic rate at which avoidance can change. The parameters  $\varphi$  and  $\chi$  determine the threshold and rate at which arousal schema affects avoidance:

$$\frac{dV}{dt} = \eta \left( \frac{S^\chi}{S^\chi + \varphi^\chi} - V \right) \quad (10.8)$$

If the level of avoidance suggested by arousal schema ( $\frac{S^\chi}{S^\chi + \varphi^\chi}$ ) is greater than the current level of avoidance (V), then the rate of change will be positive and avoidance will increase. Thus, with the incorporation of Equation 10.8 into the model, increases in arousal above a given threshold will elicit avoidance behavior. Avoidance, in turn, acts as an additional control on the rapidly-changing panic attack variables. It does so by acting as a parameter in the equations that determine stochastic variation in arousal (N) and situational context (C; see Supplementary Materials G.1), thereby altering arousal and the effect of arousal on perceived threat, respectively. As avoidance increases, the magnitude of fluctuations in arousal is reduced and the probability of being in a panic-predisposing situation is diminished. Accordingly, like escape behavior and homeostatic feedback, avoidance has a regulating effect on the vicious cycle of panic attacks.

### 10.3.5.2 The Effect of Panic Attacks on Panic Disorder Variables

By affecting arousal schema, the fast-changing panic attack variables “revolt” against the control from slow-changing Panic Disorder variables. This effect reflects the ability to learn from experiences with elevated arousal and perceived threat, such as a panic attack. Based on our review of psychological theories, we determined that this learning effect should depend not on a single panic attack variable, but on the aggregate behavior of the panic attack variables over time (which we will indicate using the subscripts  $t - \Omega$  and  $t$ ). In other words, the learning that occurs from a panic attack depends on how the panic attack unfolds. To incorporate this effect of panic attacks on arousal schema, we defined the rate of change in arousal schema ( $\frac{dS}{dt}$ ) as a function of current arousal schema (S) and three core panic attack variables (A, T, and E). Two rate parameters represent the learning rate at which beliefs and associations regarding arousal are either acquired ( $\zeta_A$ ) or extinguished ( $\zeta_E$ ; see Equation 10.9 below).

We used two conditional statements to determine the appropriate calculation of  $\frac{dS}{dt}$ . First, the parameter  $\psi$  is used to determine if arousal and perceived threat are sufficiently present to allow for learning to occur. To make this determination,

we calculated the joint level of A and T as their geometric mean ( $\sqrt{AT}$ ), which we will refer to as Fear (F) given that jointly experienced arousal and perceived threat are core ingredients of fear (Lindquist & Barrett, 2008). We chose the geometric mean to aggregate A and T to ensure that both variables are present for learning to occur. If fear is insufficient ( $\max\{F_{t-\Omega}, \dots, F_t\} < \psi$ ) there is no opportunity for learning and, arousal schema will not change ( $\frac{dS}{dt} = 0$ ; see Equation 10.9).

If arousal and perceived threat are sufficiently elevated, there is opportunity for learning and the lesson taken from this opportunity depends on the second conditional statement: whether escape behavior was also present. The parameter  $\omega$  is used to determine if escape behavior is present. If so (i.e., if  $\max\{E_{t-\Omega}, \dots, E_t\} > \omega$ ), then arousal schema will move toward the maximum level of perceived threat during the specified time period (i.e.,  $\max\{F_{t-\Omega}, \dots, F_t\}$ ), at a rate determined by  $\zeta_A$ . If escape behavior is insufficient, then the arousal schema variable will move toward 0 at a rate determined by  $\zeta_E$ .

Thus,  $\frac{dS}{dt}$  is given by:

$$\frac{dS}{dt} = \begin{cases} 0, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} < \psi \\ \zeta_A(\max\{T_{t-\Omega}, \dots, T_t\} - S), & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} > \omega \\ -\zeta_E S, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} \leq \omega \end{cases} \quad (10.9)$$

Conceptually, this means that episodes of elevated arousal and perceived threat provide an opportunity to modify arousal schema (S). If individuals engage in escape behavior during these episodes, they will counterfactually infer that the anticipated consequences of arousal would have occurred if not for the escape behavior. Thus, they will learn that arousal is as dangerous as it was perceived to be during the attack. If the individual does not engage in escape behavior, they may learn that even without escape behavior, no catastrophe occurred, and, thus, that arousal is not dangerous.

## 10.4 Evaluating the Theory of Panic Disorder

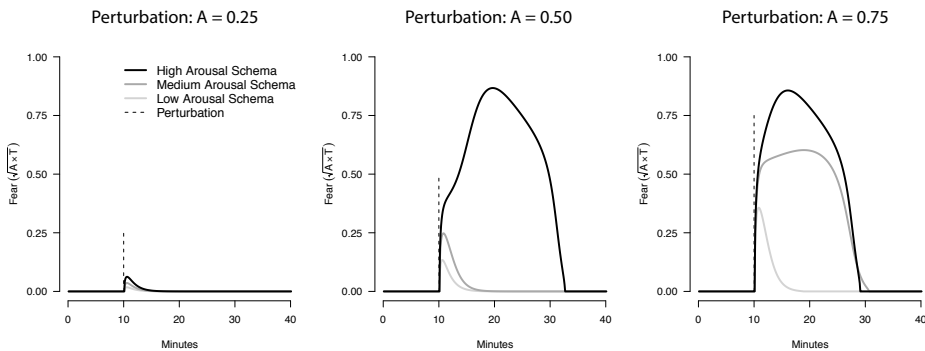
The Ordinary Differential Equations presented in Section 10.3 provide a mathematical model of panic attacks. We implemented this model in the software environment R (R Core Team, 2014) as a series of difference equations. We will refer to this implementation as a “computational model”, in contrast to the “mathematical model” defined by the differential equations. Whereas differential equations provide the instantaneous rate of change in continuous time, the difference equations used in our computational model work in discrete time, using the values of the state variables at a time step  $t$  to calculate the values of those same state variables at time  $t+1$  (for further description, see Supplementary Materials G.1). In this model, we treat each time step  $t$  as corresponding to one “minute”. We used Euler’s method (Atkinson, 2008) with time step  $t/1000$  to approximate the solutions of the system of differential equations.

The computational model of Panic Disorder provides us with a tool to simulate the behavior implied by the theory, thereby allowing us to evaluate what

it can and cannot explain (Epstein, 2008; Smaldino, 2017). In this section, we focus our evaluation of Panic Disorder theory on its ability to produce five robust features of panic attacks and Panic Disorder. First, individuals vary in the propensity to experience panic attacks. This is perhaps most clearly illustrated in the biological challenge literature, where the same perturbation to arousal-related bodily sensations elicits panic attacks in some individuals but not others (Clark, 1993; Liebowitz et al., 1984; Rapee, 1995; Woods, Charney, Goodman, & Heninger, 1988). Second, the core aspects of panic attack phenomenology: a rapid surge of arousal and perceived threat in the absence of a clear external provocation (Barlow & Craske, 1988). Third, recurrent panic attacks are often accompanied by avoidance behavior and persistent beliefs regarding the danger of panic attacks and the bodily sensations that accompany them (Buller, Maier, & Benkert, 1986; White, Brown, Somers, & Barlow, 2006). That is, these symptoms often cohere as a syndrome. Fourth, panic attacks can occur in the absence of Panic Disorder. Although only 3.7% of U.S. adults report a lifetime history of Panic Disorder, more than a quarter report having had at least one panic attack over the course of their lives (Kessler et al., 2006). Finally, cognitive behavioral therapy, a well-established treatment for Panic Disorder (Barlow, 1997), reduces symptoms of Panic Disorder. An adequate theory must be able to reproduce these fundamental features of panic attacks and Panic Disorder. Throughout our examination of these features in this section, we continue to draw on concepts from the ecology literature to deepen our understanding of the model's behavior, especially the concepts of alternative stable states, tipping points, and resilience.

#### 10.4.1 Feature 1: Individual Differences in Vulnerability to Panic Attacks

To examine individual differences in vulnerability to panic attacks, we simulated perturbations to the system and examined the system's response. These simulated perturbations are analogous to "biological challenges" in which researchers use standard procedures (e.g., CO<sub>2</sub> inhalation) to induce arousal-related bodily sensations (Clark, 1993; Liebowitz et al., 1984; Rapee, 1995; Woods et al., 1988). Here, we assigned specified levels of arousal at time step 10 (which we will refer to as "minute" 10) and evaluated the system's response to that perturbation. The results of this simulation for low, moderate, and high arousal schema ( $S=0.25, 0.50,$  and  $0.75,$  respectively) at low, moderate, and high perturbation strength ( $A=0.25, 0.50,$  and  $0.75,$  respectively) appear in Figure 10.4. The most noteworthy finding appears in the center panel depicting the effect of moderate perturbations to arousal. At low or moderate arousal schema, moderate perturbations cause only a brief and modest increase in arousal and perceived threat. However, at high arousal schema ( $S=0.75$ ), moderate perturbations lead to runaway feedback between arousal and perceived threat and the system quickly tips into an alternative stable state of extreme arousal and perceived threat. In other words, at high arousal schema, moderate perturbations to arousal suffice to send the system into a state of panic.



**Figure 10.4:** Individual Differences in Vulnerability to Panic Attacks. We simulated perturbations to arousal of varying strength (inducing arousal of 0.25, 0.50, and 0.75) in three conditions: low, medium, and high arousal schema ( $S = 0.25, 0.50,$  and  $0.75,$  respectively). To simplify this simulation, we removed natural variation in arousal and the effects of escape behavior. At low arousal schema (light grey lines), the system is highly resilient, recovering rapidly following low, moderate, or high perturbation and never flipping into an alternative stable state. In contrast, when arousal schema is moderate the time to recover is prolonged following mild perturbations ( $A = 0.25$ ) and perturbations of moderate or greater strength ( $A = 0.50$  and  $0.75$ ) are capable of pushing the system into an alternative stable state of a panic attack. In other words, when arousal schema is low, the system is resilient. When arousal schema is high, the system is vulnerable to panic attacks in reaction to even moderate perturbation in arousal.

Importantly, this simulated behavior follows directly from the behavior of the simple dynamical system created by the  $A \rightleftharpoons T$  feedback loop as depicted in Figure 10.2. At low arousal schema, the feedback between arousal and perceived threat is insufficient to create an alternative stable state, precluding a vicious cycle that culminates in panic. Hence the low arousal schema system's rapid return to baseline following even the most extreme perturbations. In contrast, when arousal schema is high, the feedback between arousal and perceived threat is sufficient to create an alternative stable state with a relatively low tipping point, rendering the system vulnerable to the vicious cycle of panic attacks following even moderate perturbations to arousal.

This simulation suggests that varying arousal schema allows the model to account for individual differences in the propensity to experience panic attacks, aligning well with the literature on anxiety sensitivity as a predictor of response to biological challenge (McNally, 2002). Importantly, however, individual differences in any other parameter in the  $A \rightleftharpoons T$  feedback loop can produce similar system behavior. For example, varying the effect of perceived threat on arousal ( $T \rightarrow A$ ) while holding the  $A \rightarrow T$  effect constant produces stability landscapes for arousal similar to those depicted in Figure 10.2 (see Supplementary Materials G.2). Regardless of which factor produces change in the system's behavior, if the strength of the  $A \rightleftharpoons T$  feedback loop is sufficient to produce an alternative stable state of high arousal and perceived threat, then the system is vulnerable to panic attacks.



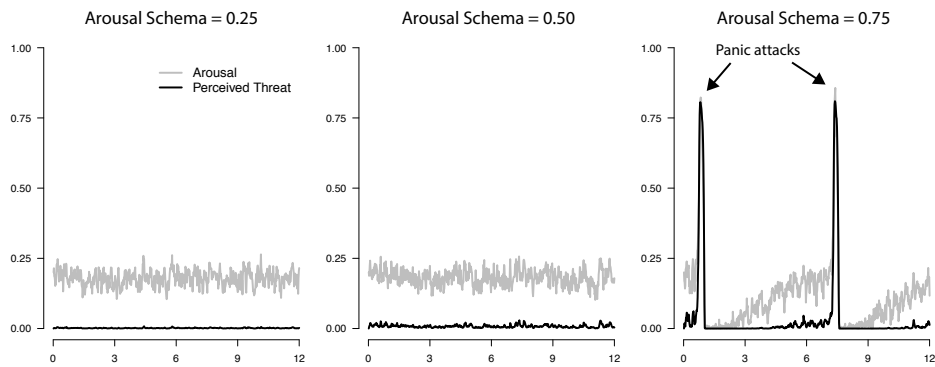
### 10.4.1.1 Resilience

The question of whether a system is vulnerable can be reframed as a question about the system's resilience: its ability to recover from disturbance. Researchers studying dynamical systems, such as ecosystems, have developed indices for quantifying a system's *resilience*. One such index is the amount of disturbance the system can tolerate before shifting into an alternative stable state (Holling, 1973; Scheffer et al., 2009). This index is referred to as *ecological resilience* because of the focus in ecology on understanding and predicting catastrophic shifts in the state of an ecosystem (e.g., from savanna to desert). For the  $A \rightleftharpoons T$  system, we can consider ecological resilience to represent the strength of perturbation (e.g., "biological challenge") needed to push the system into a state of panic. In other words, it corresponds to the location of the system's tipping point (see Figure 10.2). As depicted in Figure 10.4, when arousal schema is low, the system can recover from strong perturbations without shifting into an alternative state, thus exhibiting high ecological resilience. However, when arousal schema is high, the system flips into an alternative stable state following moderate perturbation, thus exhibiting low ecological resilience.

Resilience can also be defined as the speed with which a system returns to its stable state following perturbation. This index is sometimes referred to as *engineering resilience* because engineered systems are often designed to remain very close to a precise stable state (Holling, 1996). In the  $A \rightleftharpoons T$  model, engineering resilience represents the time for the system to return to its stable state following our simulated "biological challenge." As seen in Figure 10.4, even for relatively mild perturbations (e.g.,  $A = 0.30$ ), the time to recover is greater when arousal schema is high, illustrating that even small perturbations may be sufficient to assess individual differences in engineering resilience and, thus, individual differences in vulnerability to experience panic attacks.

### 10.4.2 Feature 2: The Phenomenology of Panic Attacks

The simple  $A \rightleftharpoons T$  feedback loop can produce surges of arousal and perceived threat in response to perturbation (see Figures 10.2 & 10.4). With the inclusion of homeostatic regulation and escape behavior, the model can constrain the "vicious cycle," circumventing or terminating the panic attack (see Figure 10.3). Here, we examined whether the model could also explain the spontaneous rise and fall of arousal and perceived threat in the absence of an external provocation, a critical feature of panic phenomenology. To do so we simulated 12 "hours" of model behavior with stochastic variation in arousal under three conditions: low, moderate, and high arousal schema ( $S = 0.25, 0.50, \text{ and } 0.75$ ). As seen in Figure 10.5, in the low and moderate arousal schema conditions, arousal varies around a relatively low baseline value. However, in the high arousal schema condition, the model produces sudden surges of arousal and perceived threat. This simulation suggests the model can explain the spontaneous and rapid onset of panic attacks in response to natural variation in arousal, a defining feature of the panic attacks observed in those with Panic Disorder.



**Figure 10.5:** The Spontaneous Surge of Arousal and Perceived Threat. To examine the features of panic attacks we simulated twelve “hours” of model behavior under conditions of low, medium, and high arousal schema ( $S = 0.25, 0.50,$  and  $0.75,$  respectively). When arousal schema is moderate or low, arousal schema varies around a consistent low mean of arousal. When arousal schema is high, sudden and rapid spikes in arousal and perceived threat.

Interestingly, the model also suggests why people may perceive panic as being qualitatively distinct from a state of anxiety. In the presence of strong positive feedback and, thus, the development of an alternative stable state, the shift between states does not occur gradually. It occurs as a sudden transition: a catastrophic shift into a state of panic (see Figure 10.5). This categorical shift in the state of the system is especially interesting because it arises from dimensional changes in the components of the positive feedback loop, illustrating a general feature of complex systems: in some parameter settings (e.g., low arousal schema), they can behave continuously, whereas in others (e.g., high arousal schema) they can only occupy a limited number of discrete states (Borsboom et al., 2016). Thus, the model may explain why panic attacks are experienced as discontinuous with the normal state of being, even though their key variables involved are all continuous.

### 10.4.3 Feature 3: Coherence of the Panic Disorder Syndrome

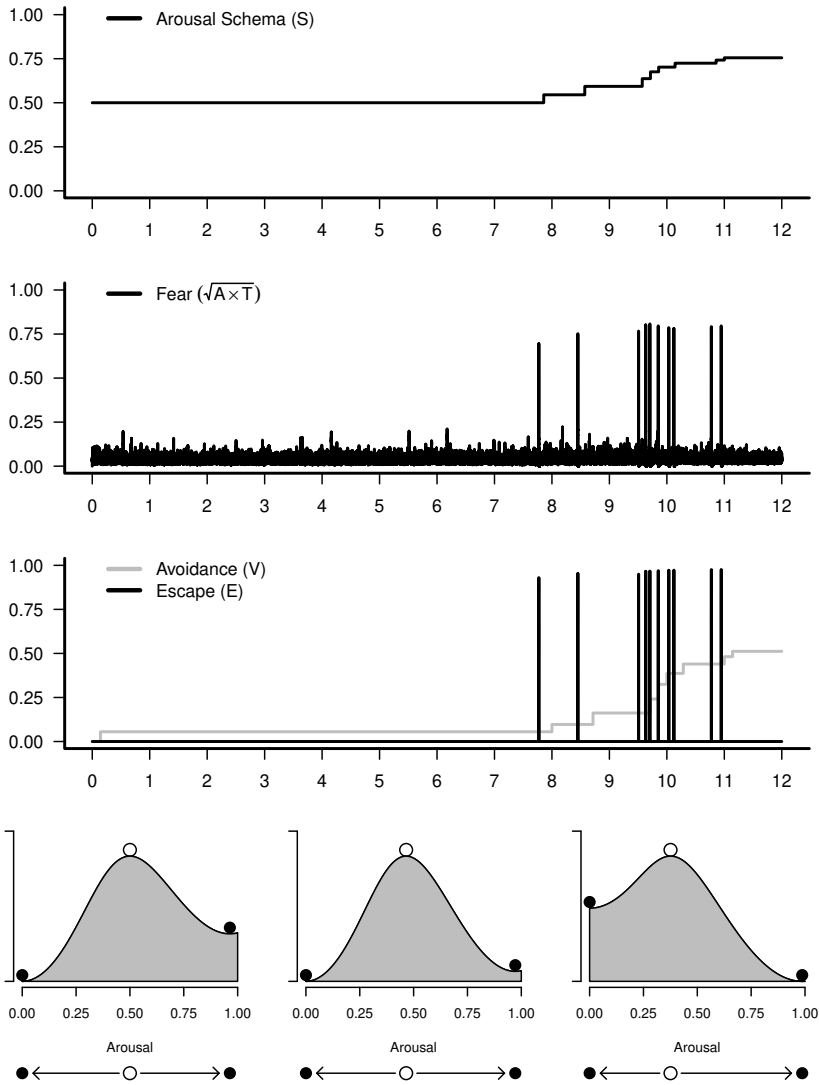
We next evaluated whether the model was able to reproduce the onset and coherence of the Panic Disorder syndrome. To do so, we specified an initial value of the arousal schema variable ( $S = 0.50$ ) and simulated twelve “weeks” of model behavior. We chose an initial value for arousal schema for which there is an alternative stable state in the system (and, thus, some vulnerability to panic attacks; see Figure 10.2 and 10.4) but for which there is not yet recurrent panic attacks or significant avoidance behavior. The results of this simulation appear in Figure 10.6. For the first seven weeks of the simulation, arousal varies but does not cross the tipping point that leads to panic. However, during the seventh week, the tipping point is crossed and runaway feedback leads to a spike in arousal and perceived threat (reported as “fear”). Critically, this spike of fear is accompanied by escape behavior (see bottom panel). The aggregate experience of fear and escape be-

havior teaches the belief that arousal is dangerous, as reflected by the increased arousal schema. This heightened arousal schema, in turn, renders the system more vulnerable to panic attacks (see inset B) and the system falls into a slow vicious cycle: panic attacks lead to higher arousal schema and higher arousal schema increases vulnerability to panic attacks. The increase in arousal schema also leads to corresponding increases in avoidance behavior (see bottom panel) which, in turn, leads to reductions in both arousal variation and the likelihood of being in a panic-predisposing situation. By week 12, arousal schema is elevated, the system is highly vulnerable to panic attacks (see inset C), and panic attacks are recurrent, constrained only by avoidance. This simulated behavior is highly reminiscent of Panic Disorder, particularly if we regard the arousal schema variable as a proxy for the “persistent concern” symptom of Panic Disorder. The model thus provides an account for how causal relations among the symptoms of Panic Disorder can mutually reinforce one another, leading them to emerge and cohere as a syndrome.

#### 10.4.3.1 Avoidance Promotes Engineering Resilience

The efforts to regulate arousal through avoidance behavior can be understood as an effort to promote the system’s engineering resilience (i.e., the time it takes to recover to the stable state; Holling, 1996). It is a “near equilibrium” strategy aimed at keeping the system as close to the desired stable state as possible (i.e., a state of low arousal and low perceived threat) by reducing variation in arousal. Escape behavior similarly pushes the system toward a desirable state as soon as a deviation from that state has been detected.

In ecology, action taken to promote engineering resilience often achieves greater near-equilibrium stability (Holling, 1986). However, these actions can also come at the long term cost of diminishing the system’s ecological resilience (i.e., its ability to withstand perturbation without shifting into an alternative stable state; Holling, 1996). For example, fish hatcheries produce more stable and predictable fish populations. However, the larger and more stable fish populations encourage commercial fishing, depleting natural stocks. Consequently, the system becomes dependent on a limited number of hatcheries and, thus, more vulnerable to catastrophic shifts in the fish population if those hatcheries fail. Analogously, in the Panic Disorder model, escape and avoidance behavior successfully reduce variability in arousal and, thus, reduce the frequency of panic attacks. However, they also create and sustain a system that is highly vulnerable to panic attacks when those strategies fail. They promote engineering resilience at the cost of ecological resilience.

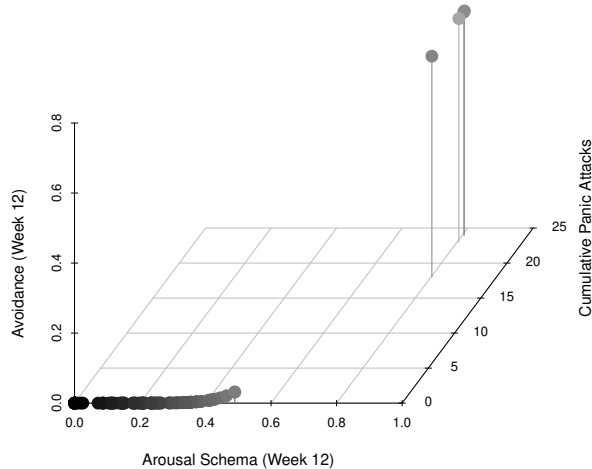


**Figure 10.6:** The Onset and Coherence of the Panic Disorder Syndrome. This figure presents a twelve “week” simulation of model behavior beginning with moderately elevated arousal schema ( $S = 0.50$ ). The bottom row depicts the stability landscape and phase line for arousal at three points during the simulation: (A) Week 6 , (B) Week 8, and (C) Week 10. Rows 1-3 plot the values of arousal schema, fear (calculated as the geometric mean of arousal and perceived threat), and escape and avoidance behavior across the full simulation, respectively.

### 10.4.4 Feature 4: Non-clinical Panic Attacks

We next evaluated whether the model adequately captures the phenomenon of non-clinical panic attacks: infrequent attacks in the absence of Panic Disorder.

We did so by again simulating 12-weeks of the model's behavior, but now repeated the simulation 100 times across a range of initial arousal schema values. Arousal schema values were drawn from a normal distribution with a relatively low mean (.25) because most individuals endorse minimal beliefs regarding the danger of arousal-related bodily sensations (Deacon, Abramowitz, Woods, & Tolin, 2003). We specified a standard deviation for the distribution ( $SD = 0.15$ ) so that the proportion of systems with an alternative stable state (and, thus, the potential to experience a panic attack) would be approximately 28%, roughly corresponding to the lifetime prevalence of panic attacks (Kessler et al., 2006). We then evaluated whether some simulations exhibited panic attacks without developing other symptoms of Panic Disorder (e.g., elevated arousal schema and avoidance).



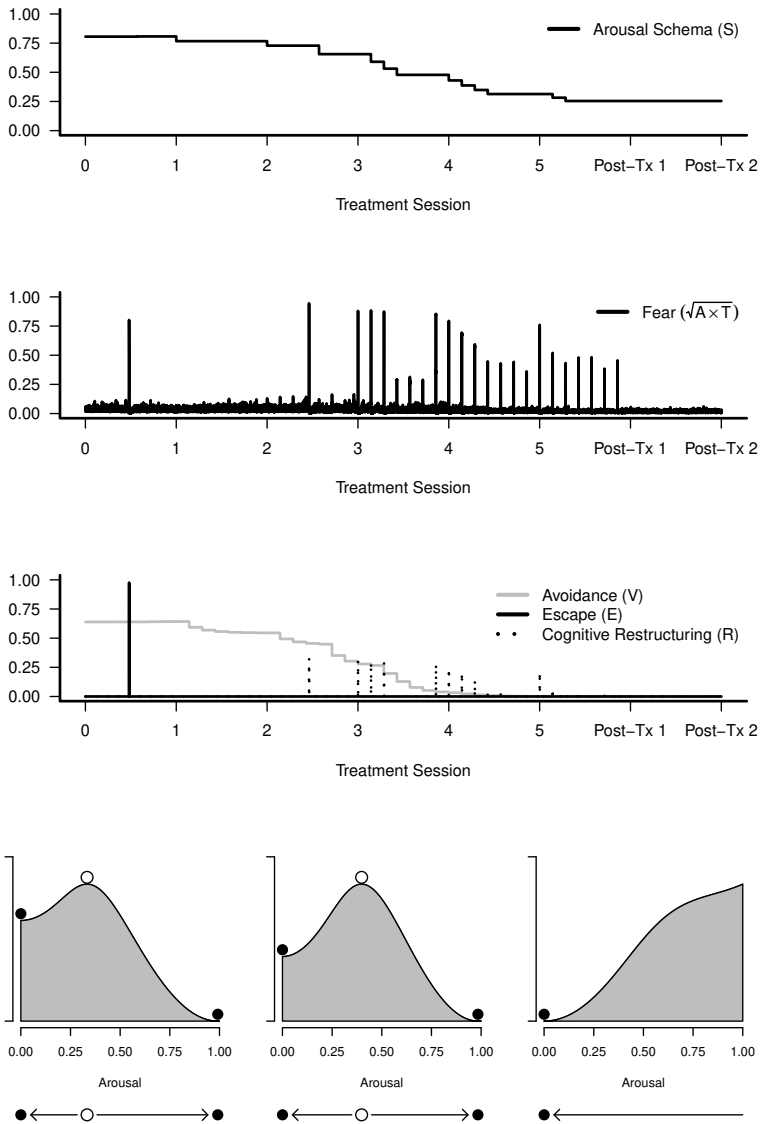
**Figure 10.7:** The Phenomenon of Non-Clinical Panic. This figure presents one-hundred iterations of the twelve “week” simulation of model behavior where each iteration began with an arousal schema value randomly drawn from a normal distribution ( $M = 0.25$ ,  $SD = 0.15$ ). The circles representing each iteration are shaded by their initial level of arousal schema. The figure depicts the final arousal schema value (x-axis), the final avoidance value (y-axis), and the cumulative number of panic attacks that occurred during the twelve-week simulation (z-axis). In no iteration was a panic attack observed in the absence of eventual elevations in arousal schema, avoidance, and recurrence of panic attacks. That is, these simulations did not find evidence that the model can produce the phenomenon of non-clinical panic attacks.

A panic attack was observed in 4 of the 100 simulations (see Figure 10.7). In each case, there were recurrent panic attacks (range = 18 – 24), elevated arousal schema (range = 0.80 – 0.81) and elevated avoidance (0.63 – 0.64) by the end of the simulation. Thus, the experience of a panic attack during the twelve-week simulations always led the system to fall into a highly similar state of Panic Disorder. Indeed, because panic attacks are always accompanied by escape behavior

in the current model (and, thus, always lead to increased arousal schema), we can anticipate that given sufficient time, any system capable of experiencing a panic attack under the conditions simulated here will eventually develop the full Panic Disorder syndrome. Accordingly, the model fails to explain the phenomenon of non-clinical panic attacks.

#### 10.4.5 Feature 5: The Efficacy of Cognitive Behavioral Therapy for Panic Disorder

Cognitive behavioral therapy is a well-established, effective treatment for Panic Disorder (Barlow, 1997). An adequate model of Panic Disorder should be able to provide some account for its efficacy. We represented treatments for Panic Disorder as external interventions on specific components of the model (cf. Lader & Mathews, 1970, p. 159). By simulating those targeted effects, the impact of the intervention on the system can be evaluated. Here, we simulated a 5-week cognitive behavioral therapy intervention (Otto et al., 2010) that included four active treatment components: psychoeducation, cognitive restructuring, interoceptive exposure, and in vivo exposure (see Supplementary Materials G.3 for a complete description of the intervention on the model). The results of this treatment simulation appear in Figure 10.8. As depicted in the first row, the simulation begins with a system that is highly vulnerable to panic attacks. In treatment sessions 1 and 2, psychoeducation and cognitive restructuring produce modest reductions in arousal schema. Cognitive restructuring (R) also introduces a new negative feedback loop for perceived threat ( $T \rightleftharpoons R$ ). Beginning in session 2, daily interoceptive exposure exercises perturb arousal with increasing perturbation strength ( $A = 0.30, 0.50, 0.70$  and  $0.70$  for the weeks beginning with Sessions 2-5, respectively). Critically, escape behavior is prohibited during these exposures. As a result, each experience with substantially elevated arousal and perceived threat leads to a reduction in the arousal schema variable. By session 3, the treatment has bolstered the system's resilience, but panic attacks remain possible (i.e., the tipping point has shifted but not been eliminated; see second row). Between sessions 3 and 5, interoceptive exposure exercises repeatedly perturb the system and lead to further reductions in arousal schema. In Session 5, interoceptive exposures are combined with in vivo exposure, in which the model "enters" panic-predisposing contexts when arousal is perturbed, producing stronger perturbations to the system. By the end of treatment, the system no longer has an alternative stable state of extreme arousal and perceived threat (see third row). In other words, the system is resilient: no longer vulnerable to the runaway feedback that gives rise to panic.



**Figure 10.8:** The Effect of Cognitive Behavioral Therapy. This figure depicts a seven week simulation of a “cognitive behavioral therapy” intervention beginning in Session 1 (Week 1). The bottom row depicts the stability landscape and phase line for arousal one week before treatment (Session 0), at third treatment session (Session 3), and one week after the final treatment session (Post-Tx 1) in insets A, B, and C, respectively. Rows 1-3 plot arousal schema, fear (calculated as the geometric mean of arousal and perceived threat), and escape and avoidance behavior over the course of the full treatment simulation, respectively.

The extent to which this treatment simulation is a valid method for drawing precise inferences about cognitive behavioral therapy for Panic Disorder is de-

pendent on the adequacy of the model. As already noted, the model has limitations and, thus, inferences drawn from the model behavior should be made with caution. Nonetheless, our simulated intervention does suggest that the model can plausibly account for the efficacy of cognitive behavior therapy.

#### 10.4.5.1 Cognitive Behavioral Therapy Promotes Ecological Resilience

Viewing Panic Disorder treatment through the lens of dynamical systems provides some insight into the mechanisms by which cognitive behavioral therapy may have its effect. In contrast to the *engineering* resilience promoting strategy of avoidance, cognitive behavioral therapy can be viewed as a strategy that promotes *ecological* resilience. Patients are encouraged to refrain from escape and avoidance behavior in favor of strategies that modify arousal schema and, thus, diminish the feedback between arousal and perceived threat. In other words, this therapy does not aim to control the position of the system on the stability landscape, but rather to change the topography of the landscape itself. By repeatedly pushing the system away from its desired stable state through exposure exercises, the intervention produces a system with greater ecological resilience: one that does not permit catastrophic shifts into panic. In future work, it may be beneficial to investigate other psychological and pharmacological interventions for Panic Disorder with the aim of identifying those that solely promote the system's engineering resilience, thereby leaving it vulnerable to relapse after the intervention is removed, and those that promote the system's ecological resilience, thereby producing gains more likely to be sustained after treatment.

## 10.5 A Theory Driven Research Agenda for Panic Disorder

The explanatory power of the model makes it tempting to conclude that theorists have largely solved Panic Disorder. Indeed, there is reason to think that such a mentality is present in the field. The number of papers published on Panic Disorder has declined in recent years (Asmundson & Asmundson, 2018) and there have been few noteworthy advances to psychological theories of Panic Disorder since the bevy of "fear of fear" theories proposed in the 1980s (for valuable exceptions, see e.g., Bouton et al., 2001; Casey et al., 2004). Yet, the model presented here reveals substantial limitations to our understanding of Panic Disorder. By requiring us to make the theory explicit, this exercise in modeling reveals that there is little theoretical or empirical guidance for defining the functional form of the relationships among Panic Disorder symptoms. Moreover, as illustrated by the model's inability to produce the phenomenon of non-clinical panic attacks, there are limitations to what the theory can explain. We consider these limitations to be among the model's most important contributions to our understanding of Panic Disorder as they identify what remains unknown and, thereby, identify key directions for future research. In this section, we describe three areas of future research suggested by the model: (a) continued development of Panic Disorder



theory, (b) empirical research that grounds the model parameters in data and (c) empirical research that tests predictions made by the model.

### 10.5.1 Developing Panic Disorder Theory

In the previous section, we identified one robust feature of Panic Disorder that is not explained by the model: non-clinical panic attacks. There are further limitations to what the model can explain. For example, in the current model, sufficiently elevated arousal always leads to a panic attack in vulnerable systems. Yet, some patients with Panic Disorder can engage in activities that significantly increase arousal without experiencing panic attacks (e.g., exercise; C. B. Taylor et al., 1987). Thus, as with any model, the current model is incomplete and further theory development is needed to improve the coherence between the theory and the observed facts about Panic Disorder.

#### 10.5.1.1 Cumulative Theory Development with Computational Modeling

We encourage researchers interested in further developing Panic Disorder theory to freely adapt the model developed here. A significant advantage of mathematical and computational modeling is that it facilitates cumulative theory development. We implemented the model in freely available software, the code implementing the model appears in our Supplementary Materials, and the code is described in further detail in Supplementary Materials G.1. We aimed to make the model sufficiently explicit that it can be divorced from any specific theorist. In other words, the model makes predictions, not the current authors, thereby allowing other researchers to independently evaluate, refute, revise, and extend the model.

We especially hope that the explication of the model in this formalized manner will allow for a “patchy reductionism” approach to developing Panic Disorder theory, with researchers able to improve upon those components in which they are expert (Kendler, 2005; Schaffner et al., 1994). Several aspects of the model are scientific fields unto themselves, suggesting that contributions from researchers across disciplines will be necessary to fully develop the theory. For example, simple learning mechanism used in this model is sufficient to reproduce the onset and treatment of Panic Disorder, but may be improved by rooting the model’s learning mechanisms in existing learning theories, especially those with well-developed computational models that could potentially be incorporated here (e.g., Rescorla, Wagner, et al., 1972). More broadly, there is considerable literature on biological factors relevant to Panic Disorder (for a review of these factors from a causal systems perspective, see Hassan, 2008) and biological theories of Panic Disorder (Klein, 1993; Ley, 1985) that could potentially be used to further develop the biological components of the current model. This work may allow the model to address phenomena that presumably operate through principally biological mechanisms (e.g., why inhaling carbon dioxide induces arousal-related bodily sensations) and could potentially allow the model to distinguish between different types of panic attack (e.g., respiratory vs. nonrespiratory; Roberson-

Nay & Kendler, 2011; Kendler et al., 2011). Similarly, we modeled homeostatic feedback here as a simple negative feedback loop. Although this feedback loop allows the model to behave in a manner that resembles a panic attack, it can likely be further informed by the literature on homeostasis and allostasis (Ramsay & Woods, 2014).

### 10.5.1.2 Theory Evaluation

The call for further theory development raises the question of how to determine whether an alternative to the current model constitutes an improvement in Panic Disorder theory. A complete review of the criteria for theory evaluation is beyond the scope of this chapter (for extended discussions, see Haig, 2014; Kuhn, 1977; Lakatos, 1970; McMullin, 1982), but we would broadly recommend that researchers consider at least four criteria when evaluating competing Panic Disorder theories: (a) explication, (b) accuracy, (c) consilience, and (d) simplicity.

We believe that theories made *explicit* through mathematics or computational modeling are, other things being equal, preferable to narrative accounts of a theory where the vagaries of language can mask unknowns and contradictions within the theory. Although not a necessary criterion for a well-developed theory, the ability to analyze or simulate the behavior produced by a theory so greatly facilitates other aspects of theory evaluation that explication should be considered when evaluating a theory's merits. In this context, it is worth noting that we are not the first to explicate a mathematical model of the vicious cycle of panic attacks. Fukano and Gunji (2012) proposed such a model in which "physical symptoms" and "fear" (cf. arousal and perceived threat) are defined by coupled logistic equations with Allee effects (an approach commonly used to model population growth Stephens, Sutherland, & Freckleton, 1999). Although not a model of the full Panic Disorder syndrome, we believe this model makes a highly valuable contribution to the theory of panic attacks by serving as the first effort to mathematically model them. By proposing a mathematical model, the authors produced a theory that can be readily reproduced and evaluated according to our remaining suggested criteria for theory evaluation.

*Accuracy* refers to agreement between the consequences deduced from a theory and the observed facts about that phenomena (Kuhn (1977, p. 103); cf. explanatory coherence; Thagard (1989)). The Fukano and Gunji (2012) model of panic attacks can explain the defining feature of a panic attack: the sudden rise of fear and arousal-related physical symptoms (see Supplementary Materials G.4). However, it also ascribes intrinsic behavior to arousal-related physical symptoms that is inconsistent with observation of those with panic attacks. For example, the model implies that above a critical threshold, physical symptoms will always rise to an elevated stable equilibrium, even in the absence of any feedback relationship with fear (see Supplementary Materials G.4). This prediction does not accord with the simple observation that physiological arousal tends to diminish toward a low stable equilibrium and is also inconsistent with the observation that intervening on the feedback relationship between fear and physical symptoms (e.g., in cognitive behavioral therapy) is sufficient to eliminate the surge of

physical symptoms characteristic of a panic attack.

The Fukano and Gunji model also exhibits limited *consilience*: the amount a theory explains (Thagard, 1978). As a model of panic attacks alone, it does not explain features of the broader Panic Disorder syndrome, such as the coherence of Panic Disorder symptoms or their treatment in cognitive behavioral therapy. Together, the limited accuracy and consilience of the Fukano and Gunji model suggest that an alternative model is needed.

The model we have proposed avoids several of the inaccuracies of the Fukano and Gunji model and it accounts for more features of Panic Disorder, thus exhibiting both greater accuracy and greater consilience. Yet, as previously noted, the current model has its own limitations in both of these criteria, suggesting further development is still needed. In Supplementary Materials G.5, we illustrate how continued development of this theory could occur. Specifically, we propose the incorporation of an Escape Schema ( $S_E$ ) variable that moderates the effect of perceived threat on escape behavior, just as the arousal schema variable in our model moderates the effect of arousal on perceived threat. To our knowledge, no theory explicitly posits that individual differences in beliefs about escape behavior plays a prominent role in the etiology of Panic Disorder (hence its omission in the model that was the focus of this chapter). Nonetheless, such a model component plausibly fits with theory and research on Panic Disorder. For example, those with higher panic self-efficacy (i.e., confidence in one's ability to cope with a panic attack) may not believe escape behavior to be uniquely helpful and, thus, may be less likely to engage in such behavior. In other words, panic self-efficacy may be a component of an escape schema variable rather than arousal schema as we originally proposed. With this revision, individuals who believe arousal is dangerous (high  $S$ ) but who do not believe that escape is their most effective means of coping with perceived threat (low  $S_E$ ) can experience surges of arousal and perceived threat without engaging in escape behavior and, thus, do not develop the belief that arousal is dangerous (indeed, the revised model suggests they will learn it is not dangerous, see Supplementary Materials G.5). By allowing the model to explain non-clinical panic attacks, the addition of escape schema improves the theory's consilience.

The final criterion, *simplicity*, aims to limit the incorporation of ad hoc hypotheses that explain no more than the facts that they were introduced to explain (Thagard, 1978). The merit of incorporating an Escape Schema variable thus hinges not only on whether it improves the theory's consilience, but also on whether this addition explains features of Panic Disorder beyond the specific feature it was designed to address (i.e., non-clinical panic attacks). For example, adding the escape schema variable may allow it to better account for the effects of cognitive behavioral therapy (Gallagher et al., 2013) or explain the observation that individuals frequently refrain from escaping the situation during panic attacks despite the urge to do so. When a modification expands the model's explanatory breadth beyond the specific facts it is originally intended to address, the theory exhibits dynamic consilience (Thagard, 1978) (Thagard (1978, p. 84); cf. progressive problem shifts, Lakatos (1970)). That is, the amount it can explain grows over time. In future research, it will be important to evaluate escape

schema and other potential revisions to the model to determine if they expand the explanatory breadth of the theory while retaining its simplicity.

### 10.5.2 Grounding the Model in Data

A critical step to facilitate theory development will be to ground the model's parameters in data. We found minimal guidance from theory or research when choosing the model parameters that shape the relationships among model components. Accordingly, we chose parameter values based on their plausibility and their ability to produce the behavior of interest rather than being derived from observation of the model components. It will be critical to rectify this shortcoming by rigorously examining the individual components of Panic Disorder and the relationships among them. Improving our understanding of the specific form of the relationships between arousal and perceived threat alone would constitute a significant advance in our understanding of Panic Disorder, as these relationships determine much of the behavior of the system.

This endeavor, as well as the practical value of the model, will be strengthened by rooting the state variables in substantively meaningful scales. For example, researchers should ground the arousal variable in measurable indices of autonomic arousal. We suspect there is unlikely to be an absolute level of arousal that triggers perceived threat, but rather a relative value that reflects a person-specific calibration (e.g., a deviation in heart rate from one's within-person mean or a discrepancy from an expected level of heart rate given the current context). Clarifying the precise nature and scale of arousal and other state variables will advance our understanding of how arousal-related bodily sensations elicit perceived threat and will allow theorists to ground the model in data. The model can then be more precisely applied to understanding individual patients by gathering behavioral data and deriving individual parameter values that characterize the individual patient's system, a process referred to as *computational phenotyping* (Patzelt, Hartley, & Gershman, 2018).

### 10.5.3 Testing Theory Predictions

The ultimate aim of further research and theory development efforts should be to produce a model sufficiently well-developed that it can generate specific numerical point predictions (e.g., a precise value of peak arousal predicted by a given perturbation under a given level of arousal schema) as such a theory would lend itself to rigorous empirical scrutiny (Meehl, 1990b). The current model cannot make such predictions. Nonetheless, the model does make broad predictions that can be tested as a means of evaluating core features of the model. For example, the model predicts that the time to recover from a biological challenge (i.e., its engineering resilience) should be a marker of the system's vulnerability and, thus, should prospectively predict panic attacks. In practice, measures of engineering and ecological resilience tend to be highly correlated, suggesting either will be appropriate tests of vulnerability (Van Nes & Scheffer, 2007). However, because assessments of ecological resilience would require perturbing the system to the

point of eliciting a panic attack, assessments of engineering resilience, which allow for more moderate perturbations, may be more appropriate. In addition to its utility as a test of Panic Disorder theory, such a tool could prove valuable for detecting risk for Panic Disorder, thereby enhancing the clinical utility of “biological challenge” paradigms (Forsyth & Karekla, 2002).

The model also posits that escape behavior plays a critical role in both the etiology and treatment of Panic Disorder. There is reason to question whether such a prominent role for escape behavior is appropriate. For example, Rachman and colleagues found that agoraphobic patients instructed to engage in escape behavior responded as well to treatment as those instructed refrain from such behavior, an observation at odds with the model’s predictions (Rachman, Craske, Tallman, & Solyom, 1986). There are limitations to this study as a test of the model presented here. For example, subjects in the ‘escape’ condition frequently refrained from escaping and the researchers used a more restrictive definition of escape behavior than we have argued for here (for further discussion, see Salkovskis, 1991). Moreover, other studies have reported greater improvements in those that refrain from escape behavior, as predicted by the model (Salkovskis et al., 1999). Nonetheless, these inconsistent results regarding the effects of escape behavior suggest caution is warranted and serves as a reminder that the model’s ability to produce the behavior of interest does not ensure that the model is accurate. If future research fails to support the crucial role of escape behavior, it will be necessary to alter the model and propose alternative mechanisms by which substantial elevations in arousal and perceived threat can produce both the development and remission of Panic Disorder.

## **10.6 Understanding and Investigating Mental Disorders as Complex Systems**

The model proposed in this chapter explains key features of Panic Disorder by modeling the causal relations among Panic Disorder symptoms. The model thus illustrates how a mental disorder can arise from mutually reinforcing relationships among symptoms. Indeed, one advantage of mathematical and computational models is that they can illustrate and even reveal behavior implied by a theory that is difficult to discern or anticipate from a narrative account alone. By modeling Panic Disorder as a non-linear dynamical system, the model presented here illustrates several noteworthy features of dynamical systems that have implications for how we understand and investigate mental disorders. These features are worthy of some attention.

### **10.6.1 Harmful Stable States and the Definition of Disorder**

The model illustrates how we might define mental disorders from a dynamical systems perspective, a definition based on two dimensions: (a) the current state of the system and (b) the presence vs. absence of a harmful alternative stable state as determined, in part, by the structure of relationships among symptoms

(cf. Borsboom, 2017, Figure 7). Panic Disorder can be characterized as a system residing in a harmful alternative stable state. Symptoms are active and will remain so barring influence from outside the Panic Disorder system (e.g., treatment or change in beliefs prompted by elicited social support). Conversely, a system that has a single desirable stable state in which it currently resides can be characterized as a resilient and healthy system. If this resilient system is pushed into a state of elevated symptom activation, it is perhaps no longer in a state of mental health, but is also not in the harmful equilibrium state that defines psychopathology. Symptoms will remit independent of external intervention. Such a system is thus better characterized as being in a state of transient symptom elevation and likely warrants a different level of clinical attention than that of Panic Disorder. Interestingly, current diagnostic criteria require the repeated occurrence of panic symptoms for at least one month, implicitly drawing a boundary between transient symptoms (symptom elevations in the absence of an alternative stable state) and persistent symptoms that do not appear likely to remit naturally (symptom elevations in a state of equilibrium). Finally, these two dimensions provide a fourth classification: a system that has a harmful alternative stable state but is currently in a stable state of low symptom activation. Although such a system appears to be in a state of health, it is nonetheless vulnerable to Panic Disorder and could benefit from preventative intervention.

Notably, defining mental disorder in this way suggests potential tools from the complexity literature that could be applied in psychiatric research. In this chapter, we used engineering and ecological resilience to quantify the vulnerability of the panic attack system, an approach that could potentially be used to identify those vulnerable to panic attacks before the onset of Panic Disorder symptoms. Relatedly, researchers investigating other dynamical systems, such as ecosystems, have used early warning signals, such as increased autocorrelation among the system's state variables, to detect systems approaching a tipping point that would push it into an alternative stable (Scheffer et al., 2009). Such early warning signals may portend a shift into a depressive episode (Bos & De Jonge, 2014; van de Leemput et al., 2014b; Wichers et al., 2016). Although preliminary, this work suggests that the toolbox used to investigate, anticipate, and control non-linear dynamical systems may be applied to identifying and treating individuals with vulnerable systems, even absent a current disorder.

### 10.6.2 Mental Disorders as Emergent Phenomena

The characterization of Panic Disorder as a system in a harmful stable state illustrates an interesting feature of the model. The most important phenomena explained by the model are not explicit components in the model, but rather emerge from interactions among components. "Panic attack" is not a component in the model nor can it be reduced to any one component. It is a systemic state emerging from the interaction between arousal and perceived threat. Likewise, Panic Disorder is not a component of the system, but a state of the system arising from interactions among the elements that define its presence: panic attacks, cognitions regarding panic-related arousal, and avoidance behavior.

### **10.6.3 Explanatory Pluralism and Equifinality**

The model allows for factors across levels of analysis to operate together in the development and treatment of Panic Disorder, thereby accommodating both psychological and biological theories of the disorder. In our evaluation of the model, we focused on psychological components (e.g., arousal schema). However, biological components (e.g., physiological regulation of arousal or the effect of perceived threat on arousal) play similarly critical roles in shaping system behavior and individual differences in these components can just as readily produce a system vulnerable to the development of Panic Disorder (see Supplementary Materials G.2). Systems with feedback across multiple levels of analysis such as this generally preclude the attribution of causal priority to one level of description. Instead, such models exploit the interactions between these levels to achieve the rich set of adaptive and maladaptive behaviors characteristic of complex systems (Barabási, 2012). Notably, because of the mutual causal relationships among its components, the system arrives at a similar state regardless of the specific factors that initiate its movement toward that state (cf. Bystritsky, Nierenberg, Feusner, & Rabinovich, 2012, p. 430). Whether instigated by impaired physiological regulation or the receipt of news about one's cardiovascular health, the feedback relationships in the model will ultimately lead the system to a state of Panic Disorder. These properties of explanatory pluralism and equifinality are common in psychopathology and in the broader domain of complex systems (Nolen-Hoeksema & Watkins, 2011; Von Bertalanffy, 1972).

### **10.6.4 Dysfunction from Function**

Dysfunction in an individual component of the model can contribute to the development of Panic Disorder. However, the model's ability to produce Panic Disorder does not require component-level dysfunction. Indeed, there is no essentially dysfunctional component in the model as we have presented it here, nor are any of the relations among components dysfunctional. On the contrary, they are utterly necessary. Any species that does not react to perceived threat with increased arousal, does not engage in escape behavior in the face of perceived threat, or that is incapable of learning that a given stimulus may be dangerous will surely and rapidly go extinct. Although the belief that autonomic arousal is dangerous may be inaccurate, the model comes by this falsehood honestly, through the appropriate functioning of its ability to learn. Thus, although this is a model of a pathological phenomenon, none of its specific ingredients need be pathological.

### **10.6.5 Developing Theories for Other Mental Disorders**

The overarching aim of this chapter was to develop a theory that posits precisely how a mental disorder may operate as a complex system. In doing so, we hewed closely to the abductive theory of scientific method (ATOM; Haig, 2005). We identified a robust phenomenon (panic attacks) and used the overarching network theory of mental disorders (Borsboom, 2017) as well as the prior litera-

ture on Panic Disorder to generate an initial theory of Panic Disorder as a complex system. We developed this theory using the analogy of mental disorders as ecosystems and the production of a computational model. Finally, we appraised the theory by examining the model's accuracy, consilience, and simplicity (an approach similar to, though distinct from, the explanatory coherence approach embraced by ATOM; Haig, 2005).

We anticipate that this framework can be successfully applied to other mental disorders. There are many resources from which one could generate or adopt initial theories about a mental disorder's causal structure. We surveyed the literature on Panic Disorder, integrating prior psychological theories. We suspect this approach will be effective for some mental disorders but that others may not have as strong a theoretical base upon which to draw as was available in the context of Panic Disorder. In such cases, alternative approaches may be fruitful, including the use of exploratory network analysis to investigate the structure of relationships among symptoms (Cramer et al., 2016), reviews of the empirical research on relationships among specific features of a disorder (van den Hout, 2014; Wittenborn, Rahmandad, Rick, & Hosseinichimeh, 2016), and the assessment of clinician or patient beliefs about the causal relationships among symptoms (Frewen, Schmittmann, Bringmann, & Borsboom, 2013; N. S. Kim & Ahn, 2002)

The computational model proposed here can also serve as a resource for developing theories. We suspect that many emotional disorders arise from the same dynamical system motifs we used to construct this model, including positive feedback loops (e.g., between rumination and depressed mood; Hosseinichimeh, Wittenborn, Rick, Jalali, & Rahmandad, 2018), negative feedback loops (e.g., between social anxiety and avoidance of social situations), and an interaction between fast-changing variables and the slow-changing variables that guide their behavior. Indeed, the similarity of system motifs across emotional disorders is implicitly posited and exploited to greater effect in recent transdiagnostic approaches to treatment, which argue that there is a similar structure in the causal relationships among cognitions, emotions, and behavior across emotional disorders (Barlow et al., 2011, 2017). In addition, many psychological theories posit that functional short-term behavior heightens long-term vulnerability to psychopathology, suggesting that our efforts to model the relationships among fast and slow-changing variables may be applicable to other mental disorders.

## 10.7 Conclusions

This chapter offers a formalized theory of Panic Disorder as a complex system. In essence, we aimed to condense a century of psychological theory and research on recurrent panic attacks into a minimal set of mathematical equations that define the causal relations among symptoms of panic attacks and Panic Disorder. The ensuing minimal model explains how causal relationships among symptoms can produce the key features of Panic Disorder, including the core phenomenological features of panic attacks, individual differences in the propensity to experience panic attacks, the development of the full Panic Disorder syndrome, and the ef-



fects of cognitive behavioral therapy on Panic Disorder.

The model also reveals significant gaps in our understanding of Panic Disorder, identifying three critical avenues for future research. First, further theory development is needed to increase the accuracy and consilience of the model. Second, the model parameters require empirical grounding. Third, the model makes predictions that should be tested. We believe future research should proceed in the order presented here, prioritizing continued development of Panic Disorder theory and careful observation of Panic Disorder symptoms, thereby facilitating the development of a model that is capable of making more precise and testable predictions.

We hope that the approach taken to theory development here can guide similar efforts for other mental disorders and, ultimately, move the field toward an ongoing exchange between theory and empirical research in which well-developed formalized theories summarize what is known about a mental disorder and guide the ongoing investigation into how the disorder operates as a complex system.



# FROM DATA MODELS TO FORMAL THEORIES

---

## Abstract

Over the past decade there has been a surge of empirical research investigating mental disorders as complex systems. In this chapter, we investigate how to best make use of this growing body of empirical research and move the field toward its fundamental aims of explaining, predicting, and controlling psychopathology. We first review the contemporary philosophy of science literature on scientific theories and argue that fully achieving the aims of explanation, prediction, and control requires that we construct formal theories of mental disorders: theories expressed in the language of mathematics or a computational programming language. We then investigate three routes by which one can use empirical findings (i.e., data models) to construct formal theories: (a) using data models themselves as formal theories, (b) using data models to infer formal theories, and (c) comparing empirical data models to theory-implied data models in order to evaluate and refine an existing formal theory. We argue that the third approach is the most promising path forward and conclude by expanding on this approach, proposing a framework for theory construction that details how to best use empirical research to generate, develop, and test formal theories of mental disorders.

## 11.1 Introduction

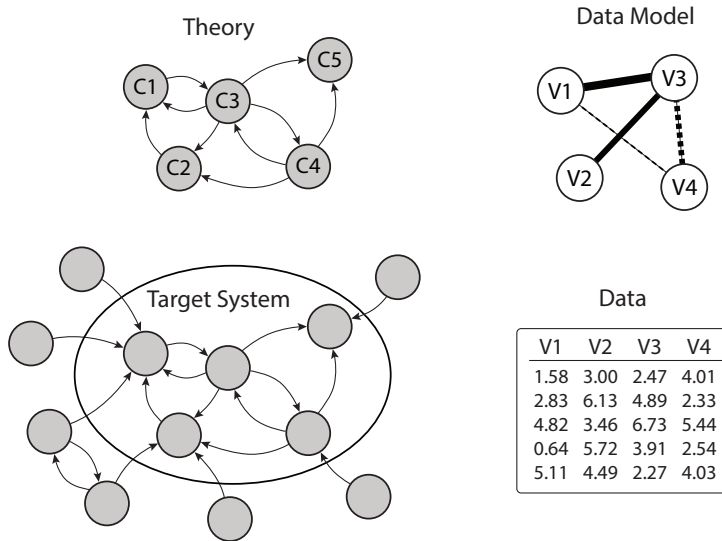
Mental disorders are complex phenomena: highly heterogeneous and massively multifactorial (e.g., Kendler, 2019). Confronted with this complex etiological and ontological picture, researchers have increasingly called for approaches to psychiatric research that embrace this complexity (Gardner & Kleinman, 2019). The “network approach” to psychopathology addresses these calls, conceptualizing mental disorders as complex systems of interacting symptoms (e.g., Borsboom & Cramer, 2013; Schmittmann et al., 2013; Borsboom, 2017). From this perspective, symptoms are not caused by an underlying disorder, rather the symptoms themselves and the causal relations among them constitute the disorder.

In recent years, empirical research within the network approach literature has rapidly grown (for reviews see e.g., Robinaugh, Hoekstra, & Borsboom, 2019; Contreras et al., 2019). Most of this work employs statistical models that allow researchers to study the multivariate dependencies among symptoms, thereby providing rich information about the relationships among those symptoms. However, this quickly expanding empirical literature has raised a critical question: how can we best make use of this growing number of empirical findings to advance the fundamental aims of psychiatric science? This problem is not unique to the network approach. Psychiatry has produced countless empirical findings, yet genuine progress in our efforts to explain, predict, and control mental disorders has remained stubbornly out of reach.

In this chapter, we will argue that empirical research can best advance these aims by supporting the development of scientific theories. We will begin in Section 11.2 by discussing the nature of scientific theories and how they achieve the explanation, prediction and control sought by psychiatric science. We will argue that to fully achieve these aims, psychiatry requires theories formalized as mathematical or computational models. In Section 11.3, we will explore how models estimated from data can best be used to develop formal theories. We examine three possible routes from data model to formal theory: first, treating data models themselves as formal theories; second, drawing inferences from data models to generate a formal theory; and third, using data models to develop formal theories with an abductive approach. We will argue that the third approach is the most promising path forward. In Section 11.4, we will expand on this approach and propose a framework for theory construction, detailing how best to use empirical research to advance the generation, development, and testing of scientific theories of mental disorders.

## 11.2 The Nature and Importance of Formal Theories

In this section we will examine the nature of scientific theories and how they support explanation, prediction, and control. We will begin by introducing four key concepts that we will use throughout the remainder of the chapter: theory, target system, data, and data models. We will illustrate each of these concepts using the example of panic disorder.



**Figure 11.1:** The figure illustrates the concepts target system, theory, data and data model. The target system is the system consisting of interacting components that gives rise to phenomena. Phenomena are robust features of the world captured by data models. Theories represent the structure of the target system, proposing a set of components  $C$  and the relations among them and positing that they give rise to the phenomena. Data for variables  $V$  are obtained by probing the target system.

### 11.2.1 Theories and Target Systems

*Theories* seek to explain phenomena: stable, recurrent, and general features of the world (Bogen & Woodward, 1988; Haig, 2008, 2014) such as the melting point of lead, the orbit of planets, and the tendency for some individuals to experience recurrent panic attacks. Well developed theories can predict these phenomena and show how they can be controlled. Although the precise nature of theories remains a subject of ongoing debate among philosophers of science, the past half century has seen a growing consensus that theories are best understood as *models*.<sup>1</sup> Specifically, theories are models that aim to represent *target systems*: the particular parts of the real world that give rise to the phenomena of interest. We use the word “system” here because we assume that the part of the real world giving rise to any psychiatric phenomena can be partitioned into components and the relations among them. We use the term “target”, because it is this system that a theory aspires to represent (cf. Elliott-Graves, 2014).

In psychiatry, the most common phenomena to be explained are symptoms

<sup>1</sup>The precise relationship between theories and models is muddled by inconsistent and often conflicting use of these terms across time, disciplines, and scientists (for a brief history of models and their relation to theory, see Bailer-Jones, 2009). In this chapter, we will adopt the perspective that theories are models (Suárez & Pero, 2019). However, the core arguments presented in this chapter do not require this precise conceptualization of theories and would similarly hold for pragmatic accounts that regard models as an intermediary between theory and the real world (e.g., Bailer-Jones, 2009; Cartwright, 1983).

and syndromes. For example, researchers seek to explain the tendency for some individuals to experience panic attacks and the tendency for recurrent panic attacks to be accompanied by persistent worry about those attacks and avoidance of situations in which they may occur (Spitzer, Md, & Williams, 1980). The target system in psychiatric research comprises the components of the real world that give rise to these symptoms and syndromes, and may include genetic, neurobiological, physiological, emotional, cognitive, behavioral or social components. Psychiatric theories aim to represent these target systems, positing a specific set of components and relationships among them that give rise to the phenomena of interest. For example, researchers have generated numerous theories of panic disorder, specifying a set of components that they believe interact to give rise to panic attacks and panic disorder. Among these, perhaps the most influential is Clark's cognitive model of panic attacks, which posits that "if [stimuli] are perceived as a threat, a state of mild apprehension results. This state is accompanied by a wide range of body sensations. If these anxiety-produced sensations are interpreted in a catastrophic fashion, a further increase in apprehension occurs. This produces a further increase in body sensations and so on round in a vicious circle which culminates in a panic attack" (Clark, 1986). This cognitive theory of panic attacks specifies components (e.g., bodily sensations and a state of apprehension) and the relations among them (e.g., the "vicious cycle" of positive causal effects), positing that this is the target system that gives rise to panic attacks.

Because theories represent the target system, we can reason from theory in order to draw conclusions about the target system. It is this capacity for *surrogate reasoning* (Swoyer, 1991) that allows theories to explain, predict, and control. For example, we can explain the rise and fall of predator and prey populations in the real world by appealing to the relationships between components specified in mathematical models representing these populations (H. I. Freedman, 1980; Nguyen & Frigg, 2017). We can predict what will occur when two atoms collide by deriving the expected outcome from models of particle physics (Higgs, 1964). We can determine how to intervene to prevent panic attacks by appealing to the relationships posited in the cognitive model of panic attacks, determining that an intervention modifying a patient's "catastrophic misinterpretations" should prevent the "vicious cycle" between arousal and perceived threat, thereby circumventing panic attacks (Clark, 1986). This ability to support surrogate reasoning makes theories such powerful tools.

### 11.2.2 The Importance of Formal Theories

Surrogate reasoning relies on a theory's structure: its components and the relations among them (Pero, 2015; Suárez & Pero, 2019). This structure can be expressed in a written or spoken language (i.e., *verbal theory*) or in the language of mathematics or computation (i.e., *formal theory*). For example, a verbal theory would state that the rate of change in an object's temperature is proportional to the difference between its temperature and the temperature of its environment. A formal theory would instead express this relationship as a mathematical equation, such as  $\frac{dT}{dt} = -k(T - E)$ , where  $\frac{dT}{dt}$  is the rate of

change in temperature,  $T$  is the object's temperature, and  $E$  is the temperature of the environment; or in a computational programming language, such as: `for(t in 1:end) { T[t+1] = T[t]-k*(T[t]-E) }`.

Expressing a theory in a mathematical or computational programming language gives formal theories many advantages over verbal theories (e.g., Smith & Conrey, 2007; Epstein, 2008; Lewandowsky & Farrell, 2010; Smaldino, 2017). There is one advantage especially relevant to the present chapter: Formalization enables precise deduction of the behaviour implied by the theory. Verbal theories can, of course, also be used to deduce theory-implied behavior. However, due to the vagaries of language, verbal theories are typically imprecise, thereby precluding their ability to make exact predictions. For example, the verbal theory of temperature cooling described in the previous paragraph allows for some general sense of how the object's temperature will evolve over time, but cannot be used to make precise predictions about how it will change or where temperature will be at any given point in time. Indeed, because of the imprecision of verbal theories, there are often multiple ways in which those theories could be interpreted and implemented, each with a potentially divergent prediction about how the target system will evolve over time. Consider the interpersonal theory of suicide, which posits that suicide arises from the simultaneous experience of perceived burdensomeness and thwarted belongingness (Van Orden et al., 2010). This theory fails to specify many aspects of this causal structure, such as the strength of these effects or the duration for which they must overlap before suicidal behavior arises (Hjelmeland & Loa Knizek, 2018). As a result, there are many possible implementations of that verbal theory, each of which could potentially lead to a different prediction about when suicidal behaviour should be expected to arise. This imprecision thus substantially limits the theories ability to support surrogate reasoning and the degree to which we can empirically test the theory.

In contrast to most verbal theories, formal theories are precise in their implementation as the mathematical notation or code in a computer programming language forces one to be specific about the structure of the theory (e.g., specifying the precise effect of one component on another).<sup>2</sup> The precision of formal theories allows for the provision of singular and precise predictions about how the target system will behave. These predictions can either be obtained analytically from the mathematical equation or computed by implementing the formula in a programming language. For example, we can use the formal theory of cooling to predict the exact temperature of our object at any given point in time. Similarly, a formal implementation of the interpersonal theory of suicide would make precise predictions that could inform the prediction of suicide attempts. In other words, formal theories substantially strengthen surrogate reasoning, the very characteristic of scientific theories upon which we wish to capitalize.

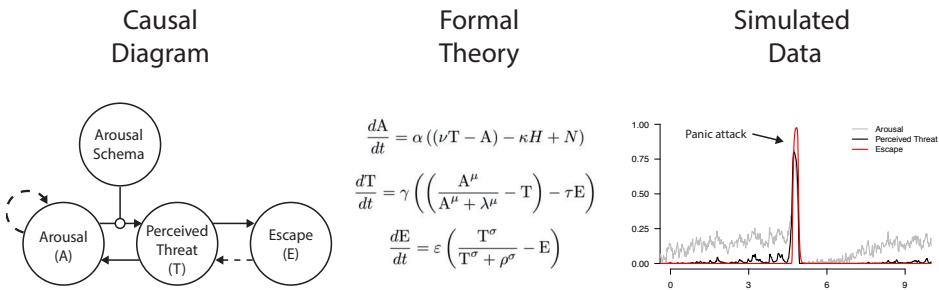
---

<sup>2</sup>It is, of course, possible to express verbal theories with the same level of precision as is provided by a mathematical equation (e.g., there are very few equations in the *Principia*, yet the laws Newton describes are not lacking in precision). Nonetheless, the specificity required by mathematics or computational programming makes them more amenable to expressing theories precisely and has the considerable practical advantage of supporting the derivation of predictions from the theory.

### 11.2.2.1 A Formal Theory of Panic Disorder

The cognitive model of panic attacks posited by Clark is a verbal theory and is limited by the imprecision characteristic of most verbal theories. Indeed, in two recent papers, Fukano and Gunji (Fukano & Gunji, 2012) and Robinaugh and colleagues (Robinaugh, Haslbeck, et al., 2019) independently proposed two distinct formal implementations of this theory, taking the verbal theory and expressing it in differential equations. Notably, these distinct implementations of the same verbal theory make divergent predictions about when panic attacks should occur, illustrating the limitations of failing to precisely specify the theory (for further detail, see Robinaugh, Haslbeck, et al., 2019).

In this chapter, we will make extensive use of the formal theory proposed by Robinaugh and colleagues. A complete description of the generation of this theory can be found in Robinaugh, Haslbeck, et al. (2019). For the purposes of this chapter, it is sufficient to note that the aim in developing this model was to take extant verbal theories, especially cognitive behavioral theories, and express them in the language of mathematics. For example, Clark’s verbal theory posits that a perception of threat can lead to arousal-related bodily sensations. However, the actual form and strength of this effect remain unspecified. In our mathematical model, we used a differential equation to precisely define this effect:  $\frac{dA}{dt} = \alpha(\nu T - A)$ . In this equation, there is a linear effect of Perceived Threat ( $T$ ) on the rate of change of Arousal ( $A$ ), with the strength of this effect specified by the parameter  $\nu$ . The product of  $\nu$  and  $T$  is the value Arousal is pulled toward: if  $\nu T$  is smaller than the current level of Arousal,  $\frac{dA}{dt}$  will be negative and Arousal will *decrease* toward  $\nu T$ ; if  $\nu T$  is greater than Arousal,  $\frac{dA}{dt}$  is positive and Arousal *increases* toward  $\nu T$ . Each model component was defined as a differential equation in this way (see middle panel in Figure 11.2).



**Figure 11.2:** The left panel displays the key components of the theory proposed by Robinaugh, Haslbeck, et al. (2019) at play during panic attacks: Arousal, Perceived Threat, Escape Behavior and arousal schema. The arrows indicate the direct causal relationships which are posited to operate between these components in the formal theory. The middle panel displays the formal theory that specifies the precise nature of the relations among these components. The right panel depicts the simulated behavior implied by the theory.

By specifying the structure of the theory in this way, we are able to solve the system numerically, thereby deducing the theory’s predictions about how the



target system will behave. For example, the theory shows that when the effect of Arousal on Perceived Threat is sufficiently strong, the positive feedback between these components is sufficient to send the system into runaway positive feedback, producing the characteristic surge of arousal, perceived threat, and escape behavior that we refer to as a panic attack (see right panel in Figure 11.2). As this example illustrates, specifying the theory as a computational model substantially strengthens our ability to deduce the behavior implied by the theory. A full realization of a theory's usefulness thus all but requires that theory be formalized. For that reason, we believe the ultimate goal of psychiatric research should not only be the production of theories, but the production of formal theories.

### 11.2.3 Data and Data Models

Our brief overview of the philosophy of science literature on theories suggests that if our aim is the explanation, prediction, and control of mental disorders, what we are after are well-developed formal theories: mathematical or computational models that represent the target system. The key question then becomes: how can we best determine such a formal theory?

The answer to this question will, of course, involve the collection and analysis of *data*. Empirical data plays at least two key roles in the development of formal theories. First, data gathered about the target system are key to establishing what our theories must explain. Yet, theories typically do not aim to explain data directly. Data are sensitive to the context in which they are acquired and subject to myriad causal influences that are not of core interest (Woodward, 2011). For example, panic disorder researchers collect data from diagnostic interviews, self-report symptom inventories, assessments of physiological arousal during panic attacks, time-series data, and a host of other methods. Data about panic attacks gathered using these methods will be influenced not only by the experienced attacks, but also by recall biases, response biases, sensor errors, and simple human error. Accordingly, theories do not aim to account for specific "raw" data. Rather, theories explain phenomena identified through robust patterns in the data that cannot be attributed to the particular manner in which the data were collected (e.g., researcher biases, measurement error, methodological artifacts, etc.). To identify these empirical regularities in data, researchers use *data models*, which are representations of the data (Suppes, 1962; Kellen, 2019). Data models can take many forms. These can range from the most basic canonical descriptive tools, such as a mean score, a correlation, or a fitted curve, to more complex statistical tools which are common in different areas of psychology and beyond; such as structural equation models, time-series models, hierarchical models, network models, mixture models, loglinear models and so forth. Essentially, we can consider a data model to be any (statistical) model that summarizes the data in some way. Thus, data models, particularly robust and replicable data models, play a key role in determining what a theory must explain.

Second, data models also inform our understanding of the components and the relations among them that are posited to give rise to a phenomenon (i.e., the theory's structure). It is this role which we will focus on in the next part of this

chapter. This role is especially noteworthy in the context of the data models most commonly used in the network approach literature: the Ising model, the Gaussian Graphical Model, and the Vector Autoregressive model. In Section 11.3 we will describe each of these models in more detail, but here it is sufficient to note their most salient feature: these analyses estimate the structure of relationships among a set of variables; specifically, the structure of conditional dependence relationships (see Figure 11.1; Top Right). There is a strong intuitive appeal to these analyses as they seem to hold the promise of directly informing the very thing we are after: the structure of relations among components of the mental disorder (see Figure 11.1; Top Left). In Section 11.3, our overarching aim will be to critically evaluate that promise and determine how best to use (network) data models to guide the development of theories about specific mental disorders.

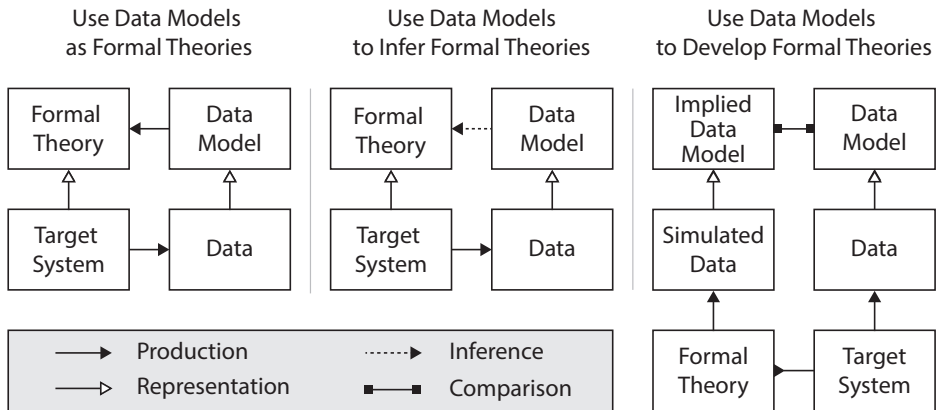
### 11.3 Identifying Formal Theories from Data

In this section we will explore how data models can best contribute to the development of formal theories. We will do so within the broader theoretical framework of conceptualizing mental disorders as complex systems and will focus on three data models that have become popular among researchers adopting this framework: the Ising model, the Gaussian Graphical Model (GGM), and the Vector Autoregressive (VAR) model. Specifically, we evaluate three routes that make use of data models in different ways to obtain a formal theory. We believe that the first two routes describe how data models are currently used in the literature, and the third route is an alternative that addresses some of the shortcomings of the first two approaches.

The first route arrives at formal theories directly by treating these data models as formal theories. In this case, the transition from data model to formal theory is largely an act of interpretation. Instead of interpreting a data model as a representation of the data, we interpret it as a representation of the target system (see Figure 11.3, Left Panel). Specifically, the variables of the data model are treated as the components of the target system, and the statistical relationships are treated as the structural relationships among the components. From this perspective, research is carried out by conducting an empirical study, estimating a data model, and treating the data model as a theory. If viable, this approach would be extremely powerful, because a well-developed theory would be just one well-designed study away. We evaluate this route in Section 11.3.1.

The second possible route arrives at formal theories by drawing inferences from data models (Figure 11.3, Middle Panel). That is, the data model is not directly treated as a theory, but rather is used to inform the theory. From this perspective, research is carried out by conducting an empirical study, estimating a data model, and using the data model to infer characteristics of the target system, thereby informing the development of a theory. For example, one could observe a conditional dependence relationship between two variables and infer the presence of a causal relationship between the corresponding components in the target system. To evaluate this approach we need to know the data-generating

target system. We do so in Section 11.3.2 by treating the Panic Model introduced in the previous section as the target system of interest, simulating data from that target system, and examining how well inferences drawn from data models can be used to inform our understanding of the target system.



**Figure 11.3:** The figure provides an overview of three routes to developing formal theories using data models. In the left panel, data models are treated as formal theories. In the middle panel, data models are used to draw inferences about the target system and, thereby, to generate formal formal theories of that system. In the right panel, data models used to develop formal theories by deducing implied data models and comparing them with empirical data models.

The third possible route puts formal theories at the heart of theory development. From this perspective, research is carried out by first generating an initial formal theory. From this formal theory we simulate data which we use to obtain the theory-implied data model. We subsequently compare the implied data model with the empirical data model, and adapt the formal theory based on the discrepancy between the two. This route thus leverages the “immense deductive fertility” of formal theories to make precise predictions that clarify how the model must be revised to be brought in line with empirical data (Meehl, 1978). From this perspective, formal theory is not only the ultimate goal of the research process, but also plays an active role in theory development. We evaluate this route in Section 11.3.3 by deriving predicted data models from a formal theory of Panic disorder, and showing how the model can be improved by comparing the predicted data models to empirical data models.

### 11.3.1 Using Data Models as Formal Theories

If data models are to serve as formal theories of a target system, the properties of those data models must be able to represent the properties we expect in the target system. Accordingly, in this section, we discuss the properties we expect in the target systems of mental disorders from the complex systems perspective (Section 11.3.1.1) and evaluate whether these properties are captured by the properties

of three data models: the VAR model, the GGM, and the Ising model (Section 11.3.1.2).

### 11.3.1.1 Properties of Mental Disorder Target Systems

Target systems consist of components and the relations among them. From the network perspective there are a number of properties we would expect to be present in the target systems of mental disorders. First, feedback loops among components are likely present. Researchers have frequently posited “vicious cycles”, where the initial activation of one component (e.g., arousal) elicits activation of other components (e.g., perceived threat) and, in turn, is reinforced by the activation of those components. Second, causal effects between components are likely to be asymmetrical. That is, the effect of component A on component B may differ from the effect of component B on component A. For example, it is unlikely that concentration has the same effect on sleep as sleep has on concentration or that compulsions have the same effect on obsessions that obsessions have on compulsions.

Third, interactions among components are likely to occur at different time scales. For example, the effect of intrusive memories on physiological reactivity in Post-traumatic Stress Disorder is likely to occur on a time scale of seconds to minutes, whereas an effect of energy on depressed mood may play out over the course of hours to days, and the effect of appetite on weight gain may occur on a time scale of days to weeks. Fourth, it is likely that there are higher order interactions among components. For example, the presence of sleep difficulties may strengthen the effect of feelings of worthlessness on depressed mood or the effect of intrusive trauma memories on physiological reactivity. If data models are to serve as formal theories of the target system, they must be able to represent these types of causal structures.

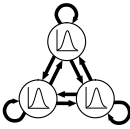


We would further suggest that most, perhaps all, mental disorder target systems are likely to have multiple stable states, that is, states into which the system settles and will remain in the absence of external perturbation. In the simplest case, the system will be characterized by the presence of two stable states: an unhealthy state (i.e., a state of elevated symptom activation, such as a depressive episode), and a healthy state (e.g., a state without elevated symptom activation). In other cases, there may be multiple stable states (e.g., healthy, depressed, and manic states in Bipolar Disorder). The presence of multiple stable states is, in turn, accompanied by other behavior often observed in mental disorders, including spontaneous recovery and sudden shifts into or out of a state of psychopathology, further suggesting that a model of any given mental disorder will almost certainly need to be able to produce alternative stable states.

### 11.3.1.2 Comparing Target System Properties with Data Model Properties

The first model we will consider is the VAR model. The VAR model for multivariate continuous time series data linearly relates each variable at time point  $t$  to all other variables and itself at previous time points (Hamilton, 1995), typically the time point immediately prior  $t - 1$  (i.e., a first order VAR, or VAR(1),

model; e.g., Bringmann et al., 2013; Pe et al., 2015; Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019). The estimated lagged effects of the VAR models indicate conditional dependence relationships among variables over time. The dynamics of the VAR model is such that the variables are perturbed by random input (typically Gaussian noise) and the variables return to their means, which represent the single stable state of the system.

As depicted in Figure 11.4, the VAR model is able to represent some key characteristics likely to be present in mental disorder target systems. Most notably, it allows for feedback loops. Variables can affect themselves both directly (e.g.,  $X_t \rightarrow X_{t+1}$ ), or via their effects on other variables in the system (e.g.,  $X_t \rightarrow Y_{t+1} \rightarrow X_{t+2}$ ). The VAR model also allows for asymmetric relationships, since the effect  $X_t \rightarrow Y_{t+1}$  does not have to be the same effect as  $Y_t \rightarrow X_{t+1}$  in direction or magnitude. However, because the lag-size (i.e., the distance between time points) is fixed and consistent across all relationships, the VAR model does not allow for different time scales. Moreover, because the VAR model only includes relations between pairs of variables, it is unable to represent higher-order interactions involving more than two variables. Finally, the VAR model has a single stable state defined by its mean vector and thus cannot represent multiple stable states of a system, such as a healthy state and unhealthy state.

	Feedback Loops	Asymmetric Relationships	Different Time Scales	Higher-order Interactions	Multiple Stable States
 VAR Model	✓	✓	✗	✗	✗
 GGM	✓*	✗	✗	✗	✗
 Ising Model	✓	✗	✗	✗	✓

**Figure 11.4:** The figure shows whether the five properties of mental disorders discussed above can be represented by the three most popular network data models, the VAR(1) model, the GGM, and the Ising model with Glauber dynamics. Note that there is a check mark at feedback loops for GGMs because one could in principle endow the GGM with a dynamic similar to the Ising model, which would essentially lead to a restricted VAR model but with symmetric relations. The asterisk is present because this endowment of dynamics is not done in practice.

The second model we will consider is the Gaussian Graphical Model (GGM). The GGM linearly relates pairs of variables in either cross-sectional (Haslbeck & Fried, 2017) or time series data (Epskamp, Waldorp, et al., 2018). In the case of time series data the GGM models the relationships between variables at the same time point. Because it does not model any dependency across time, it is typically not considered a dynamic model and, thus, could not be used to represent the behavior of a mental disorder target system as it evolves over time. In principle the GGM could be augmented by a dynamic rule similar to one commonly used with the Ising model (i.e., “Glauber dynamics”, see below). However, in that case, the GGM would become a model similar to, but more limited than, the VAR model described above (e.g., it would be limited to symmetric relationships). Accordingly, the GGM is similarly unable to represent key features we expect to observe in a mental disorder target system.

The final model we will consider is the Ising model. The Ising model again represents pairwise conditional dependence relations between variables (Ising, 1925), however, it is a model for multivariate binary data. While the original Ising model does not model dependencies over time, it can be turned into a dynamic model by augmenting it with Glauber dynamics (Glauber, 1963).<sup>3</sup> Like the VAR model, the Ising model is able to represent feedback loops. Moreover, due to its non-linear form it is able to exhibit multiple stable states (and the behavior that accompanies such stable states, such as hysteresis and sudden shifts in levels of symptom activation, see e.g., Cramer et al., 2016; Lunansky et al., 2019; Dalege et al., 2016). It is perhaps not surprising then, that the Ising model is used as a theoretical model across many sciences (Stutz & Williams, 1999), and to our knowledge, is the only of the three data models examined here that has been used as a formal theory of a mental disorder target system (Cramer et al., 2016). Unfortunately, the Ising model falls short in its ability to represent the remaining characteristics likely to be present in mental disorders. The relationships in the Ising models are exclusively symmetric; with the standard Glauber dynamics, there is only a single time scale; and the Ising model includes exclusively pairwise relationships, precluding any representation of higher-order interactions.

### 11.3.1.3 Data Models as Formal Theories?

The analysis in this section shows that the VAR, GGM, and Ising models are unable to represent most key properties we would expect in the target systems giving rise to mental disorders, and therefore cannot serve as formal theories for those disorders. Of course, more complex models would be able to produce more of the characteristics likely to be present in mental disorders. For example, one could extend the VAR model with higher-order interactions or a latent state (Tong & Lim, 1980; Hamaker et al., 2010), thereby allowing it to represent multiple stable states. However, estimating data models is subject to fundamental con-

---

<sup>3</sup>This dynamics works as follows: After specifying an initial value for each variable, it randomly picks one variable  $X_i$  at  $t = 1$  and takes a draw from the distribution of  $X_i$  conditioned on the values of all other variables. This value (either 0 or 1) is set to be the new value of  $X_i$  and then the same process is repeated, thereby allowing the model to evolve over time.

straints. More complex models require more data, and larger sample sizes which are often unavailable in psychiatric research. For example, around 90 observations (about 2.5 weeks of a typical ESM study) are needed for a VAR model to outperform the much simpler AR model (Dablander et al., 2019). Models more complex than the VAR model would require even more data to be estimated reliably. In addition, the sampling frequency (e.g., measurement every 2 hours) might be too low to capture the structure of the target system of interest (Haslbeck & Ryan, 2019). In this situation a data model still contains some information about the target system, but cannot capture the structure of the target system to the extent that it can serve as a formal theory. Even if large amounts of high frequency data were widely available, it is unclear how to estimate many complex models. For example, one could extend the Ising model with a second time scale (e.g., Lunansky et al., 2019), but it would be unclear how to estimate such a model from data. Finally, even if such models could be estimated, more complex models are often uninterpretable. For example, nonparametric models (e.g., splines; J. Friedman et al., 2001, p. 139), which can capture extremely complex behavior, typically consist of thousands of parameters, none of which can be interpreted individually. Accordingly, it is unlikely that any data model estimated from the type of data typically available in psychiatric research will be both interpretable and capable of capturing the characteristics of psychopathology in such a way that would allow it to serve as a formal theory of a mental disorder.

### 11.3.2 Using Data Models to Infer Formal Theories

An alternative route from data models to formal theories is to use data models to draw inferences about a target system, inferences that we can use to construct a formal theory. There is good reason to think that this approach could work. Because the data are generated by the target system, and data models summarize these data, the parameters of any data model certainly *somehow* reflect characteristics of the target system. This means that it should be possible, in principle, to infer something about the target system and its characteristics from data and data models. Although we have seen already that the GGM, Ising and VAR models cannot directly reproduce the key characteristics of the target system, their parameters could potentially still yield insights into the structure or patterns of relationships between components. In line with this intuition, it has frequently been suggested that the GGM, the Ising model, and the VAR models can serve as “hypothesis-generating tools” for the causal structure of the target system (e.g., Borsboom & Cramer, 2013; van Rooijen et al., 2017; Fried & Cramer, 2017; Epskamp, van Borkulo, et al., 2018; Epskamp, Waldorp, et al., 2018; Jones, Mair, Riemann, Mugno, & McNally, 2018).

Although this approach seems intuitive, in practice it is unclear how this inference from data model to target system should work. For example, if we observe a strong negative cross-lagged effect of  $X_t$  on  $Y_{t+1}$  in a VAR model, what does that imply for the causal relationship between the corresponding components in the target system? A precise answer to this question would require a rule that connects parameters in particular data models to the structure of the target system.

For some simple systems, such a rule is available. For example, if the target system can be represented as a Directed Acyclic Graph (DAG), then under certain circumstances its structure can be inferred from conditional (in)dependence relations between its components: Conditional independence implies causal independence, and conditional dependence implies either direct causal dependence or a common effect (Pearl, 2009; Ryan, Bringmann, & Schuurman, 2019). However, it is generally unclear how we can use the parameters of typical data models to make inferences about the types of non-linear dynamic systems we expect in a psychiatric context (although Mooij, Janzing, & Schölkopf, 2013 and Forré & Mooij, 2018 have established some links in this regard). The consequences of this are twofold. First, any inference from data model to target system must rely instead on some simplified heuristic(s) in an attempt to approximate the link between the two. Second, it is unclear how well the combination of common data models and simple heuristics perform in allowing us to make inferences about the target system.

In this section, we evaluate whether the three data models introduced above can be used to make inferences about mental disorder target systems. To do this, we treat the Panic Model discussed in Section 11.2 as the data-generating target system and compare the causal structure inferred from the data models to the true causal structure. To yield these inferences we use a very simple and intuitive set of heuristics: a) if two variables are conditionally dependent in the data model, we will infer that the corresponding components in the target system are directly causally dependent; b) if there is a positive linear relationship, we will infer that the causal relation between the corresponding components is positive (i.e., reinforcing); c) if there is a negative linear relationship, we will infer that the causal relationship among components is negative (i.e., suppressing).

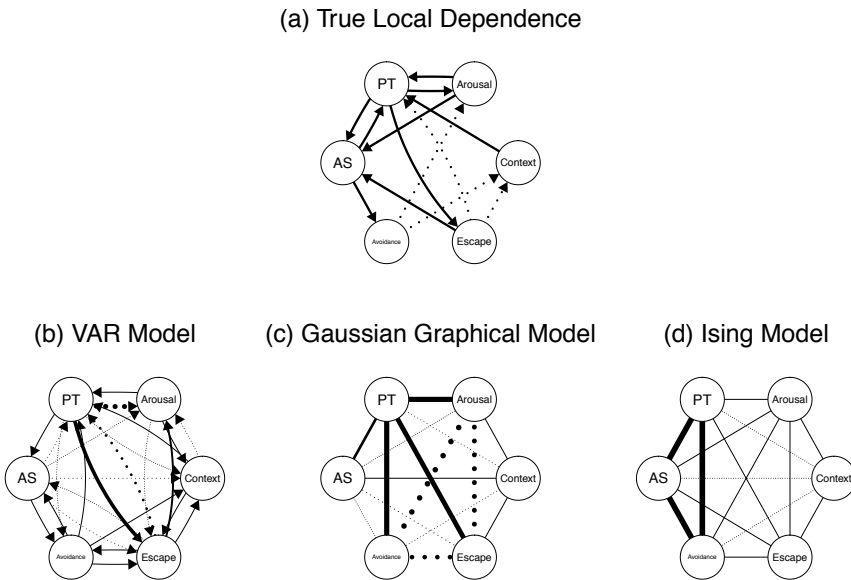
### 11.3.2.1 Inferring the Panic System from Network Data Models

To be able to evaluate the success of the simple heuristics described above, we must first represent the structure of the Panic Model (see Section 11.2) in the structure of a square matrix, that is, in the same form as the parameters of the VAR, GGM, and Ising models. Since the relationships between components are formalized through *differential equations*, a natural choice is to represent the Panic Model as a network of moment-to-moment dependencies, drawing an arrow  $X \rightarrow Y$  if the rate of change of  $Y$  is directly dependent on the value of  $X$  (known as a *local dependence graph*; Didelez, 2007). Figure 11.5 (a) displays these moment-to-moment dependencies. Note that this structure cannot capture many aspects of the true model, such as the presence of two time scales or the moderating effect of Arousal Schema (AS) (see Section 11.2 for details). It is, thus, already clear that the models cannot recover the *exact* causal structure of the Panic Model. Nonetheless, we can still investigate whether applying the simple heuristics to these three data models allows us to infer this less detailed pattern of direct causal dependencies.

We next compare this true causal structure to the causal structure inferred based on the three data models. To obtain the three data models, we first gen-



erate data from the target system (See Appendix H.1). Specifically, we use four weeks of minute-to-minute time-series data for 1000 individuals. These individuals differ in their initial value of Arousal Schema, with the distribution chosen so that the proportion of individuals for whom a panic attack is possible was equivalent to the lifetime history prevalence of panic attacks in the general population (R. R. Freedman et al., 1985). For the VAR model analysis, we create a single-subject experience-sampling-type dataset by choosing the individual who experiences the most (16) panic attacks in the four-week period. To emulate ESM measurements, we divide the four week period into 90-minute intervals, taking the average of each component in that interval, yielding 448 measurements. For the GGM analysis, we create a continuous cross-sectional dataset by taking the mean of each component for each individual over the four weeks. For the Ising model analysis, we obtain cross-sectional binary data by taking a median split of those same variables. The resulting VAR, GGM and Ising model networks are displayed in Figure 11.5 panels (b), (c) and (d), respectively.<sup>4</sup>



**Figure 11.5:** Panel (a) shows the true model in terms of local dependencies between components; panel (b) shows the VAR model estimated from ESM data sampled from the true model; panel (c) shows the GGM estimated from the cross-sectional data of 1000 individuals, generated from the true model; panel (d) shows the Ising model estimated on the same data after being binarized with a median split. Solid edges indicate positive relationships, dotted indicate negative relationships. For panels (b) to (d), the widths of edges is proportional to the absolute value of the corresponding parameter. Note that in panel (b) we do not depict the estimated auto-regressive parameters as the primary interest is in inferring relationships between variables.

<sup>4</sup>Note that in the Ising model the parameter estimates are somewhat unstable due to near-deterministic relationships between some binarized variables.

We will focus our evaluation on two important causal dependencies in the target system: the positive (i.e., reinforcing) moment-to-moment feedback loop between Perceived Threat and Arousal, and the positive effect of Arousal Schema (i.e., beliefs that arousal-related bodily sensations are dangerous) on Avoidance (i.e., efforts to avoid situations or stimuli that may elicit panic attacks). In the VAR model (panel (b) in Figure 11.5) we see a lagged positive relationship of Arousal to Perceived Threat, a strong *negative* lagged relationship from Perceived Threat to Arousal, and a weak positive effect of Arousal Schema on Avoidance. Applying the heuristics, we would infer a reinforcing relationship from Arousal to Perceived Threat, a suppressing relationship from Perceived Threat to Arousal, and a reinforcing effect of Arousal Schema on Avoidance. In the GGM (panel (c) in Figure 11.5) we see a positive conditional dependency between mean values of Arousal and Perceived Threat, but we also see a weak negative dependency between mean values of Arousal Schema and Avoidance. Applying the heuristics to the GGM, we would infer a reinforcing relationship between Arousal and Perceived Threat, and a suppressing relationship between Arousal Schema and Avoidance. Finally, in the Ising model (panel (d) in Figure 11.5), we see a strong positive dependency between Arousal Schema and Avoidance, and a very weak positive relationship between PT and Arousal. This leads us to infer two reinforcing relationships, between Arousal and Perceived Threat, and Arousal Schema and Avoidance.

For the VAR model, the heuristics yield one correct and one incorrect inference. For the GGM, we make exactly the opposite inferences, with again one correct and one incorrect. In the Ising Model, we yield two correct inferences. However, inspecting the rest of the Ising Model edges we can see a variety of incorrect inferences about other relationships, with independent components in the target system connected by strong edges in the Ising model, and the valences of various true dependencies flipped. At best, we can say that in each of the three network models, some dependencies do reflect the presence and/or direction of direct causal relationships, and some do not. Unfortunately, it is not possible to distinguish which inferences are trustworthy and which are not without knowing the target system, and in any real research context, the target system will be unknown. Consequently, these data models cannot be used to confidently and reliably draw inferences about the target system using these simple heuristics.

### 11.3.2.2 The Mapping between Data Model and Target System

Importantly, our inability to draw accurate inferences from these data models is not a shortcoming of the data models themselves. Each data model correctly captures some form of statistical dependency between the components in a particular domain (e.g., lagged 90 minute windows). Moreover, the statistical dependencies in the data models are produced by causal dependencies in the target system, so we know there is *some* mapping from the causal dependencies in the target system to statistical dependencies in the data model. The fundamental barrier to inference is that the form of this mapping is unknown and considerably more complex than the simple heuristics we have used to draw inferences here.

For example, consider the relationships between Perceived Threat and Arousal. The VAR model (panel (b) in Figure 11.5), identifies a negative lagged relationship from Perceived Threat to Arousal in the data generated by the target system. Yet in the target system, this effect is positive. This “discrepancy” occurs because of a very specific dynamic between these components: After a panic attack (i.e. a brief surge of Perceived Threat and Arousal) there is a “recovery” period in which arousal dips below its mean level for a period of time. As a result, when we average observations over a 90 minute window, a high average level of Perceived Threat is followed by a low average level of Arousal whenever a panic attack occurs. That same property of the system produces the observed findings for the GGM and Ising Model through yet another mapping (for details, see Appendix H.2).

As this example illustrates, the mapping between target system and data model is intricate, and it is unlikely that any simple heuristics can be used successfully to work backwards from the data model to the exact relationships in the target system. We can expect this problem to arise whenever we use relatively simple statistical models to directly infer characteristics or properties of a complex system (c.f. the problem of under-determination or indistinguishability; Eberhardt, 2013; Spirtes, 2010). Indeed, the same problem arises even for simpler dynamical systems when analyzed with more advanced statistical methods (e.g., Haslbeck & Ryan, 2019). Of course, in principle, it must be possible to make valid inferences from data and data models to some properties of a target system using a more principled notion of how one maps to the other. For example, under a variety of assumptions, it has been shown that certain conditional dependency relationships can potentially be used to infer patterns of local causal dependencies in certain types of dynamic system (Mooij et al., 2013; Bongers & Mooij, 2018; Forré & Mooij, 2018). However, the applicability of these methods to the type of target system we expect to give rise to psychopathology (see Section 11.3.1) is as yet unclear and even under the strict assumptions under which they have been examined, these methods still do not recover the full structure of the target system.<sup>5</sup> This means that the intricacy of the mapping between target system and data model currently precludes us from making reliable inferences about the target system. Accordingly, we cannot use those inferences to build formal theories.

---

<sup>5</sup>Specifically, Mooij et al. (2013) and Bongers and Mooij (2018) have shown that cyclic causal models can be conceptualized as encoding causal dependencies between the equilibrium positions of deterministic differential equations and differential equations with random initial values. Forré and Mooij (2018) formally link the conditional dependencies between equilibrium position values to the causal dependencies in these cyclic causal models using a considerably more complex mapping rule than that which holds for DAGs. Their applicability to the current context is limited in the sense that 1) to our knowledge these rules have not been extended to dynamic systems with time-varying stochastic terms (SDEs) as we would expect to see in complex psychological systems (and on which the Panic Model is based), and 2) the use of these methods is reliant on data that reflects equilibrium positions. Future developments in this area may prove to yield useful tools for psychological theory development however, and we consider this area to be ripe for future research beyond the scope of the present chapter.

### 11.3.3 Using Data Models to Develop Formal Theories

In Section 11.3.2, we saw that the mapping between target system and data model is intricate and would be nearly impossible to discern when the target system is unknown. However, we also saw that when the target system is known, we can determine exactly which data models the target system will produce. Indeed, this is precisely what we did when we simulated data and fit data models to it in the previous section. In this section we consider a third route to formal theories, which makes use of this ability to determine which data models are implied by a given target system (or formal theory).

This third route works as follows. First and foremost, we must propose *some* initial formal theory which we take as a representation of the target system. The quality or accuracy of this representation may be good or bad, but crucially the theory must be formalized in such a way as to yield unambiguous predictions (see Section 11.2.2). Second, we can use this initial theory to deduce a theory-implied data model. This can be done by simulating data from the formal theory and fitting the data model of interest. Third, we can learn about and adapt the formal theory by comparing implied data models with their empirical counterparts. This approach is represented in schematic form in the right-hand panel of Figure 11.3. It can be seen as a form of inference, but it is *abductive inference*: inference to the best explanation (Haig, 2005). We first infer the best explanation for the core phenomena to generate an initial theory. We then infer the best explanation for any discrepancies between empirical and theory-implied data models, inferences which inform subsequent theory development. Given the importance of abductive inference to this approach, we adopt this term to refer to this third route. In this sub-section, we will illustrate this approach using the example of panic disorder (for an overview, see Figure 11.6).

#### 11.3.3.1 Obtaining Theory-Implied and Empirical Data Models

In this section, we will treat the Panic Model introduced in Section 11.2 as our initial formal theory, which should represent the target system that gives rise to panic disorder (Figure 11.6, bottom row). The Panic Model can be used to simulate data and, in turn, to derive predictions made by the theory in the form of theory-implied data models (left-hand column of Figure 11.6). While in principle many data models can (and should) be used to perform the abductive inference described above, here we will examine the implied cross-sectional Ising model of the three core panic disorder symptoms: 1) Recurrent Panic Attacks (PA), 2) Persistent Concern (PC) following a panic attack and 3) Avoidance (Av) behaviour following a panic attack (American Psychiatric Association, 2013). If our formal theory of panic disorder is an accurate representation of the target system that gives rise to panic disorder, the implied Ising model derived from this theory should be in agreement with a corresponding Ising model derived from empirical data.

Critically, obtaining an implied data model requires not only a formal theory from which we can simulate data, but also a formalized process by which variables are “measured” from those data. The Panic Model is used to generate



symptom present if an individual has a panic attack, and their average levels of avoidant behaviour following that attack are higher than we would expect to see in the healthy sample. A more detailed account of how we generated these data can be found in Appendix H.3. This simulated cross-sectional data was then used to estimate the implied Ising Model (top left-hand corner, Figure 11.6).<sup>6</sup>

We obtained the corresponding empirical Ising model (right-hand column of Figure 11.6) using the publicly available Collaborative Psychiatric Epidemiology Surveys (CPES) 2001-2003 (Alegria, Jackson, Kessler, & Takeuchi, 2007). The CPES is a nationally representative survey of mental disorders and correlates in the United States, with a total sample size of over twenty thousand participants (of which  $n = 11367$  are used in the current analysis; for details see Appendix H.3). The CPES combines more than 140 items relating to panic attacks and panic disorder, with a diagnostic manual describing how these items can be re-coded into binary symptom variables reflecting Recurrent Panic Attacks, Persistent Concern and Avoidance. PA is present if the participant reported more than three lifetime panic attacks. PC is present if, following an attack, the participant experienced a month or more of persistent concern or worry. Av is present if the participant reports either a month of avoidance behaviour following an attack, or a general avoidance of activating situations in the past year.

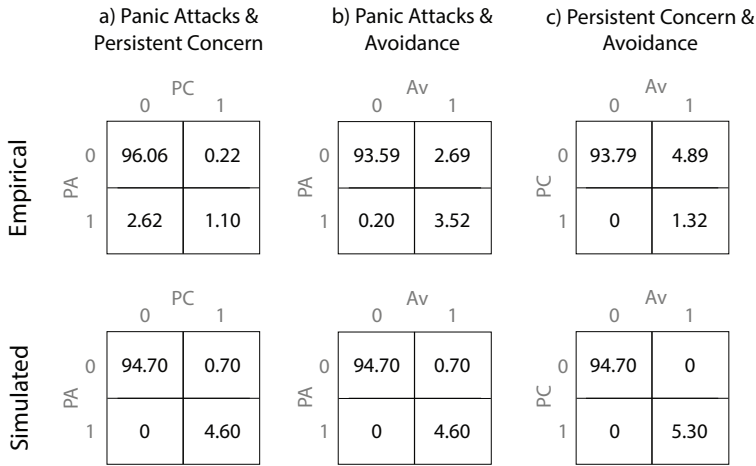
### 11.3.3.2 Theory Development: Comparing Model-Implied and Empirical Data Models

As seen in Figure 11.6, there is a similar pattern of conditional dependencies in the implied and empirical data models. In both, all pairwise dependencies are positive, and all thresholds are negative. There is also a similar ordering of conditional dependencies in terms of their magnitude. Within each model, the conditional relationships of PA with Av and PA with PC are of the same order of magnitude, and the conditional relationship between Av and PC is an order of magnitude greater. However, we also see some differences between the models. First, the absolute value of pairwise dependencies and thresholds are much greater in the implied Ising Model (Figure 11.6 (a)) than the empirical Ising Model (Figure 11.6 (b)). Second, we see that the relationships in the implied model are perfectly symmetric, with exactly the same thresholds for Av and PC, and precisely the same weights relating PA to both. In the empirical network, these weights and thresholds are much smaller.

The bivariate contingency tables of all symptom-symptom relationships clarify the nature of these relationships (see Figure 11.7). In both the implied and empirical data models only a small proportion of individuals experience Recurrent Panic Attacks (Empirical 4.3%, Simulated 3.72%). Crucially, in the simulated dataset, the symptom relationships are almost deterministic: If one symptom is present, so too are all others, and vice versa for the absence of symptoms (apart from three individuals who experience less than three panic attacks in the time window). This is because there is a deterministic relationship between the com-

---

<sup>6</sup>Here again the Ising model estimates are somewhat unstable due to near-deterministic relationships between the variables.



**Figure 11.7:** Contingency Tables showing percentages for each pair of symptom variables (one per column) for the empirical data (top row) and simulated data (bottom row). The CPES contingency tables are based on  $n_{CPES} = 11367$  observations. The simulated dataset contains  $n_{sim} = 1000$  observations.

ponents underlying these symptoms in the Panic Model: All participants who experience one panic attack have Persistent Concern and Avoidance behaviour after those attacks. In contrast, there are non-deterministic relationships in the empirical data. For example, it is actually more common to have Recurrent Panic Attacks without Persistent Concern than with Persistent Concern (column (a)). Similarly, more individuals experience Avoidance without Persistent Concern, than with Persistent Concern (column (c)) Conversely, there are no individuals who experience Persistent Concern but not Avoidance.

The discrepancies between the implied and the empirical data model could arise at any step in the process from formal theory/target system to implied and empirical data model illustrated in Figure 11.6. It could be the case that a discrepancy is due to inaccuracies in how we emulate the measurement process. For example, perhaps Persistent Concern and Avoidance co-occur equally, but the former suffers from a greater degree of recall bias than the latter (for an example of differential symptom recall bias in depressed patients, see Ben-Zeev & Young, 2010). There are also different time scales at which the simulated and empirical symptoms are defined. The simulated symptoms are defined over a month period whereas the CPES items are defined over lifetime prevalence. Due to the deterministic nature of the Panic Model, we believe a month period is a good approximation for lifetime experience of panic symptoms in this case. Nonetheless, it is a discrepancy in measurement that could lead to discrepancies between the implied and empirical data models. It could also be that the discrepancy is due to estimation issues. However, due to the large sample sizes and simple models used, we suspect it is unlikely that sampling variance is a problem in this instance. For present purposes, we will assume here that discrepancies are to due

to inaccuracies in how the theory represents the target system, and as such we can use these discrepancies to directly evaluate our theory.<sup>7</sup>

On a global level there is a good match between the empirical and implied models: The theory implies positive symptom-symptom dependencies, which we also observe in the empirical data. However, the implied model over-estimates the strength of these relationships. This is largely explained by the deterministic causal effects in the theory. In the simulated data, everybody who experiences panic attacks also develops Persistent Concern and, in turn, Avoidance. As seen in Figure 11.7, this is inconsistent with empirical data, identifying a serious shortcoming in the theory. To improve the model, we must include some mechanism by which individuals can experience a panic attack without developing the remaining symptoms of panic disorder. In the empirical data, there is a near deterministic effect of Persistent Concern on Avoidance, suggesting that once Persistent Concern develops, Avoidance will follow; an observation that is consistent with the formal theory. However, inconsistent with the formal theory, the empirical data suggests a relatively low probability of Persistent Concern following panic attacks, suggesting that this is where the theory must be revised if it is to better account for the observation that some individuals experience recurrent panic attacks without developing the full panic disorder syndrome.

This is just one discrepancy in these data that can inform model development and more insights may be gained by focusing on others. Many more insights can be gained by considering different data models based on different data. For example, experimental data on the relation between Arousal and Perceived Threat may allow us to refine the specification of the feedback between those two variables. In general this route offers a great deal of flexibility in theory development. Although the theory is likely to be complex, dynamic and non-linear, the form of the data models used to learn about that theory need not be. Instead, by starting with an initial theory, the researcher can use any data about the phenomena of interest to further develop that theory. In the following section we will provide a full account of this third route by discussing the full process of theory development from establishing the phenomenon, to making novel predictions with a well-developed formal theory.

## 11.4 An Abductive Approach to Constructing Formal Theory

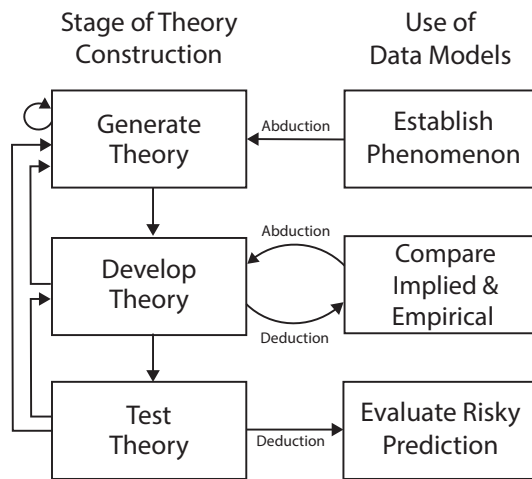
In Section 11.3.3 we illustrated a clear approach to use empirical data to develop an existing formal theory. However, our description of this approach so far has omitted several critical steps, including how to generate an initial formal theory

---

<sup>7</sup>In practice, inaccurate conceptualizations of how measurements represent the target system will be problematic for any approach to theory development or indeed any scientific endeavour, as evidenced by the growing attention on measurement in psychopathology literature (e.g., Fried & Flake, 2018). Our proposed approach to formal theory development forces us to be specific about the measurement process, just as we are forced to explicate the theory itself. Although we focus on the formalized theory itself here, we consider the formalization of measurement to be a significant advantage of this approach, and one that warrants further development.



and how to test that theory. In this section, we propose a three-stage process of formal theory construction, with an emphasis on the role data models play at each stage (see Figure 11.8). First, in the theory generation stage, we establish the phenomenon to be explained, generate an initial verbal theory, and formalize that theory. Second, in the theory development stage, the theory is developed beyond this initial proposal by adapting it such that it is consistent with as many empirical findings as possible. Finally, in the theory evaluation stage, the theory is subjected to strong tests within a hypothetico-deductive framework. The approach to theory construction proposed here places considerable emphasis on the theory’s ability to explain phenomena, especially during the generation and development of the theory. Accordingly, the framework we have proposed is a largely abductive approach (Haig, 2005).



**Figure 11.8:** Flowchart depicting the process of developing a formal theory with the abductive approach put forward in this section. In the theory generation step we first establish the phenomenon (Section 11.4.1.1) and then generate an initial verbal theory (Section 11.4.1.2) which is subsequently formalized (Section 11.4.1.3). In the second step (Section 11.4.2) the theory is validated by testing whether it is consistent with existing empirical findings that are not part of the core phenomenon. If the formal theory is not consistent with some findings, it is adapted accordingly. If these adaptations lead to a “degenerative” theory (Meehl, 1990a) we return to the first step; otherwise we continue to the final step, in which we test the formal theory using risky predictions (Section 11.4.3). If many tests are successful, we tentatively accept the theory. If not, the theory must either be adapted (step two) or a new theory generated (step one).

## 11.4.1 Generating Theory

### 11.4.1.1 Establishing the Phenomenon

The goal of a formal theory is to explain phenomena. Accordingly, the first step of theory development is to specify the set of phenomena to be explained. Establishing phenomena is a core aim of science and a full treatment of how best to achieve this aim is beyond the scope of this chapter (for a possible way to organize

this process see Haig, 2005). However, we suspect that the most appropriate phenomena for initial theory development will often include things that researchers would not think to subject to empirical analysis, as the most robust phenomena may simply be taken for granted as features of the real world. For example, in the case of panic disorder, the core phenomena to be explained are simply the observations that some people experience panic attacks and recurrent attacks tend to co-occur with persistent worry or concern about those attacks and avoidance of situations where such attacks may occur. These are empirical phenomena so robust that it is typically not empirically examined but instead simply assumed to be a feature of the real world.

#### 11.4.1.2 Generate Initial Verbal Theory

Once the phenomena to be explained have been established, how do we go about generating an initial theory to explain them? A brief survey of well-known scientific theories reveals that this initial step into theory is often unstructured and highly creative. For example, in the 19th century August Kekulé dreamt of a snake seizing its own tail, leading him to generate the theory of the benzene ring, a major breakthrough in chemistry (Read, 1995). In the early 20th century, Alfred Wegener noticed that the coastlines of continents fit together similar to puzzle pieces, and consequently developed the theory of continental drift (Wegener, 1966), which formed the basis for the modern theory of plate tectonics (Mauger, Tarbuck, & Lutgens, 1996). In the late 20th century, Howard Gardner explained that he developed his theory of multiple intelligences in the 1980s using “subjective factor analysis” (Walters & Gardner, 1986, p. 176). Although more codified approaches to theory development exist (e.g., Grounded Theory; Strauss & Corbin, 1994), we are unaware of any evidence to suggest that any one approach to theory generation is superior to any other.

Nonetheless, the nature of theories does provide some guidance for how they might initially be generated. Theories achieve their aim of explaining phenomena by representing a target system. Accordingly, generating an initial theory will require that we specify the components thought to compose the target system. This process entails dividing the domain of interest into its constituent components (i.e., “partitioning”) and selecting those components one believes must be included in the theory (i.e., “abstraction”), thereby producing a system of components that will be the object of the theory (cf. Elliott-Graves, 2014). For researchers adopting a “network perspective”, the target system is typically presumed to comprise cognitive, emotional, behavioral, or physiological components, especially those identified in diagnostic criteria for mental disorders (Borsboom, 2017). For example, as we have seen in Section 11.2, the target system could consist of Arousal, Perceived Threat, Avoidance and other symptoms of panic attacks or panic disorder. Having identified the relevant components we next specify the posited relations among them. For example, specifying that Perceived Threat leads to Arousal. Within the domain of the network approach, this second step will typically entail specifying causal relations among symptoms or momentary experiences (e.g., thoughts, emotions, and behavior).

Notably, in psychiatry, we do not necessarily need to rely on creative insight about the components and relations among them in order to generate an initial theory. There are already a plethora of verbal theories about mental disorders. If the initial verbal theory is well supported and specific, it will lend itself well to formalization and subsequent theory development. However, even poor verbal theories can be a useful starting point to developing a successful formal theory (Wimsatt, 1987; Smaldino, 2017).

### 11.4.1.3 Formalize Initial Theory

Once a verbal theory has been specified, the next step is to formalize the theory. To do so, we first need to choose a formal framework. A common formal framework is the use of difference or differential equations, which model how variables change across discrete time steps and continuous time, respectively (e.g., Strogatz, 2015). Specifically, the relations between components is specified by defining the rate of change of each component as a function of all other components and itself. The Panic Model, which we used as an example throughout this chapter, uses this formal framework. Another common framework is Agent based Modeling (ABM), in which autonomous agents interact with each other using a set of specified rules (e.g., Grimm & Railsback, 2005). Here, each agent has local rules on how to interact with other agents. Both frameworks can be implemented in essentially any computer programming language and both are likely to be relevant to psychiatric and psychological research as a whole.

Having chosen a formal framework, the next step is to specify the relations between each component in the language of that framework. This process of formalizing relations is an exercise in being specific. Mathematics and computational programming languages require theorists to specify the precise nature of the relationship between variables. Requiring this level of specificity is one advantage of computational modeling, as it has the effect of immediately clarifying what remains unknown about the target system of interest, thereby guiding future research. However, this also means that theorists will often be in the position of needing to explicate relationships when the precise nature of those relationships is uncertain. We believe that, even in the face of this uncertainty, it is better to specify a precise relationship and be wrong than to leave the relationship ambiguously defined, as it is in a verbal theory. Nonetheless, we suspect that theorists will be on firmer foundation for subsequent theory development the more that they are able to draw on empirical data and other resources to inform this initial formal theory. There are several sources of information that can guide the formalization process.

First, empirical research can inform specification of components and the relations among them. For example, one could use the finding that sleep quality predicts next-day affect, but daytime affect does not predict next-night sleep (de Wild-Hartmann et al., 2013) to constrain the set of plausible relationships between those two variables in the formal theory. There could also be empirical data on the rate of change of variables, for example, Siegle, Steinhauer, Thase, Stenger, and Carter (2002) and Siegle, Steinhauer, Carter, Ramel, and Thase (2003) have

shown that depressed individuals exhibit longer sustained physiological reactions to negative stimuli than healthy individuals, a finding which is echoed in self-report measures of negative affect (Houben et al., 2015).

Second, we can possibly derive reasonable scales for variables and relationships between variables from basic psychological science. For example, classical results from psychophysics show that increasing the intensity of stimuli in almost all cases leads to a nonlinear response in perception (e.g., Fechner, Howes, & Borning, 1966): When increasing the volume of music to a very high level, individuals cannot hear an additional increase.

Third, in many cases we can use definitions, basic logic, or common sense to choose formalizations. For example, by definition emotions should change at a time scale of minutes (Houben et al., 2015), while mood should only change at a time scale of hours or days (Larsen, 2000). And we can choose scales of some variables using common sense, for example one cannot sleep less than 0 and more than 24 hours a day, and heart rate should be somewhere between 50 and 180.

Fourth, we could use an existing formal model of another target system, which we expect to have a similar structure as the target system giving rise to the phenomenon of interest. This approach is called “analogical modeling”. For example, Cramer et al. (2016) formulated a model for interactions between symptoms of Major Depression using the Ising model, which was originally formulated to model magnetism on an atomic level (Ising, 1925). Similarly, Fukano and Gunji formulated a model for interactions among core components of panic attacks using a Lotka-Volterra model originally formulated to represent predator-prey relationships (Fukano & Gunji, 2012).

Fifth, it is also important to note that there are methods by which we can potentially estimate the parameters for a formal theory from empirical data.<sup>8</sup> These approaches require considerable development of the formal theory (e.g., the form of a differential equation), suitable data (typically intensive longitudinal data), and a clear measurement model relating observed variables to theory components (as we did in Section 11.3.3). Accordingly, this approach already requires considerable progress in generating a formal theory and may be limited by practical considerations. Nonetheless, it remains a valuable resource that, if successfully carried out, would likely strengthen subsequent efforts at theory development.

The aim of this initial stage is to generate a formal theory able to explain a set of core phenomena. As we have emphasized throughout this chapter, formal theories precisely determine the behavior implied by their theory. Accordingly, explanation in this context means that the theory has demonstrated its ability to produce the behavior of interest. For example, a theory of panic attacks must be able to produce sudden surges of arousal and perceived threat; a theory of depression must be able to produce sustained periods of low mood; and a theory of

---

<sup>8</sup>For example, if the theory is formalized in a system of differential equations, the parameters of such equations can in principle be estimated from time series data using, amongst others, Kalman filter techniques and state-space approaches (e.g., Einicke, 2019; Kulikov & Kulikova, 2013; Durbin & Koopman, 2012). For implementations of these estimation methods see Ou et al. (2019); Carpenter et al. (2017); King, Nguyen, and Ionides (2015)

borderline personality disorder must be able to produce affective instability. We would note that there are very few theories in psychiatry that have reached this stage of not merely positing, but demonstrating, that the theory can explain the phenomena of interest. Accordingly, completing this stage of theory construction would constitute a significant advance in psychiatric theories. Once a theory has reached this stage, it is ready for the next stage of theory construction.

### 11.4.2 Developing Theory

The formal theory produced in the first stage of theory construction will have demonstrated its ability to explain the core phenomena of interest. However, the fact that the formal theory provides *some* explanation does not mean it is the *correct* explanation. In other words, demonstrating an ability to explain the phenomena of interest is a critical first step, but does not guarantee that the formal theory is a good representation of the target system. To achieve this aim, we propose a stage of theory development in which the theory is refined by deducing implied data models and comparing them to empirical data models. If the two data models align we take this as evidence that the current formal theory is adequately representing the target system; if there are discrepancies between the two data models, we analyze the nature of those discrepancies, consider the best explanation for how they arose, and adapt the formal theory to be able to account for these discrepancies.

This process of further developing a theory through comparisons between implied and empirical data models is exactly the process that we have illustrated already in Section 11.3.3. In that illustration, we derived the implied Ising model for the three variables Panic Attacks, Persistent Concern, and Avoidance, and we compared it to the empirical Ising model for the same variables. We discussed possible explanations for this discrepancy and corresponding adaptations of the Panic Model, for example including a mechanism by which individuals can experience Panic Attacks without experiencing Persistent Concern and Avoidance. If we were to continue in this stage of theory development, we would iterate this process, adapting the Panic Model to include such a mechanism, deriving the implied Ising model from this adapted Panic Model, and determining whether it better accounts for the implied Ising model.

This stage of theory development can make use of many different types of data such as physiological, psychological and behavioral measurements from individuals, cross-sectional data such as clinical interviews and questionnaire data, or experimental data. Depending on the empirical phenomenon we would like to account for, different kinds of data and data models will be appropriate. In general, however, more complex data models tend to be more powerful tools to tease apart competing theories. For example, a great many formal theories might be consistent with a set of means, but it is likely that fewer are consistent with the means *and* the conditional relationships between the variables, captured, for example, by a GGM or Ising model. In other words, there are a more constrained number of possible formal theories that may account for more complex data models, thereby doing more to guide theory development.

There are two important considerations when working through this stage. First, it is important to consider how much trust to place in the empirical data at hand. Discrepancies between the empirical data model and theory-implied model may be due to shortcomings of the formal theory, but they may also be due to poor measurement, insufficient samples, or poor estimation of the parameters of the data model. How do we know when to adapt the formal theory in the face of some discrepancy? For some guidance on this question, we can draw on the large literature on model evaluation and model comparison. A straightforward way to decide whether to adapt the theory would be to derive an implied model of the adapted theory, and then compare the likelihood of the empirical data given the initial theory and the adapted theory. In order to decide whether to accept the adaptation we can use, for instance, a likelihood ratio test or a Bayes factor. This procedure ensures that we only make adaptations to our theory if we are certain enough that they actually lead to a better representation of the target system, and not only the idiosyncratic features of the empirical data at hand. That is, whether we accept an adaptation of the formal theory depends both on how large the improvement is, and how certain we are about it (i.e. how large the sample size is).

Second, it is important to consider the danger of making too many ad hoc revisions to the model that account only for idiosyncratic features of a given data model or, worse, yield new implications that are inconsistent with other empirical findings. For some guidance on this question, we can draw on the literature on theory evaluation from the philosophy of science literature (Meehl, 1990a; Lakatos, 1976), which would suggest that the theory development phase has two possible outcomes. If the theory is adapted almost every time it is tested against empirical data, if those adaptations are making the theory increasingly unwieldy, and if additional changes are increasingly difficult to make without causing the theory to be inconsistent with earlier tested empirical findings, the theory can be considered to be “degenerative” (Meehl, 1990a; Lakatos, 1976). In such a situation, the initial theory was inappropriate and we return to the first step to generate a different initial formal theory. On the other hand if modifications to a theory expand, rather than contract, its ability to account for other empirical data beyond those it was originally introduced to explain, then we can have greater confidence in those modifications and, in turn, the formal theory. Ultimately, theorists must strive for a balance between the simplicity of the model and its consistency with empirical data models.

The aim of the theory development stage is a formal theory that not only explains the core phenomena of interest, but is also consistent with a range of empirical data models. Our confidence in such a theory will grow the more data models and the more complex the data models that are consistent with theory, especially if the theory is able to achieve this consistency with minimal ad hoc adjustments. In other words, if a theory has achieved these aims, we can be increasingly confident that it is a good representation of the target system and can prepare to subject the theory to more rigorous testing.

### 11.4.3 Testing Theory

Much of psychiatric research begins at this stage: positing and testing a hypothesis, typically through null hypothesis testing. However, the theories from which these hypotheses are derived are often unclear and, as we have argued elsewhere in this chapter, the process by which hypotheses (i.e., predictions) are derived from these theories is opaque and likely prone to error. This is perhaps not surprising as the hypothetico-deductive framework in which much of this research is conducted has very little to say about where these theories come from or how they should be developed (Haig, 2005). In the framework proposed here, we have attempted to detail a process by which theories can be generated and developed and by which clear hypotheses can be unambiguously derived from those theories. Accordingly, while we have characterized this framework as being primarily abductive in nature with a focus on its ability to explain phenomena, we also believe it substantially strengthens hypothesis testing as a tool for evaluating theories.

Importantly, the theory testing stage calls for strong tests of a theory: risky predictions (Meehl, 1990a) that render the theory vulnerable to refutation. Strong tests have at least two key features. First, strong tests entail the prediction of observations that, absent the theory, we would not otherwise expect. For example, the panic disorder model we have discussed throughout this chapter predicts that the time to recover from an induction of arousal-related bodily sensations should indicate vulnerability to panic attacks and, thus, should prospectively predict the onset of panic attacks (for details, see Robinaugh, Haslbeck, et al., 2019). To our knowledge, this is not a prediction that has arisen in the context of any other theory and has never been tested. A study testing and finding support for this prediction would lend more credence to this theory than a study testing a prediction we would otherwise expect (e.g., that recurrent panic attacks will be correlated with avoidance behavior).

Second, strong tests entail precise predictions. That is, a prediction that goes beyond merely positing a refutation of the null hypothesis (e.g., a statistically significant association) or even a directional prediction (e.g., a positive association) to instead make precise point predictions about what should be observed. For example, a very well-developed theory of panic disorder would be able to predict the precise value of perceived threat (or interval of perceived threat values) which are likely to result from a particular arousal-inducing manipulation (e.g., by breathing CO<sub>2</sub> enriched air; Roberson-Nay et al., 2017). In other words, just as in the theory development stage, the theory testing stage calls for us to deduce the precise data models implied by our theory and to compare those implied data models to empirical data models. There is, thus, a fine distinction between theory testing and theory development in this framework. In the theory development stage, these comparisons are carried out in the spirit of improving upon and refining the theory. In the theory testing stage, these comparisons are carried out with the aim of subjecting the theory to refutation. A discrepancy between theory-implied and empirical data models in the development stage calls for refinement of the theory. A discrepancy between the theory-implied and em-

irical data models in the testing stage calls for the theorist to deeply consider the appropriateness of the theory.

Importantly, we are not proposing a “naive falsificationism” approach to theory testing in which a failed test requires abandonment of the theory (Meehl, 1990a). A discrepancy between theory implied and empirical data models can provide an opportunity to improve upon a theory, returning us to the stage of theory development. Nonetheless, a risky test should entail risk and repeated failures at this stage should push the theorist toward the generation of a new competing theory. For that reason, we believe that these risky tests should be engaged in only when the theorist is sufficiently confident in the theory that they would be willing to stake its survival on the outcome of the test. To that end, we would make several recommendations. First, as we have stressed throughout this chapter, formal theories will strengthen confidence in the predictions being tested by ensuring that they have been correctly deduced. Indeed, the level of specificity required for predictions to constitute a strong test of the theory all but requires that the theory be formalized. Second, the stage of theory development should be used to not only improve upon the theory, but also the assumptions about the instruments, measurements, and analyses that may also be responsible for any discrepancies between the theory-implied and empirical data models. The approach we have argued for in producing implied data models is helpful in this regard, as it forces the theory to formalize not only the theory, but also the measurement of variables; what we have termed “emulated measurement” (see Section 11.3.3). Strengthening confidence in these “emulated measurements” will strengthen the test of the theory, as such issues cannot so readily be blamed should the test fail. Third, research at this stage must be confirmatory in the strictest sense of the term (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). These studies should be preregistered, ideally with model simulations showing the precise theory, measurement, and analysis that will be used in the study.

If a theory fails a strong test, the decision of how to proceed depends upon what Meehl referred to as the “money in the bank” principle (Meehl, 1990a): If a theory has a track record of success, it would be unwise to discard the theory in the face of a single, or even several, failed tests. For Meehl, money in the bank was accumulated by passing risky tests. A theory that has passed many such tests should be retained more readily than a theory with no such record. We would argue for a broader conceptualization that draws on a wider range of criteria for theory appraisal, with particular emphasis on explanatory breadth. A formal theory that can explain a range of phenomena should be retained more readily than a theory that accounts for only a narrow set of phenomena. Nonetheless, we believe that any failure of a strong test should be taken as a serious challenge to the theory that, at a minimum, warrants careful consideration about how to proceed.

If a theory passes a strong test, it is corroborated, with the strength of corroboration proportional to the strength of the test. Notably, because strong tests all but require the evaluation of predictions made by the theory, a theory that has passed several such tests will have demonstrated a strong capacity for support-



ing prediction. Accordingly, a theory that has moved from generation, through refinement, and testing will emerge well equipped to support not only the explanation, but also the prediction and control of mental disorders.

## 11.5 Conclusions

In this chapter, we have argued that psychiatry needs formal theories and we have examined how data models can best inform the development of such theories. We focused especially on the network approach to psychopathology and considered three possible routes by which conditional dependence networks may inform formal theories about how mental disorders operate as complex systems. We found that these data models were not themselves capable of representing the structure we presume will be needed for a theory of mental disorders. Perhaps more surprisingly, we also found that we were unable to draw clear and reliable inferences from data models about the underlying system. Together, these findings suggest that merely gathering data models alone is unlikely to readily inform a well-developed formal theory. Instead, we found that the most promising use of empirical data models for theory development was to compare them to “implied data models” derived from an initial formal theory. In this approach, formal theories play an active role in their own development, with initial formalized theories being refined over time through ongoing comparison of implied and empirical data models.

Importantly, our analysis is not a critique of the specific data models we examined here, nor is it a dismissal of their value. Quite the opposite. We believe these data models provide rich and valuable information about the relationships among components of a system. However, our analysis strongly suggests that the network approach to psychopathology cannot survive on these data models alone. Formal theory is needed if the network approach is to move toward the explanation, prediction, and control of mental disorders. Indeed, there is growing recognition that formal theories are needed if we are to avoid problems associated with conflicting empirical results (i.e., the “Replication Crisis”; Collaboration et al., 2015) and move toward an accumulation of knowledge in scientific research (e.g., Muthukrishna & Henrich, 2019; Szollosi & Donkin, 2019; Yarkoni, 2019; Ioannidis, 2014). Accordingly, as a field psychiatry must grapple not only with methods for the collection and analysis of data, but also methods for the generation and development of formal theories

The research framework we proposed in Section 11.4 is intended to be a first step toward such a method of theory construction. At the heart of this approach is the use of formalized initial theories to start a cycle of theory development in which empirical data informs ongoing theory development and these improved theories inform subsequent empirical research. Critically, this research framework is not intended to suggest that all researchers must develop expertise in computational modeling. Data and the detection of robust empirical phenomena are central to psychiatric research in our proposed framework. However, our framework does suggest that, as a field, psychiatry must do more to develop ex-

expertise in computational modeling within its ranks. We suspect that it will only be through ongoing collaboration among theorists and empirical researchers that we will be able to leverage the empirical literature to produce genuine advances in our ability to explain, predict, and control psychopathology.

## **Acknowledgements**

We would like to thank Tessa Blanken, Fabian Dablander, Nathan Evans, Madelyn Frumkin, Ellen Hamaker, Richard McNally, Alex Millner, and Dirk Wulff for their feedback on earlier versions of this chapter.

## **Part III**

# **Conclusion**



---

# DISCUSSION

---

In the first part of my dissertation I developed a number of data models to capture the multivariate dependencies between symptoms and other variables related to psychopathology. In the second part my goal was to convince the reader that fitting data models is not enough to explain, predict and control psychopathology. To achieve these goals we need to construct formal theories. Of course, data and data models are critical to constructing such formal theories. I therefore concluded with a chapter on how to use data models to construct formal theories. However, the necessity of moving from data models to formal theories is not something I saw clearly when I started working on my dissertation. In fact, this transition reflects how my own thinking developed during the four years of my PhD. In this last chapter I would like to retrace this path and then conclude by suggesting directions for future research that I believe would benefit the development of formal theories of psychopathology.

## 12.1 Data Models

When I started working on my dissertation in 2015, applied researchers who were working within the network approach to psychopathology were limited to only a few models: The Ising model for binary-valued variables, the multivariate Gaussian distribution for continuous variables, and the Vector Autoregressive (VAR) model for continuous time series data. The estimation routines implemented in the R-package *mgm*, which I developed, extended the range of available models considerably. In Chapter 2 I introduced Mixed Graphical Models (MGMs) which allow one to capture the dependency structure between variables defined on different domains, such as continuous or categorical. Such mixed data occur often in psychopathology research. For example, while symptom severity scores and psychological constructs are typically defined on a continuous or ordinal scale, variables relating to social context, work environment, or treatment are often (nominal) categorical. In addition, allowing one to model continuous and ordinal variables as categorical variables provides a way to detect non-linear interactions. Finally, I adapted the estimation routines for MGMs to mixed Vector autoregressive (mVAR) models, in which different types of variables predict each other over periods of time.

Statistical network models are typically reported using network visualizations. In these figures, the absolute values of parameters are represented by the width of edges, which are scaled relative to the largest parameter in the model. While these relative edge-widths lead to an optimal visual representation of the

relative sizes of parameters, they do not convey how well the variables predict each other on an absolute scale. To address this problem, in Chapter 3 I proposed to compute a nodewise predictability measure and suggested a way to incorporate it into network visualizations. In Chapter 4 I re-analyzed the then nascent literature of empirical network papers on psychopathology with a focus on predictability and discussed how predictability can be a theoretically interesting quantity that indicates the extent to which a system is self-sustained.

A crucial limitation of the statistical network models used in the network literature was that they only included interactions between pairs of variables. In other words, they cannot model moderation effects. This is a major limitation because moderation effects are very plausible in a contextualized field such as psychology. They are also of central interest in the network approach to psychopathology, as they may point to possible interventions at the edge-level (Borsboom, 2017). To detect such moderation effects in network models, in Chapter 5 I introduced Moderated Network Models (MNM), which allow one to model each pairwise interaction as a function of all other variables in the model. Since MNMs are implemented within the MGM framework, they allow one to fit a large variety of models. For example, one can have pairwise interactions between continuous and categorical variables, which are simultaneously moderated by both continuous and categorical variables. This implies that an MNM with a single categorical moderator provides an alternative way to discover differences in network models across groups, which does not require multiple steps such as resampling and significance testing.

The central idea of the network approach to psychopathology is that mental disorders arise from causal interactions between symptoms. This suggests that the interactions between symptoms in healthy and unhealthy individuals are different, and that those interactions change within an individual when they transition between sustained healthy and unhealthy states. To detect such changes in individual time series, in Chapter 6 I introduced a method to estimate time-varying VAR models. In addition, I mapped out in an extensive simulation study to which extent time-varying parameters can be detected in realistic data. While this chapter focused on VAR models for simplicity, the R-package *mgm* also implements the estimation of time-varying mVAR models and MGMs. Time-varying models are crucial for answering a variety of research questions. They allow one to detect changes in the structure of interactions in observational studies and explain those changes using additional variables. Another application would be to monitor patients and use the time-varying models as multivariate Early Warning Signals (EWS; Scheffer et al., 2009), which could pinpoint periods in which treatment is most effective (Olthof et al., 2019). Finally, time-varying models can be used to study how the structure of interactions change in response to treatments (Wichers et al., 2016).

While Chapters 2 - 6 were concerned with making new data models available to applied researchers, the last two chapters on data models tackled more specific methodological problems. In Chapter 7 I discussed the topics of bias-variance trade-off and model selection in the context of choosing between the VAR model and a special case of the VAR model, the AR model, which only includes autore-

gressive effects. Next to discussing these theoretical issues, this chapter included a simulation study which indicated how many observations are necessary for the VAR model to outperform the AR model, when simulating from different types of VAR models. Finally, in Chapter 8 I discussed how the interpretation and dynamics of the Ising model change if one switches the domain of the binary variables from  $\{0, 1\}$  to  $\{-1, 1\}$ . Amongst other results, this chapter showed that the common belief that dense networks lead to elevated symptom levels is not generally true, but that this result is contingent on the characteristics of the model at hand.

## 12.2 From Data Models to Formal Theories

While at the beginning of my PhD I focused on statistical problems, the constant exposure to psychopathology research in the Psychosystems group had its effect on me and I could not help but become interested in the subject matter myself. This developing interest led me to think about whether the data models I have been working on are actually good models for mental disorders. Key phenomena of psychopathology include that individuals can transition between states with different symptom patterns (e.g., leading to diagnosis / no diagnosis), that underlying processes evolve at different time scales (e.g., momentary experience vs. learning), and that causes are to be found at many levels (e.g., symptom level, cognitive-behavioral level, societal level). Accordingly, a good model for psychopathology should be able to capture these phenomena. However, the multivariate Gaussian distribution and the VAR model do not achieve this. And while the Ising model produces a surprising amount of interesting behavior given its simplicity, it is still insufficient to match the complexities of mental disorders (e.g., in its standard formulation it does not allow for different time scales). It became clear to me that while the data models I have been working on are crucial to capture data patterns and to evaluate their robustness, they do not allow one to model mental disorders at the level of detail that is most likely necessary to provide satisfying explanations and to develop efficacious treatments. This realization motivated the second part of my dissertation, which investigated the nature of models that represent mental disorders in sufficient detail — which I call formal theories — and how to use data and data models to construct them.

My initial reaction to the problem that current data models were insufficient to fully capture mental disorders was to use more complex data models. In Chapter 9 I set out to explore this possibility by simulating data from a simple but non-trivial model for psychopathology and evaluating how difficult it is to recover the model in different settings. The results showed that if the data are sampled at an extremely high sampling frequency, one is able to recover the true set of differential equations. This is the basic logic of statistical estimation and in no way surprising. However, when sampling at a rate that is typical for the Experience Sampling Method (ESM), the dependencies across time were largely lost and it was therefore impossible to recover the true model. While this result is contingent on how we set up our model, it illustrates that low sampling frequencies can render the recovery of certain processes impossible. And it is plausible that

this presents a problem in many applications. For example, it seems intuitive that emotion dynamics, which are defined on a time scale of seconds or minutes (Houben et al., 2015), are difficult to recover from ESM measurements taken at an interval of 90 minutes. The dynamics of mood, on the other hand, which are defined on a time scales of hours or days (Larsen, 2000) may well be feasible to study with ESM time series. Assuming that the sampling frequency is sufficient to recover the process of interest, one still needs to sample a long enough time series. However, the simulation results in Chapter 7 suggest that it is unrealistic to reliably estimate models that are more complex than the VAR model with the 100 observations or less of a typical ESM study.

Since it is difficult to estimate a more appropriate complex model from data, can we instead fit simpler models (like the VAR model) and make inferences about the underlying complex model? This is the second question I explored in Chapter 9. Such inferences would require knowledge of the mapping between the parameters in the estimated statistical model and the characteristics of the true model. However, since the true model is unknown, no exact mapping is available. I illustrated this problem by showing that even straightforward inferences such as “the effect between A and B is stronger than the effect between C and D in the statistical model, therefore the same has to be the case in the true model” do not need to be valid. This paints a sobering picture for the emerging literature that claims to be able to use statistical time series models to uncover the “complex dynamics” underlying mental disorders. While the goal in this literature is typically to obtain a model that appropriately captures the complexities of mental disorders, Chapter 9 suggests that it is currently unclear how such a model should be inferred from popular time series models such as the VAR model.

The first take away from Chapter 9 is that for most mental disorders it is unrealistic to fit a model directly from data that satisfies its complexities; the second one is that it is extremely difficult to make direct inferences from statistical models about the underlying system. Together, these findings suggest that the goal of obtaining formal theories for mental disorders cannot be reached with the framework of statistical estimation alone.

### 12.3 Formal Theories

I was very fortunate that at around the same time that I became interested in general formal modeling, Don Robinaugh visited our lab to construct a formal theory of panic disorder. I participated in this project by taking over some of its technical aspects, and therefore had the chance to be involved in most stages of theory development. To develop the theory, we employed an approach that has already been used by van der Maas et al. (2006) to create the mutualism model for intelligence and by Dalege et al. (2016) to develop the Causal Attitude Network (CAN) model for attitudes. Instead of starting out by fitting a data model to a specific data set, this approach begins by listing facts that are considered to be established in the literature. In the case of panic disorder, this involved list-



ing the components underlying the disorder and the relations between them, but also relatively simple empirical facts such as that panic attacks typically have a duration of 5-20 minutes. We first implemented the core dynamics of panic attacks by adapting a model from ecology such that it could produce realistic panic attacks, and then added a slower learning process which is necessary to explain panic disorder. Finally, the formal theory was tweaked to produce plausible predictions and to be consistent with basic empirical facts about panic disorder. The resulting formal theory has been presented in Chapter 10 of this dissertation. Participating in this project was extremely insightful because it provided me with experience in how to create a formal theory in a framework that is much more general than statistical estimation.

Working on the formal theory of panic disorder further strengthened my belief that fitting data models and interpreting their parameters alone is unlikely to produce a good theory for panic disorder and much less for other, typically more complex, mental disorders. What does this mean for the status of data models? Is the focus of the growing network literature on data models misguided? While it is indeed not entirely clear how to use data models to create formal theories, such a verdict seems unwarranted. Formal theories have to be grounded in empirical data, and therefore data models have to play an important role in their development. To wrap our heads around this question, Don Robinaugh, Oisín Ryan and myself spent a full month in Boston discussing this issue which eventually led to the final Chapter 11 of this dissertation.

In Chapter 11 we consider three different routes from data models to formal theories. First, treating data models as formal theories; second, drawing inferences from data models to create a formal theory; and third, using data models to develop theories with an abductive approach. As discussed above, the first route is unlikely to work because data models typically either do not match the complexity of mental disorders, or it is unrealistic to estimate them from data. The second route is problematic because it is generally unclear how to draw such inferences from data models to formal theories. We therefore advocate the third route, which uses data models in an abductive approach to constructing theories. In this route, data models are derived from competing formal theories, which are then used to evaluate which formal theory fits the data best. In addition, we provided an explicit description of the theory construction methodology used in Chapter 10 by laying out a four-step procedure for theory construction. This approach involves the steps of establishing the phenomenon, formulating, developing, and testing the formal theory, and thereby provides a general methodology to construct formal theories of mental disorders.

## 12.4 Future Directions

The abductive approach laid out in Chapter 11 provides in broad strokes a general methodology for theory construction. I conclude by suggesting three directions for future work that build on this approach.

### 12.4.1 Refining Theory Development

In Chapter 11 we suggested that formal theories should be developed by deriving implied data models, comparing them against empirical data models and using discrepancies to come up with changes to the original theory. Subsequently, one derives a data model from the altered theory and evaluates whether the empirical data are more likely given the original implied data model or the one derived from the adapted theory. However, this approach leaves many important questions unanswered.

For example, it remains unclear which data models should be derived (i.e., which predictions should be made) from the formal theory to optimally develop the theory. Should we always prefer a more complex model since it makes more specific predictions? Or should we derive predictions to test specific assumptions in the formal theory? And which of those assumptions should we test first? To answer questions related to how to optimally develop a formal theory requires specifying what makes a formal theory a good theory. Ultimately, we would hope to end up with formal theories which provide satisfying explanations, allow us to develop effective preventions and interventions, and neatly fit into existing theoretical frameworks. However, such judgments are often not possible during the early stages of theory development. For example, it will typically take many years before one can decide whether a theory led to the development of effective treatments. We therefore require more tangible ways to evaluate theories in early stages of development. An obvious way to evaluate the quality of a formal theory would be to evaluate how well it fits the data. However, in contrast to data models, a formal theory typically predicts many different data sets (e.g., different sets of variables, coupled with different measurement models, experimental vs. observational, within- vs. between-subjects data). This means that there is no straightforward way to compute a prediction error that can be used to compare the predictive adequacy of theories. Clearly, this is a problem that must be solved in any field that employs formal theories that are not data models. It is therefore likely that much can be learned from fields with a stronger modeling tradition such as biology or ecology.

The accuracy of predictions, however, is not the only dimension along which one can evaluate the quality of a formal theory. An alternative strategy, which we already used in the approach suggested in Chapter 11 and which is referred to as abduction (Peirce, 1931) or inference to the best explanation (Harman, 1965), evaluates theories based on their explanatory worth. While these approaches are necessarily qualitative, some operationalizations and concrete examples of applications do exist. For example, the theory of explanatory coherence (Thagard, 1989) provides a system of seven principles which allow to evaluate theories on the dimensions explanatory breadth, simplicity and analogy. Note that one's definition of explanation can also include predictive adequacy, which means that quantitative and qualitative measures can be considered together to evaluate a formal theory (cf., Haig, 2009). Since abductive approaches are based on explanation, and the nature of explanation itself is subject to active debate in the philosophy of science literature (Salmon, 1989; Woodward, 2014), they will al-

ways remain somewhat vague. Nonetheless, exploring the literature on theory evaluation seems a promising way to develop heuristics that productively steer theory development.

Another step in theory development that requires further specification is how to adapt a formal theory when its predictions do not line up with the empirical data. In Chapter 11 we only suggested to take the nature of the discrepancy between the predicted and empirical data models as a starting point and somehow come up with an adaptation to the formal theory that improves its prediction. Importantly, this process does not only concern tweaking parameters but can also involve changing the functional form of some relationship or even adopting a different overall architecture of the formal theory. Knowledge about the theory at hand and its dynamics will always be crucial to generating candidate adaptations. However, this process can probably be optimized by following a structured methodology. Again, a promising way forward would be to evaluate how disciplines with a strong formal modeling tradition such as biology or ecology deal with this problem. One solution from those fields that facilitates tweaking parameters is to fix a large number of parameters to simplify the formal theory sufficiently so that a likelihood function can be formulated over the remaining parameters. Then, at least a subset of parameters can be estimated directly from data. In some situations parameters can even be estimated when no likelihood can be defined by using Approximate Bayesian Computing (ABC; Csilléry, Blum, Gaggiotti, & François, 2010; Sunnåker et al., 2013). These methods also provide an avenue for adapting parameters to (groups of) individuals, which would improve predictions in clinical practice.

### 12.4.2 Formalizing Measurement Models

Every formal theory is based on a number of variables, such as arousal, mood, or avoidance. However, we never observe these variables directly, but we observe measurements of them instead. This can present a problem for drawing inferences about formal theory from data: If the measurement model used to compute the predictions from the formal theory is systematically different from the true measurement process, then differences in predicted data models and empirical data models could be due to this difference and not due to inadequacies in the formal theory. This problem has been recognized for a long time. For example, Meehl (1978) referred to such measurement models in the context of hypotheses tests as auxiliary hypotheses, which may render inferences from statistical hypothesis tests to theories invalid.

Formalizing measurement models does not immediately solve this problem. However, it addresses the problem that the measurement process often remains unexamined. Formalizing measurement models allows one to formulate the problem precisely and to determine what is known and unknown, or even knowable. For example, in Chapter 9 we simulated data from a dynamical system and showed that this particular system cannot be recovered with data collected at a sampling frequency of 90 minutes. In this instance the measurement function was extremely simple: We took a measurement every 90 minutes from the

continuous-time state variables. Such methodological studies are useful for planning empirical studies, because it allows one to get an idea about whether the process of interest can in principle be recovered from a certain type of data.

The most interesting questions about measurement, however, concern the response dynamics within an individual when being probed by an experiment or a questionnaire. While much of these response dynamics are unknown, at least rudimentary measurement models seem to be within reach. For example, different formulations of questions in ESM studies suggests different types of measurement functions: “What is your current level of anxiety?” refers to a snapshot of the variable anxiety, while “What was your level of anxiety since the last measurement?” will be a function of the anxiety values since the last measurement, such as its mean. However, using the ample psychological knowledge about memory and recall, we can probably do better than that. For example, one could compute a weighted mean using a realistic forgetting curve (Ebbinghaus, 1913/2013). Another example are shifting reference points for scales across time. For example, a mood rating of 4 on a 1-5 scale probably refers to a different mood level when mood has been very low in the preceding weeks than when mood would have been very high. Hasselman and Bosman (2020) suggest to address this issue by modeling the differences between responses and a time-varying mean instead of the raw responses. Finally, formalizing measurement models would help to open new lines of research into problems that are less well understood such as measurement reactivity (French & Sutton, 2010; Barta, Tennen, & Litt, 2012). In a distant future such efforts could lead to a measurement theory alike in physics, which, for instance, allows to take the temperature of the thermometer into account when measuring the temperature of a glass of water.

### 12.4.3 Dysfunction from Function

The network approach conceptualizes mental disorders as systems of mutually reinforcing symptoms (Borsboom, 2017). Consequently, most work within this framework has aimed at identifying possible causal links between symptoms, typically by analyzing the dependencies between symptoms with statistical network models (for reviews see Robinaugh, Hoekstra, Toner, & Borsboom, 2019; Contreras et al., 2019). However, in what follows I will point out limitations of the symptom level and argue that powerful theories of mental disorders are probably better developed at a level of cognitive, affective and behavioral variables.

A general problem of symptoms is the fact that they are typically aggregate variables. For example, take the following depression symptom: “*Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day*” (American Psychiatric Association, 2013). It seems intuitive that it is difficult to define mechanistic relationships with a variable that actually includes two variables, both codes their increase and decrease, and is additionally subject to two temporal qualifiers (“nearly every day” above and “for at least two weeks” from the overall definition of major depression). Another consequence of these aggregate variables is that posited relationships do not make very specific predictions:

For example, the causal effect between two binary symptoms can be captured in a simple  $2 \times 2$  table. Instead of modeling this symptom, one could model the two separate variables weight and appetite in an individual across time. This modeling choice both greatly simplifies the specification of mechanistic relations and makes much more precise predictions. Instead of predicting only that the above symptom correlates with, say, the symptom depressed mood, we can precisely specify how each variable evolves over time, the exact nature and lag of the effect of, say, depressed mood on appetite, the distribution of all variables, etc. Modeling on a cognitive-behavioral level therefore provides much more precise predictions, which makes it easier to develop and test a formal theory using empirical data.

While causal relations between symptoms can explain why an individual can get stuck in a state of elevated symptom activation (Borsboom, 2017), they do not explain why any single symptom appears initially, or what the exact mechanics of these direct causal effects between symptoms are. This means that analyzing the symptom level alone provides little insight into *how* to perform an intervention, either on symptoms directly or the causal relations between them. Efficacious interventions on symptoms typically require to deconstruct the symptom itself. For example, interventions such as sleep restriction therapy (SRT; Kyle et al., 2015) for insomnia are based on a careful investigation of the underlying variables related to sleep and the variety of problems associated with it. Similarly, any intervention on the causal relation between two symptoms requires us to deconstruct the causal relation into its cognitive, affective and behavioral parts and their interplay.

Another limitation of the symptom level is that it provides no answer to the question of why individuals differ in their causal relationships between symptoms. Even worse, it is not actually clear whether the causal relations *do* differ between healthy and diagnosed individuals. One reason for this lack of knowledge is that studying statistical relations between symptoms requires active symptoms, which are by definition hardly present in healthy individuals. A popular attempt to sidestep this problem is to analyze groups with different severity levels. However, selecting groups based on symptom sum scores can in some situations lead to biases which render results difficult to interpret (de Ron, Fried, & Episkamp, 2019). These limitations of the symptom level may be part of the explanation of why the evidence on the relationship between connectivity of symptom networks and symptom activation does not seem to converge (Robinaugh, Hoekstra, Toner, & Borsboom, 2019). However, being able to identify causal relations between symptoms before they actually appear would be crucial to determine an individual's resilience and to design any type of prevention. To sharpen the concept of resilience and to generate ideas for prevention and intervention, the best way forward is probably to deconstruct causal relations between symptoms into cognitive, affective and behavioral variables and their interrelations. This provides us with a more flexible framework to identify the nature of the causal relation, both because we can describe them in much more detail and because the non-symptom variables are more likely than symptoms to show variation in healthy individuals.

Formulating theories of mental disorders at a level of cognitive, affective, and behavioral variables opens up a more refined perspective on the causes of mental disorders on an individual level. Processes such as perception, memory, decision-making, emotion, mood or behaviors such as sleeping, exercising, working or engaging in social interaction are operating in any individual. In healthy individuals the system comprised of all these processes is operating in a functional way, thereby not giving rise to any (elevated levels of) symptoms. Consequently, in unhealthy individuals who experience symptoms the system has to be different in some dysfunctional way. This suggests that we can explain the presence of symptoms and the causal relations between them with one or several dysfunctional changes to an in principle functional system. The formal theory of panic disorder in Chapter 10 illustrates this perspective. This model consists of variables such as arousal, perceived threat, and avoidance, which arguably can be measured in any individual. However, the dynamics of the formal theory imply that if one experiences a panic attack, one will escape from the current situation, and consequently have a higher arousal schema (S), which is a composite variable describing to which extent arousal is perceived as a threat, and determines the probability of future panic attacks. The fact that panic attacks always lead to escape, and that the escape behavior always leads to a higher S-value are the dysfunctional parts of the system. If either of the two is fixed, according to our theory it is not possible anymore to develop panic disorder. This illustrates how explaining a mental disorder in terms of dysfunctional relations between cognitive and behavioral variables can readily provide ideas for tangible interventions. In Appendix G.3 of Chapter 10 we expanded on this idea by deriving a treatment plan with several interventions from our theory.

Formal theories which explain mental disorders with dysfunctional systems of cognitive, affective and behavioral variables allow one to also include a functional state of the system. Indeed, this is necessary if the formal theory is supposed to be able to predict successful treatment outcomes. Including both dysfunctional and functional states allows us to remove dysfunctions from the system and demonstrate how the change impacts the presence of symptoms. In other words, the formal theory allows us to simulate interventions. The fact that a formal theory should include a functional state suggests that formal theories of mental disorders can be constructed based on models of functional cognition, affect and behavior. This possibly allows one to leverage the vast body of research in cognitive psychology on processes such as perception, memory, learning, belief formation, and decision-making, all of which are involved in, or even at the core of all mental disorders. For example, body dysmorphic disorder is clearly related to perception, manic episodes impede decision-making, and delusions and paranoid ideation are related to problems with the formation of beliefs. Fortunately, cognitive and mathematical psychology have a strong tradition in formalization, which means that formalizing theories of mental disorders can build on a wealth of formalized models of cognition and behavior. However, many formalized theories of relevant functional systems exist also outside the core areas of cognitive psychology. For example, there is a rich literature of well-developed formal theories of sleep (Borbely & Achermann, 1992), there is work that models emotional

processes in terms of inferences about interoceptive bodily states (Seth, Suzuki, & Critchley, 2012), and there are formal theories of mood that are rooted in evolutionary theory and game theory (Nettle, 2009; Eldar, Rutledge, Dolan, & Niv, 2016).

I have argued that we should model mental disorders as dysfunctional versions of typically functional systems on a cognitive, affective and behavioral level. However, explanations for mental disorders also exist at other levels. For example, the field of computational psychiatry uses formal theories of brain function to characterize the mechanisms of psychopathology (Wang & Krystal, 2014; Stephan & Mathys, 2014; Friston, Stephan, Montague, & Dolan, 2014; Friston, Redish, & Gordon, 2017). This field is especially interesting due to its strong tradition in formal modeling. Certainly, theories at the cognitive-behavioral level can be informed by theories at the synaptic or circuit level in the brain and vice versa. In addition to the cognitive and brain level, there are also social, economic and cultural variables that are related to mental disorders (e.g., Kendler, 2008). Ultimately, formal theories on all of these levels should be integrated in order to obtain an explanation of mental disorders that does justice to their complexity.

## 12.5 Conclusions

This dissertation dealt with the problem of modeling psychopathology. Its first part considerably extended the range of data models that are available to applied researchers by introducing and implementing estimation routines for Mixed Graphical Models (MGMs), mixed VAR (mVAR) models, Moderated Network Models (MNM), nodewise predictability and time-varying models. The second part aimed to shift the focus of our field towards developing formal theories as opposed to only fitting data models. It included a methodological study of the problems of undersampling and misspecification in recovering systems from data, and presented a formal theory of panic disorder. The last chapter provided a general analysis of the relationship between data models and formal theories and put forward a general approach for how to construct formal theories.

In this concluding chapter I outlined three directions for future research. First, the general approach to theory construction described in the last chapter should be worked out in much more detail, for example answering the questions of how to optimally derive predictions from theories, and how to adapt theories in case their predictions fail. This requires developing new methodology or adapting existing ones from fields with a strong formal modeling tradition. Second, the measurement process should be formalized next to the theory itself in order to increase the validity of inferences drawn from the data about formal theories. Finally, I proposed to create formal theories about mental disorders at the level of cognitive, affective, and behavioral variables. This shift in the level of analysis may lead to a better understanding of mental disorders, facilitate the discovery of intervention targets, allow for a sharper definition of resilience and vulnerability, and allow integration with neighboring fields such as cognitive psychology and neuroscience.





**Part IV**

**Appendices**



## References

- Aalen, O. O., Røysland, K., Gran, J. M., & Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(4), 831–861.
- aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review*, *32*(6), 510–523. doi: 10.1016/j.cpr.2012.05.007
- Abegaz, F., & Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, *14*(3), 586–599. doi: 10.1093/biostatistics/kxt005
- Afshartous, D., & Preston, R. A. (2011). Key results of interaction models with centering. *Journal of Statistics Education*, *19*(3). doi: 10.1080/10691898.2011.11889620
- Afzali, M. H., Sunderland, M., Batterham, P. J., Carragher, N., Calear, A., & Slade, T. (2017). Network approach to the symptom-level association between alcohol use disorder and posttraumatic stress disorder. *Social Psychiatry and Psychiatric Epidemiology*, *52*(3), 329–339.
- Agresti, A. (2003). *Categorical data analysis* (Vol. 482). John Wiley & Sons.
- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Sage. doi: 10.1057/jors.1994.16
- Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47.
- Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (2007). *Collaborative psychiatric epidemiology surveys (cpes), 2001-2003 [united states]*. Inter-university Consortium for Political and Social Research.
- American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders (dsm-ii)*. American Psychiatric Pub.
- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (dsm-iii)*. American Psychiatric Pub.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (dsm-iv)*. American Psychiatric Pub.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed. ed.). Washington, DC: Autor.
- Andersen, R. (2009). Nonparametric methods for modeling nonlinearity in regression analysis. *Annual Review of Sociology*, *35*, 67–85. doi: 10.1146/annurev.soc.34.040507.134631
- Anderson, G. M., Montazeri, F., & de Bildt, A. (2015). Network approach to autistic traits: group and subgroup analyses of ados item scores. *Journal of Autism and Developmental Disorders*, *45*(10), 3115–3132.
- Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., ... White, K. P. (2002). Gene expression during the life cycle of *drosophila melanogaster*. *Science*, *297*(5590), 2270–2275.
- Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of dsm-5 posttraumatic stress disorder symptoms and corre-

- 
- lates in us military veterans. *Journal of Anxiety Disorders*, 45, 49–59.
- Asmundson, G. J., & Asmundson, A. J. (2018). Are anxiety disorders publications continuing on a trajectory of growth? a look at boschen's (2008) predictions and beyond. *Journal of Anxiety Disorders*, 56, 1–4.
- Asmundson, G. J., Taylor, S., & Smits, J. A. J. (2014). Panic disorder and agoraphobia: An overview and commentary on dsm-5 changes. *Depression and Anxiety*, 31(6), 480–486.
- Association, A. P., & for Mental Hygiene. Bureau of statistics, N. C. (1918). *Statistical manual for the use of institutions for the insane*. National committee for mental hygiene.
- Atkinson, K. E. (2008). *An introduction to numerical analysis*. John Wiley & Sons.
- Bailer-Jones, D. M. (2009). *Scientific models in philosophy of science*. University of Pittsburgh Press.
- Bak, M., Drukker, M., Hasmi, L., & van Os, J. (2016). An n= 1 clinical network analysis of symptoms and treatment in psychosis. *PLoS ONE*, 11(9), e0162811. doi: 10.1371/journal.pone.0162811
- Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9, 485–516. Retrieved 2015-06-08, from <http://dl.acm.org/citation.cfm?id=1390696>
- Barabási, A.-L. (2012). The network takeover. *Nature Physics*, 8(1), 14–16.
- Barlow, D. H. (1997). Cognitive-behavioral therapy for panic disorder: current status. *The Journal of Clinical Psychiatry*.
- Barlow, D. H., & Craske, M. G. (1988). The phenomenology of panic. In R. S & M. J. D (Eds.), *Panic: Psychological perspectives* (p. 11-35).
- Barlow, D. H., Ellard, K., Fairholme, C., Farchione, C., Boisseau, C., Allen, L., & Ehrenreich-May, J. (2011). The unified protocol for transdiagnostic treatment of emotional disorders: Client workbook.
- Barlow, D. H., Farchione, T. J., Bullis, J. R., Gallagher, M. W., Murray-Latin, H., Sauer-Zavala, S., ... others (2017). The unified protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: A randomized clinical trial. *JAMA Psychiatry*, 74(9), 875–884.
- Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (p. 108-123). The Guilford Press.
- Bates, D., & Maechler, M. (2017). Matrix: Sparse and dense matrix classes and methods [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=Matrix> (R package version 1.2-8)
- Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M., ... Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, 46(16), 3359–3369.
- Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.
- Beck, A. T. (1988). Cognitive approaches to panic disorder: theory and therapy. In R. S & M. J. D (Eds.), *Panic: Psychological perspectives* (p. 11-35).

- Begeer, S., Wierda, M., & Venderbosch, S. (2013). Allemaal Autisme, Allemaal Anders. Rapport NVA enquête 2013 [All Autism, All Different. Dutch Autism Society Survey 2013]. *Bilthoven: NVA.*, 83.
- Belsley, D. A., & Kuti, E. (1973). Time-varying parameter structures: An overview. In *Annals of economic and social measurement, volume 2, number 4* (pp. 375–379). NBER.
- Ben-Zeev, D., & Young, M. A. (2010). Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. *The Journal of Nervous and Mental disease, 198*(4), 280–285.
- Berrios, G. E. (1996). *The history of mental symptoms: descriptive psychopathology since the nineteenth century*. Cambridge University Press.
- Berrios, G. E. (1999). *Journal of affective disorders 56* (1999) 245–247 [www.elsevier.com/locate/jad](http://www.elsevier.com/locate/jad). *Journal of Affective Disorders, 56*, 245–247.
- Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics, 41*(3), 1111. doi: 10.1214/13-AOS1096
- Blevins, S. M., & Bronze, M. S. (2010). Robert Koch and the 'golden age' of bacteriology. *International Journal of Infectious Diseases, 14*(9), e744–e751.
- Bogen, J., & Woodward, J. F. (1988). Saving the phenomena. *The Philosophical Review, 97*(3), 303–352.
- Boker, S. M., Deboeck, P. R., Edler, C., & Keel, P. K. (2010). Generalized local linear approximation of derivatives from time series. In *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (pp. 161–178). Routledge.
- Bongers, S., & Mooij, J. M. (2018). From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*.
- Borbely, A. A., & Achermann, P. (1992). Concepts and models of sleep regulation: an overview. *Journal of sleep research, 1*(2), 63–79.
- Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64*(9), 1089–1108.
- Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*(1), 5–13.
- Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology, 9*, 91–121.
- Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L., & Cramer, A. O. (2017). False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017).
- Borsboom, D., Rhemtulla, M., Cramer, A. O. J., van der Maas, H. L., Scheffer, M., & Dolan, C. (2016). Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine, 46*(8), 1567–1579.
- Bos, E. H., & De Jonge, P. (2014). "Critical slowing down in depression" is a great idea that still needs empirical proof. *Proceedings of the National Academy of Sciences, 111*(10), E878–E878.
- Boschloo, L., Schoevers, R. A., van Borkulo, C. D., Borsboom, D., & Oldehinkel,

- 
- A. J. (2016). The network structure of psychopathology in a community sample of preadolescents. *Journal of Abnormal Psychology, 125*(4), 599.
- Boschloo, L., van Borkulo, C. D., Borsboom, D., & Schoevers, R. A. (2016). A prospective study on how symptoms in a network predict the onset of depression. *Psychotherapy and Psychosomatics, 85*(3), 183–184.
- Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D., & Schoevers, R. A. (2015). The network structure of symptoms of the diagnostic and statistical manual of mental disorders. *PLoS ONE, 10*(9), e0137621.
- Bouton, M. E., Mineka, S., & Barlow, D. H. (2001). A modern learning theory perspective on the etiology of panic disorder. *Psychological Review, 108*(1), 4.
- Breiman, L., et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.
- Bringmann, L. F., & Albers, C. J. (2019). Inspecting gradual and abrupt changes in emotion dynamics with the time-varying change point autoregressive model.
- Bringmann, L. F., Ferrer, E., Hamaker, E. L., Borsboom, D., & Tuerlinckx, F. (2018). Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model. *Multivariate Behavioral Research, 53*(3), 293–314. doi: 10.1080/00273171.2018.1439722
- Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods, 22*(3), 409.
- Bringmann, L. F., & Haslbeck, J. M. B. (2017). *tvvargam*. <https://github.com/LauraBringmann/tvvarGAM>. GitHub.
- Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., ... Tuerlinckx, F. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PLoS ONE, 8*(4), e60188.
- Brown, T. A., & Cash, T. F. (1990). The phenomenon of nonclinical panic: Parameters of panic, fear, and avoidance. *Journal of Anxiety Disorders, 4*(1), 15–29.
- Brush, S. G. (1967). History of the lenz-ising model. *Reviews of Modern Physics, 39*(4), 883.
- Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application, 1*, 255–278.
- Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media. doi: 10.1007/978-3-642-20192-9
- Buller, R., Maier, W., & Benkert, O. (1986). Clinical subtypes in panic disorder: Their descriptive and prospective validity. *Journal of Affective Disorders, 11*(2), 105–114.
- Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR (1) Based Models do not Always Outpredict AR(1) Models in Typical Psychological Applications. *Psychological Methods, 1*–17. doi: 10.1037/met0000178
- Burger, J., Robinaugh, D., Quax, R., Riese, H., Schoevers, R. A., Epskamp, S., et al.

- (2019). Bridging the gap between complexity science and clinical practice by formalizing idiographic theories: A computational model of functional analysis.
- Byrne, B. M. (2013). *Structural equation modeling with mplus: Basic concepts, applications, and programming*. Routledge.
- Bystritsky, A., Nierenberg, A., Feusner, J., & Rabinovich, M. (2012). Computational non-linear dynamical psychiatry: a new methodological paradigm for diagnosis and course of illness. *Journal of Psychiatric Research*, 46(4), 428–435.
- Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., Ito, T. A., et al. (2000). The psychophysiology of emotion. *Handbook of emotions*, 2, 173–191.
- Cameron, O. G., Lee, M. A., Curtis, G. C., & McCann, D. S. (1987). Endocrine and physiological changes during “spontaneous” panic attacks. *Psychoneuroendocrinology*, 12(5), 321–331.
- Cannon, W. B. (1916). *Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement*. New York: D. Appleton and Company.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Carr, D. S., Nesse, R. M., & Wortman, C. B. (2005). *Spousal bereavement in late life*. Springer Publishing Company.
- Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.
- Casas, I., & Fernandez-Casal, R. (2018). tvreg: Time-varying coefficients linear regression for single and multiple equations [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=tvReg> (R package version 0.3.0)
- Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3), 167–174.
- Casey, L. M., Oei, T. P., & Newcombe, P. A. (2004). An integrated cognitive model of panic disorder: The role of positive and negative cognitions. *Clinical Psychology Review*, 24(5), 529–555.
- Chambless, D. L., Caputo, G. C., Jasin, S. E., Gracely, E. J., & Williams, C. (1985). The mobility inventory for agoraphobia. *Behaviour Research and Therapy*, 23(1), 35–44.
- Chen, S., Witten, D. M., et al. (2015). Selection and estimation for mixed graphical models. *Biometrika*, 102(1), 47.
- Chen, S., Witten, D. M., & Shojaie, A. (2014). Selection and estimation for mixed graphical models. *Biometrika*, 102(1), 47–64. doi: 10.1093/biomet/asu051
- Chen, X., & He, Y. (2015). Inference of high-dimensional linear models with time-varying coefficients. *arXiv preprint arXiv:1506.03909*. Retrieved 2016-04-06, from <http://arxiv.org/abs/1506.03909>
- Chow, S.-M. (2019). Practical tools and guidelines for exploring and fitting linear and nonlinear dynamical systems models. *Multivariate Behavioral Research*, 1–29.

- 
- Chow, S.-M., Ou, L., Ciptadi, A., Prince, E. B., You, D., Hunter, M. D., ... Messinger, D. S. (2018). Representing sudden shifts in intensive dyadic interaction data using differential equation models with regime switching. *Psychometrika*, 83(2), 476–510.
- Clark, D. M. (1986). A cognitive approach to panic. *Behaviour Research and Therapy*, 24(4), 461–470.
- Clark, D. M. (1993). Cognitive mediation of panic attacks induced by biological challenge tests. *Advances in Behaviour Research and Therapy*, 15(1), 75–84.
- Clark, D. M. (1999). Anxiety disorders: Why they persist and how to treat them. *Behaviour Research and Therapy*, 37(1), S5-S27.
- Cohen, A. S., Barlow, D. H., & Blanchard, E. B. (1985). Psychophysiology of relaxation-associated panic attacks. *Journal of Abnormal Psychology*, 94(1), 96.
- Collaboration, O. S. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660. doi: 10.1177/1745691612462588
- Collaboration, O. S., et al. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Consortium, I. S., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256), 748.
- Contreras, A., Nieto, I., Valiente, C., Espinosa, R., & Vazquez, C. (2019). The study of psychopathology from the network analysis perspective: a systematic review. *Psychotherapy and psychosomatics*, 88(2), 71–83.
- Costantini, G., & Epskamp, S. (2017). Estimategroupnetwork: Perform the joint graphical lasso and selects tuning parameters [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=EstimateGroupNetwork> (R package version 0.1.2)
- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the art personality research: A tutorial on network analysis of personality data in r. *Journal of Research in Personality*, 54, 13–29. doi: 10.1016/j.jrp.2014.07.003
- Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2017). Stability and variability of personality networks. a tutorial on recent developments in network psychometrics. *Personality and Individual Differences*. doi: 10.1016/j.paid.2017.06.011
- Coste, J., & Granger, B. (2014). Mental disorders in ancient medical writings: Methods of characterization and application to french consultations (16th–18th centuries). In *Annales médico-psychologiques* (Vol. 172, pp. 625–633).
- Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38), E7900–E7909.
- Cramer, A. O. J., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, 26(4), 414–431.



- Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS One*, *11*(12), e0167490.
- Cramer, A. O. J., Waldorp, L. J., Van Der Maas, H. L., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and brain sciences*, *33*(2-3), 137–150.
- Craske, M. G., Kircanski, K., Epstein, A., Wittchen, H.-U., Pine, D. S., Lewis-Fernández, R., & Hinton, D. (2010). Panic disorder: a review of dsm-iv panic disorder and proposals for dsm-v. *Depression and Anxiety*, *27*(2), 93–112.
- Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418.
- Curtiss, J., & Klemanski, D. H. (2016). Taxonicity and network structure of generalized anxiety disorder and major depressive disorder: An admixture analysis and complex network analysis. *Journal of Affective Disorders*, *199*, 99–105.
- Dablander, F., Ryan, O., & Haslbeck, J. M. B. (2019). *Choosing between AR(1) and VAR(1) models in typical psychological applications*. PsyArXiv. Retrieved from psyarxiv.com/qgewy doi: 10.31234/osf.io/qgewy
- Dakos, V., & Lahti, L. (2013). R early warning signals toolbox. *The R Project for Statistical Computing*. Retrieved from Publication URL here (<http://cran.r-project.org/web/packages/earlywarnings/index.html>)
- Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The causal attitude network (can) model. *Psychological Review*, *123*(1), 2.
- Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 373–397. doi: 10.1111/rssb.12033
- Davis, G. (2008). *"The Cruel Madness of Love": Sex, Syphilis and Psychiatry in Scotland, 1880-1930* (Vol. 85). Rodopi.
- Deacon, B. J., Abramowitz, J. S., Woods, C. M., & Tolin, D. F. (2003). The anxiety sensitivity index-revised: psychometric properties and factor structure in two nonclinical samples. *Behaviour Research and Therapy*, *41*(12), 1427–1449.
- De Beurs, E., Garssen, B., Buikhuisen, M., Lange, A., Van Balkom, A., & Van Dyck, R. (1994). Continuous monitoring of panic. *Acta Psychiatrica Scandinavica*, *90*(1), 38–45.
- Dechambre, A. (1864). *Dictionnaire encyclopédique des sciences médicales* (Vol. 2). Asselin.
- De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2016). Get over it! a multilevel threshold autoregressive model for state-dependent affect regulation. *Psychometrika*, *81*(1), 217–241.
- de Ron, J., Fried, E. I., & Epskamp, S. (2019). Psychological networks in clinical populations: A tutorial on the consequences of berkson's bias.

- 
- Deserno, M. K., Borsboom, D., Begeer, S., & Geurts, H. M. (2017). Multicausal systems ask for multicausal approaches: A network perspective on subjective well-being in individuals with autism spectrum disorder. *Autism, 21*(8), 960–971.
- de Wild-Hartmann, J. A., Wichers, M., van Bemmelen, A. L., Derom, C., Thiery, E., Jacobs, N., ... Simons, C. J. (2013). Day-to-day associations between subjective sleep and affect in regard to future depression in a female population-based sample. *The British Journal of Psychiatry, 202*(6), 407–412.
- Didelez, V. (2007). Graphical models for composable finite markov processes. *Scandinavian Journal of Statistics, 34*(1), 169–185.
- Dretske, F. I. (1997). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software, 77*(5), 1–35. doi: 10.18637/jss.v077.i05
- Dujmić, Z., Machielse, E., & Treur, J. (2018). A temporal-causal modeling approach to the dynamics of a burnout and the role of physical exercise. In *Biologically inspired cognitive architectures meeting* (pp. 88–100).
- Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.
- Ebbinghaus, H. (1913/2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences, 20*(4), 155.
- Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science, 80*(5), 684–696.
- Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics* (pp. 569–593). Springer.
- Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119–127.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Ehlers, A., & Margraf, J. (1989). The psychophysiological model of panic attacks. *Fresh Perspectives on Anxiety Disorders, 1–29*.
- Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L., Engelhard, I., ... others (2017). Replicability and generalizability of ptsd networks: A cross-cultural multisite study of ptsd symptoms in four trauma patient samples. *Clinical Psychological Science*. doi: 10.17605/OSF.IO/2T7QP
- Einicke, G. A. (2019). Smoothing, filtering and prediction: Estimating the past, present and future second edition.
- Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in cognitive sciences, 20*(1), 15–24.
- Elliott-Graves, A. (2014). The role of target systems in scientific practice.
- Epskamp, S. (2015). Isingsampler: Sampling methods and distribution functions for the ising model [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=IsingSampler> (R package version 0.2)
- Epskamp, S. (2016). Brief Report on Estimating Regularized Gaussian Networks from Continuous and Ordinal Data. *ArXiv e-prints*.

- Epskamp, S. (2017). `graphicalvar`: Graphical var for experience sampling data [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=graphicalVAR> (R package version 0.2)
- Epskamp, S., Borsboom, D., & Fried, E. I. (2016). Estimating psychological networks and their stability: a tutorial paper. *arXiv preprint arXiv:1604.08462*.
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212.
- Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., Borsboom, D., et al. (2012). `qgraph`: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18.
- Epskamp, S., Deserno, M. K., & Bringmann, L. F. (2017). `mlvar`: Multi-level vector autoregression. Retrieved from <https://cran.r-project.org/web/packages/mlVAR/index.html>
- Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*.
- Epskamp, S., Maris, G., Waldorp, L., Borsboom, D., Irwing, P., Hughes, D., & Booth, T. (2016). Handbook of psychometrics. *Wiley, New York, NY*.
- Epskamp, S., Rhemtulla, M., & Borsboom, D. (2016). Generalized network psychometrics: Combining network and latent variable models. *arXiv preprint arXiv:1605.09288*.
- Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927.
- Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. J. (2018). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science*, *6*(3), 416–427.
- Epskamp, S., Waldorp, L. J., Möttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480.
- Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12.
- Fabio Di Narzo, A., Aznarte, J. L., & Stigler, M. (2009). `tsDyn`: Time series analysis based on dynamical systems theory [Computer software manual]. Retrieved from <https://cran.r-project.org/package=tsDyn/vignettes/tsDyn.pdf> (R package version 0.7)
- Fairchild, A. J., & McQuillin, S. D. (2010). Evaluating mediation and moderation effects in school psychology: A presentation of methods and review of current practice. *Journal of School Psychology*, *48*(1), 53–84. doi: 10.1016/j.jsp.2009.09.001
- Fan, J., & Gijbels, I. (1996). Applications of local polynomial modelling. In *Local polynomial modelling and its applications* (pp. 159–216). Springer.
- Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*.

- 
- tion, 96(456), 1348–1360.
- Fava, L., & Morton, J. (2009). Causal modeling of panic disorder theories. *Clinical Psychology Review, 29*(7), 623–637.
- Fechner, G. T., Howes, D. H., & Boring, E. G. (1966). *Elements of psychophysics* (Vol. 1). Holt, Rinehart and Winston New York.
- Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry, 26*(1), 57–63.
- Feldman, D. P. (2012). *Chaos and fractals: an elementary introduction*. Oxford University Press.
- Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 201711978*. doi: 10.1073/pnas.1711978115
- Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the Idiographic Dynamics of Mood and Anxiety via Network Analysis. *Journal of Abnormal Psychology, 126*(8), 1044–1056. doi: 0.1037/abn0000311
- Forré, P., & Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*.
- Forsyth, J. P., & Karekla, M. (2002). Biological challenge in the assessment of anxiety disorders. In *Practitioner's guide to empirically based measures of anxiety* (pp. 31–36). Springer.
- Foygel, R., & Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems* (pp. 604–612).
- Foygel, R., & Drton, M. (2014). High-dimensional Ising model selection with Bayesian information criteria. *arXiv preprint arXiv:1403.3374*. Retrieved 2014-07-20, from <http://arxiv.org/abs/1403.3374>
- Frances, A., Miele, G. M., Widiger, T. A., Pincus, H. A., Manning, D., & Davis, W. W. (1993). The classification of panic disorders: from freud to dsm-iv. *Journal of psychiatric research, 27*, 3–10.
- Freedman, H. I. (1980). *Deterministic mathematical models in population ecology* (Vol. 57). Marcel Dekker Incorporated.
- Freedman, R. R., Ianni, P., Ettetdgui, E., & Puthezhath, N. (1985). Ambulatory monitoring of panic disorder. *Archives of General Psychiatry, 42*(3), 244–248.
- French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: how much of a problem is it? what can be done about it? *British Journal of Health Psychology, 15*(3), 453–468.
- Freud, S. (1962). On the grounds for detaching a particular syndrome from neurasthenia under the description 'anxiety neurosis'. In *The standard edition of the complete psychological works of sigmund freud, volume iii (1893-1899): Early psycho-analytic publications* (pp. 85–115).
- Frewen, P. A., Schmittmann, V. D., Bringmann, L. F., & Borsboom, D. (2013). Per-

- ceived causal relations between anxiety, posttraumatic stress and depression: extension to moderation, mediation, and network analysis. *European Journal of Psychotraumatology*, 4(1), 20656.
- Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O. J., ... Stroebe, M. (2015). From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology*, 124(2), 256.
- Fried, E. I., & Cramer, A. O. J. (2016). *Moving forward: challenges and directions for psychopathological network theory and methodology*. Open Science Framework. Retrieved from [osf.io/bnekp](https://osf.io/bnekp) doi: 10.17605/OSF.IO/BNEKP
- Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, 12(6), 999–1020.
- Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? comparing the centrality of dsm and non-dsm symptoms of depression in a network analysis. *Journal of Affective Disorders*, 189, 314–320.
- Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, 31(3).
- Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the star\* d study. *Journal of Affective Disorders*, 172, 96–102.
- Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, 52(1), 1–10.
- Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time... or not? lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, 28(11), 1354.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008a). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441. doi: 10.1093/biostatistics/kxm045
- Friedman, J., Hastie, T., & Tibshirani, R. (2008b, July). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441. doi: 10.1093/biostatistics/kxm045
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1.
- Friedman, J., & Tibshirani, T. H. a. R. (2014, July). *glasso: Graphical lasso- estimation of Gaussian graphical models*. Retrieved 2015-09-16, from <https://cran.r-project.org/web/packages/glasso/index.html>
- Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1), 55–77.
- Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using bayesian net-

- 
- works to analyze expression data. *Journal of Computational Biology*, 7(3-4), 601–620.
- Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. *Computational Psychiatry*, 1, 2–23.
- Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014, July). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2), 148–158. Retrieved 2020-02-15, from <https://linkinghub.elsevier.com/retrieve/pii/S2215036614702755> doi: 10.1016/S2215-0366(14)70275-5
- Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11), 1129–1164.
- Fukano, T., & Gunji, Y.-P. (2012). Mathematical models of panic disorder. *Non-linear dynamics, psychology, and life sciences*, 16(4), 457–470.
- Gallagher, M. W., Payne, L. A., White, K. S., Shear, K. M., Woods, S. W., Gorman, J. M., & Barlow, D. H. (2013). Mechanisms of change in cognitive behavioral therapy for panic disorder: The unique effects of self-efficacy and anxiety sensitivity. *Behaviour Research and Therapy*, 51(11), 767–777.
- Gardner, C., & Kleinman, A. (2019). Medicine and the mind — the consequences of psychiatry’s identity crisis. *New England Journal of Medicine*, 381(18), 1697–1699. Retrieved from <https://doi.org/10.1056/NEJMp1910603> doi: 10.1056/NEJMp1910603
- Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*, 48(3), 432–435. doi: 10.1198/004017005000000661
- Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness Training Increases Momentary Positive Emotions and Reward Experience in Adults Vulnerable to Depression: A Randomized Controlled Trial. *Journal of Consulting and Clinical Psychology*, 79(5), 618–628. doi: 10.1037/a0024595
- Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., ... Horvath, S. (2006). Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. *PLoS Genetics*, 2(8), e130. Retrieved 2014-09-19, from <http://dx.plos.org/10.1371/journal.pgen.0020130> doi: 10.1371/journal.pgen.0020130
- Gibberd, A. J., & Nelson, J. D. (2015). Estimating dynamic graphical models from multivariate time-series data. *Proceedings of AALTD 2015*, 63.
- Gibberd, A. J., & Nelson, J. D. (2017). Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3), 623–634.
- Gibbert, A. (2017). *GraphTime*. <https://github.com/GlooperLabs/GraphTime>. GitHub.
- Glauber, R. J. (1963). Time-dependent statistics of the ising model. *Journal of Mathematical Physics*, 4(2), 294–307.
- Goekoop, R., & Goekoop, J. G. (2014). A network view on psychiatric disorders: network clusters of symptoms as elementary syndromes of psychopathology. *PLoS ONE*, 9(11), e112734.
- Goetz, R. R., Klein, D. F., Gully, R., Kahn, J., Liebowitz, M. R., Fyer, A. J., & Gor-

- man, J. M. (1993). Panic attacks during placebo procedures in the laboratory: physiology and symptomatology. *Archives of general psychiatry*, 50(4), 280–285.
- Goldstein, A. J., & Chambless, D. L. (1978). A reanalysis of agoraphobia. *Behavior Therapy*.
- Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 215–223. doi: 10.1080/00401706.1979.10489751
- Gorman, J., Liebowitz, M., Fyer, A., & Klein, D. (1987). Biological changes during lactate induced panic. In *International Journal of Neuroscience* (Vol. 32, pp. 781–781).
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.
- Griez, E., et al. (1983). Some remarks on the nosology of anxiety states and panic disorders. *Acta Psychiatrica Belgica*, 83(1), 33–42.
- Grimm, V., & Railsback, S. F. (2005). *Individual-based modeling and ecology*. Princeton university press.
- Groen, R. N., Snippe, E., Bringmann, L. F., Simons, C. J., Hartmann, J. A., Bos, E. H., & Wichers, M. (2019). Capturing the risk of persisting depressive symptoms: A dynamic network investigation of patients' daily symptom experiences. *Psychiatry Research*, 271, 640–648. doi: 10.1016/j.psychres.2018.12.054
- Grömping, U. (2012). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*.
- Grömping, U., et al. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, 17(1), 1–27.
- Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, 10(4), 371–388.
- Haig, B. D. (2008). Precis of 'an abductive theory of scientific method'. *Journal of Clinical Psychology*, 64(9), 1019–1022.
- Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *The American journal of psychology*, 219–234.
- Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. MIT Press.
- Hainmueller, J., Mummolo, J., & Xu, Y. (2018). How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis*. doi: 10.1017/pan.2018.46
- Hamaker, E. L. (2012). Why researchers should think “within-person”: A paradigmatic rationale. *Handbook of research methods for studying daily life*, 43, 61.
- Hamaker, E. L., & Grasman, R. (2012). Regime switching state-space models applied to psychological processes: Handling missing data and making inferences. *Psychometrika*, 77(2), 400–422.
- Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2010). Regime-switching

- 
- models to study psychological process. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (p. 155-168). Washington, DC: American Psychological Association.
- Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2016). Modeling bas dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment, 23*(4), 436–446.
- Hamaker, E. L., Zhang, Z., & van der Maas, H. L. (2009). Using threshold autoregressive models to study dyadic interactions. *Psychometrika, 74*(4), 727.
- Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine, 363*(4), 301–304. doi: 10.1056/NEJMp1006304
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society, 357*–384.
- Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton university press Princeton.
- Hamilton, J. D. (1995). Time series analysis. *Economic Theory, II, Princeton University Press, USA, 625*–630.
- Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review, 74*(1), 88–95.
- Hartmann, J. A., Wichers, M., Menne-Lothmann, C., Kramer, I., Viechtbauer, W., Peeters, F., ... others (2015). Experience sampling-based personalized feedback and positive affect: a randomized controlled trial in depressed patients. *PLoS ONE, 10*(6), e0128095. doi: 10.1371/journal.pone.0128095
- Haslbeck, J. M. B., Borsboom, D., & Waldorp, L. (2019). Moderated network models. *Multivariate Behavioral Research*.
- Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2020). A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research*.
- Haslbeck, J. M. B., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? a reanalysis of 18 published datasets. *Psychological Medicine, 47*(16), 2767–2776.
- Haslbeck, J. M. B., & Ryan, O. (2019). Recovering bistable systems from psychological time series. *Manuscript in preparation*.
- Haslbeck, J. M. B., & Waldorp, L. J. (2015). Structure estimation for mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.05677*.
- Haslbeck, J. M. B., & Waldorp, L. J. (2018). How well do network models predict observations? on the importance of predictability in network models. *Behavior Research Methods, 50*(2), 853–861. doi: 10.3758/s13428-017-0910-x
- Haslbeck, J. M. B., & Waldorp, L. J. (2020). mgm: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software, 93*(8), 1–46. doi: 10.18637/jss.v093.i08
- Hassan, J. (2008). *Capturing the dynamics of panic disorder. a system dynamics translation of the contemporary biological and psychological conceptualization of panic disorder*. Unpublished master's thesis, The University of Bergen.



- Hasselmann, F., & Bosman, A. (2020). Studying complex adaptive systems with internal states: A recurrence-based analysis strategy for multivariate time series data representing self-reports of human experience. *Frontiers in Applied Mathematics and Statistics-Recurrence Analysis of Complex Systems Dynamics*.
- Hasselmann, K. (1976). Stochastic climate models part i. theory. *Tellus*, 28(6), 473–485.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.
- Hek, K., Demirkan, A., Lahti, J., Terracciano, A., Teumer, A., Cornelis, M. C., ... others (2013). A genome-wide association study of depressive symptoms. *Biological Psychiatry*, 73(7), 667–678.
- Hibbert, G. A. (1984). Ideational components of anxiety: Their origin and content. *The British Journal of Psychiatry*, 144(6), 618–624.
- Higgs, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, 13(16), 508.
- Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3), 329–343. doi: 10.1093/imanum/22.3.329
- Hinton, D., So, V., Pollack, M. H., Pitman, R. K., & Orr, S. P. (2004). The psychophysiology of orthostatic panic in cambodian refugees attending a psychiatric clinic. *Journal of Psychopathology and Behavioral Assessment*, 26(1), 1–13.
- Hinton, D., Um, K., & Ba, P. (2001). Kyol goeu (‘wind overload’) part i: A cultural syndrome of orthostatic panic among khmer refugees. *Transcultural Psychiatry*, 38(4), 403–432.
- Hirsch, M. W., Smale, S., & Devaney, R. L. (2012). *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.
- Hjelmeland, H., & Loa Knizek, B. (2018). The emperor’s new clothes? a critical look at the interpersonal theory of suicide. *Death studies*, 1–11.
- Hoehn-Saric, R., McLeod, D. R., Funderburk, F., & Kowalski, P. (2004). Somatic symptoms and physiologic responses in generalized anxiety disorder and panic disorder: An ambulatory monitor study. *Archives of General Psychiatry*, 61(9), 913–921.
- Holling, C. S. (1973). Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4(1), 1–23.
- Holling, C. S. (1986). The resilience of terrestrial ecosystems: local surprise and global change. *Sustainable Development of the Biosphere*, 14, 292–317.
- Holling, C. S. (1996). Engineering resilience versus ecological resilience. *Engineering within ecological constraints*, 31(1996), 32.
- Holmes, E., Ward, E., & Wills, K. (2013). Marss: Multivariate autoregressive state-space modeling [Computer software manual]. Retrieved from <http://cran.r-project.org/web/packages/MARSS/> (R package version 3.9)

- 
- Holmes, E. E., Ward, E. J., & Wills, K. (2012). Marss: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal*, 4(1), 30.
- Hoorelbeke, K., Marchetti, I., De Schryver, M., & Koster, E. H. (2016). The interplay between cognitive risk and resilience factors in remitted depression: a network analysis. *Journal of Affective Disorders*, 195, 96–104.
- Hosenfeld, B., Bos, E. H., Wardenaar, K. J., Conradi, H. J., van der Maas, H. L., Visser, I., & de Jonge, P. (2015). Major depressive disorder as a nonlinear dynamic system: bimodality in the frequency distribution of depressive symptoms over time. *BMC Psychiatry*, 15(1), 222.
- Hosseinichimeh, N., Wittenborn, A. K., Rick, J., Jalali, M. S., & Rahmandad, H. (2018). Modeling and estimating the feedback mechanisms among depression, rumination, and stressors in adolescents. *PLoS ONE*, 13(9), e0204389.
- Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, 141(4), 901.
- Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., ... Reiman, E. (2010, April). Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3), 935–949. Retrieved 2016-04-13, from <http://www.sciencedirect.com/science/article/pii/S1053811909014281> doi: 10.1016/j.neuroimage.2009.12.120
- Hyland, M. E. (2011). *The origins of health and disease*. Cambridge University Press.
- Immer, A., & Gibberd, A. (2017). *GraphTime*. <https://github.com/GlooperLabs/GraphTime>. GitHub.
- Ioannidis, J. P. (2014). How to make more published research true. *PLOS Medicine*, 11(10). doi: e1001747
- Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1), 253–258.
- Jones, P. J., Mair, P., Riemann, B. C., Mugno, B. L., & McNally, R. J. (2018). A network perspective on comorbid depression in adolescents with obsessive-compulsive disorder. *Journal of Anxiety Disorders*, 53, 1–8.
- Kalisch, R., Cramer, A. O. J., Binder, H., Fritz, J., Leertouwer, I., Lunansky, G., ... Van Harmelen, A.-L. (2019). Deconstructing and reconstructing resilience: a dynamic network approach. *Perspectives on Psychological Science*, 14(5), 765–777.
- Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12), 1174.
- Keele, L. J. (2008). *Semiparametric regression for the social sciences*. Chichester, England: John Wiley & Sons.
- Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, 2(3-4), 160–165.
- Kendler, K. S. (2005). “a gene for...”: the nature of gene action in psychiatric disorders. *American Journal of Psychiatry*, 162(7), 1243–1252.
- Kendler, K. S. (2008). Explanatory models for psychiatric illness. *American Journal of Psychiatry*, 165(6), 695–702.

- Kendler, K. S. (2012). Levels of explanation in psychiatric and substance use disorders: implications for the development of an etiologically based nosology. *Molecular Psychiatry*, 17(1), 11.
- Kendler, K. S. (2017). Progressive validation of psychiatric syndromes: The example of panic disorder. *Philosophical Issues in Psychiatry IV: Psychiatric Nosology*, 314–331.
- Kendler, K. S. (2019). From many to one to many—the search for causes of psychiatric illness. *JAMA Psychiatry*.
- Kendler, K. S., Aggen, S. H., Flint, J., Borsboom, D., & Fried, E. I. (2018). The centrality of dsm and non-dsm depressive symptoms in han chinese women with major depression. *Journal of Affective Disorders*, 227, 739–744. doi: 10.1016/j.jad.2017.11.032
- Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, 41(6), 1143–1150.
- Kessler, R. C., Chiu, W. T., Jin, R., Ruscio, A. M., Shear, K., & Walters, E. E. (2006). The epidemiology of panic attacks, panic disorder, and agoraphobia in the national comorbidity survey replication. *Archives of General Psychiatry*, 63(4), 415–424.
- Kim, J., Zhu, W., Chang, L., Bentler, P. M., & Ernst, T. (2007). Unified structural equation modeling approach for the analysis of multisubject, multivariate functional mri data. *Human Brain Mapping*, 28(2), 85–93.
- Kim, N. S., & Ahn, W. (2002). The influence of naive causal theories on lay concepts of mental illness. *American Journal of Psychology*, 115(1), 33–66.
- King, A. A., Nguyen, D., & Ionides, E. L. (2015). Statistical inference for partially observed markov processes via the r package pomp. *arXiv preprint arXiv:1509.00503*.
- Klein, D. F. (1993). False suffocation alarms, spontaneous panics, and related conditions: an integrative hypothesis. *Archives of General Psychiatry*, 50(4), 306–317.
- Klein, D. F., & Klein, H. M. (1989). The definition and psychopharmacology of spontaneous panic and phobia.
- Koenders, M., De Kleijn, R., Giltay, E., Elzinga, B., Spinhoven, P., & Spijker, A. (2015). A network approach to bipolar symptomatology in patients with different course types. *PLoS ONE*, 10(10), e0141420.
- Kolar, M., Song, L., Ahmed, A., & Xing, E. P. (2010, March). Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1), 94–123. Retrieved 2016-04-06, from <http://projecteuclid.org/euclid.aos/1273584449> doi: 10.1214/09-AOAS308
- Kolar, M., & Xing, E. P. (2009). Sparsistent estimation of time-varying discrete Markov random fields. *arXiv preprint arXiv:0907.2337*. Retrieved 2016-04-06, from <http://arxiv.org/abs/0907.2337>
- Kolar, M., & Xing, E. P. (2012). Estimating networks with jumps. *Electronic Journal of Statistics*, 6(0), 2069–2106. Retrieved 2016-04-06, from <http://projecteuclid.org/euclid.ejs/1351865118> doi: 10.1214/12-EJS739

- 
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.
- Kossakowski, J. J., Epskamp, S., Kieffer, J. M., van Borkulo, C. D., Rhemtulla, M., & Borsboom, D. (2016). The application of a network approach to health-related quality of life (hrqol): introducing a new method for assessing hrqol in healthy adults and cancer patients. *Quality of Life Research*, 25(4), 781–792.
- Kossakowski, J. J., Groot, P., Haslbeck, J. M. B., Borsboom, D., & Wichers, M. (2017). Data from ‘critical slowing down as a personalized early warning signal for depression’. *Journal of Open Psychology Data*, 5(1). doi: 10.5334/jopd.29
- Kraepelin, E. (1913). *General paresis* (No. 14). Journal of Nervous and Mental Disease Publishing Company.
- Kramer, I., Simons, C. J., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., ... others (2014). A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry*, 13(1), 68–77. doi: 10.1002/wps.20090
- Kroeze, R., Van Veen, D., Servaas, M. N., Bastiaansen, J. A., Oude Voshaar, R., Borsboom, D., & Riese, H. (2016). Personalized feedback on symptom dynamics of psychopathology: a proof-of-principle study. *J Person-Orient Res*. doi: 10.17505/jpor.2017.01
- Krueger, F. (2015). bvarsv: Bayesian analysis of a vector autoregressive model with stochastic volatility and time-varying parameters [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=bvarsv> (R package version 1.1)
- Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. *Arguing About Science*, 74–86.
- Kuiper, R. M., & Ryan, O. (2018). Drawing conclusions from cross-lagged relationships: Re-considering the role of the time-interval. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5), 809–823.
- Kulikov, G. Y., & Kulikova, M. V. (2013). Accurate numerical implementation of the continuous-discrete extended kalman filter. *IEEE Transactions on Automatic Control*, 59(1), 273–279.
- Kyle, S. D., Aquino, M. R. J., Miller, C. B., Henry, A. L., Crawford, M. R., Espie, C. A., & Spielman, A. J. (2015). Towards standardisation and improved understanding of sleep restriction therapy for insomnia disorder: a systematic examination of cbt-i trial content. *Sleep Medicine Reviews*, 23, 83–88.
- Lader, M. (1991). Bio-psycho-social interactions in anxiety and panic disorders: a speculative perspective. *Irish Journal of Psychological Medicine*, 8(2), 154–159.
- Lader, M., & Mathews, A. (1970). Physiological changes during spontaneous panic attacks. *Journal of Psychosomatic Research*, 14(4), 377–382.
- Lakatos, I. (1970). Falsificationism and the methodology of scientific research programs’ in i. lakatos and a. musgrave. *Criticism and the growth of knowledge*, 91–196.

- Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?* (pp. 205–259). Springer.
- Larsen, R. J. (2000). Toward a science of mood regulation. *Psychological Inquiry*, 11(3), 129–141.
- Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.
- Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 1350–1360.
- Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.
- Lewis-Fernández, R., Hinton, D. E., Laria, A. J., Patterson, E. H., Hofmann, S. G., Craske, M. G., ... Liao, B. (2011). Culture and the anxiety disorders: recommendations for dsm-v. *Focus*, 9(3), 351–368.
- Ley, R. (1985). Blood, breath, and fears: A hyperventilation theory of panic attacks and agoraphobia. *Clinical Psychology Review*, 5(4), 271–285.
- Liebowitz, M. R., Fyer, A. J., Gorman, J. M., Dillon, D., Appleby, I. L., Levy, G., ... others (1984). Lactate provocation of panic attacks: I. clinical and behavioral findings. *Archives of General Psychiatry*, 41(8), 764–770.
- Lindquist, K. A., & Barrett, L. F. (2008). Constructing emotion: The experience of fear as a conceptual act. *Psychological Science*, 19(9), 898–903.
- Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct), 2295–2328.
- Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems* (pp. 1432–1440).
- Loh, P.-L., & Wainwright, M. J. (2012). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In *Advances in neural information processing systems* (pp. 2087–2095).
- Ludwig, D., Jones, D. D., Holling, C. S., et al. (1978). Qualitative analysis of insect outbreak systems: the spruce budworm and forest. *Journal of Animal Ecology*, 47(1), 315–332.
- Lunansky, G., van Borkulo, C. D., & Borsboom, D. (2019). Personality, resilience, and psychopathology: A model for the interaction between slow and fast network processes in the context of mental health.
- Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C., ... Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, 23(1), 14–24.
- MacKinnon, D. P., & Luecken, L. J. (2008). How and for whom? mediation and moderation in health psychology. *Health Psychology*, 27(2S), S99. doi: 10.1037/0278-6133.27.2(Suppl.).S99
- Margraf, J., Ehlers, A., & Roth, W. (1986). Sodium lactate infusions and panic attacks: A review and critique. *Psychosomatic Medicine*, 48(1/2), 23.
- Margraf, J., Taylor, B., Ehlers, A., Roth, W. T., & Agras, W. S. (1987). Panic attacks in the natural environment. *Journal of Nervous and Mental Disease*.

- 
- Marks, R. J. I. (2012). *Introduction to shannon sampling and interpolation theory*. Springer Science & Business Media.
- Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank ising networks. *Scientific Reports*, 5, 9050.
- Marsman, M., Tanis, C., Bechger, T., & Waldorp, L. (2019). Network psychometrics in educational practice. In *Theoretical and practical advances in computer-based educational measurement* (pp. 93–120). Springer.
- Mauger, R., Tarbuck, E. J., & Lutgens, F. K. (1996). *Earth: An introduction to physical geology*. Prentice-Hall.
- McMullin, E. (1982). Values in science. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (pp. 3–28).
- McNally, R. J. (1990). Psychological approaches to panic disorder: A review. *Psychological Bulletin*, 108(3), 403.
- McNally, R. J. (1994). *Panic disorder: A critical analysis*. Guilford Press.
- McNally, R. J. (2002). Anxiety sensitivity and panic disorder. *Biological Psychiatry*, 52(10), 938–946.
- McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental disorders as causal systems a network approach to posttraumatic stress disorder. *Clinical Psychological Science*, 3(6), 836–849.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806.
- Meehl, P. E. (1990a). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108–141.
- Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195–244.
- Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436–1462. doi: 10.1214/009053606000000281
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society B*, 72(4), 417–473.
- Meyer, K. (2016). A mathematical review of resilience in ecology. *Natural Resource Modeling*, 29(3), 339–352.
- Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4), 201–218.
- Monti, R. (2014). *pysingle*. <https://github.com/piomonti/pySINGLE>. GitHub.
- Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., & Montana, G. (2014). Estimating time-varying brain connectivity networks from functional MRI time series. *Neuroimage*, 103, 427–443. Retrieved 2015-10-29, from <http://www.sciencedirect.com/science/article/pii/S1053811914006168>

- Mooij, J. M., Janzing, D., & Schölkopf, B. (2013). From ordinary differential equations to structural causal models: the deterministic case. *arXiv preprint arXiv:1304.7920*.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 1.
- Nelder, J. A., & Baker, R. J. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*. doi: 10.2307/2344614
- Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T. W., & Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder: a review. *JAMA Psychiatry*, 74(5), 528–534.
- Nettle, D. (2009). An evolutionary model of low mood states. *Journal of Theoretical Biology*, 257(1), 100–103.
- Newman, M. (2010). *Networks: an introduction*. Oxford university press.
- Newman, M. E., Barabási, A.-L. E., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton University Press.
- Nguyen, J., & Frigg, R. (2017). Mathematics is not the only language in the book of nature. *Synthese*, 1–22.
- Nicholson, W., Matteson, D., & Bien, J. (2017). Bigvar: Dimension reduction methods for multivariate time series [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=BigVAR> (R package version 1.0.2)
- Nolen-Hoeksema, S., & Watkins, E. R. (2011). A heuristic for developing transdiagnostic models of psychopathology: Explaining multifinality and divergent trajectories. *Perspectives on Psychological Science*, 6(6), 589–609.
- Norton, P. J., Zvolensky, M. J., Bonn-Miller, M. O., Cox, B. J., & Norton, G. R. (2008). Use of the panic attack questionnaire-iv to assess non-clinical panic attacks and limited symptom panic attacks in student and community samples. *Journal of Anxiety Disorders*, 22(7), 1159–1171.
- Olthof, M., Hasselman, F., Strunk, G., van Rooij, M., Aas, B., Helmich, M. A., ... Lichtwarck-Aschoff, A. (2019). Critical fluctuations as an early-warning signal for sudden gains and losses in patients receiving psychotherapy for mood disorders. *Clinical Psychological Science*, 2167702619865969.
- Ottaviani, R., & Beck, A. T. (1987). Cognitive aspects of panic disorders. *Journal of Anxiety Disorders*, 1(1), 15–28.
- Otto, M. W., Tolin, D. F., Simon, N. M., Pearlson, G. D., Basden, S., Meunier, S. A., ... Pollack, M. H. (2010). Efficacy of d-cycloserine for enhancing response to cognitive-behavior therapy for panic disorder. *Biological Psychiatry*, 67(4), 365–370.
- Ou, L., Hunter, M. D., & Chow, S.-M. (2019). dynr: Dynamic modeling in r [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=dynr> (R package version 0.1.14-9)
- Papoulis, A., & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.
- Patzelt, E. H., Hartley, C. A., & Gershman, S. J. (2018). Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience*, 1.

- 
- Pauli, P., Marquardt, C., Hartl, L., Nutzinger, D. O., Hölzl, R., & Strian, F. (1991). Anxiety induced by cardiac perceptions in patients with panic attacks: A field study. *Behaviour Research and Therapy*, 29(2), 137–145.
- Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... others (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, 3(2), 292–300.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Peirce, C. S. (1931). *Collected papers of charles sanders peirce*. Harvard University Press.
- Pero, F. (2015). *Whither structuralism for scientific representation (dsc)*. Unpublished doctoral dissertation, Univeristy of Florence.
- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT Press.
- Pfaff, B. (2008a). *Analysis of Integrated and Cointegrated Time Series with R* (2nd edition ed.). New York: Springer-Verlag.
- Pfaff, B. (2008b). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, 27(4). Retrieved from <http://www.jstatsoft.org/v27/i04/>
- Pilecki, B., Arentoft, A., & McKay, D. (2011). An evidence-based causal model of panic disorder. *Journal of Anxiety Disorders*, 25(3), 381–388.
- R Core Team. (2014). *A language and environment for statistical computing. vienna, austria: R foundation for statistical computing*. ISBN 3-900051-07-0. <http://www.R-project.org>.
- Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
- Rachman, S., Craske, M., Tallman, K., & Solyom, C. (1986). Does escape behavior strengthen agoraphobic avoidance? a replication. *Behavior Therapy*, 17(4), 366–384.
- Radomsky, A. S., Rachman, S., Teachman, B. A., & Freeman, W. S. (1998). Why do episodes of panic stop? *Journal of Anxiety Disorders*, 12(3), 263–270.
- Ramsay, D. S., & Woods, S. C. (2014). Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychological Review*, 121(2), 225.
- Rapee, R. M. (1987). The psychological treatment of panic attacks: Theoretical conceptualization and review of evidence. *Clinical Psychology Review*, 7(4), 427–438.
- Rapee, R. M. (1995). Psychological factors influencing the affective response to biological challenge procedures in panic disorder. *Journal of Anxiety Disorders*, 9(1), 59–74.
- Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010, June). High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3), 1287–1319. Retrieved 2014-07-20, from <http://projecteuclid.org/euclid.aos/1268056617> doi: 10.1214/09-AOS691
- Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3), 1287–1319.



- Read, J. (1995). *From alchemy to chemistry*. Courier Corporation.
- Reiss, S., & McNally, R. (1985). *Expectancy model of fear*. u: Reiss s. i bootzin rr [ur.]. New York: Academic Press.
- Reiss, S., Peterson, R. A., Gursky, D. M., & McNally, R. J. (1986). Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. *Behaviour Research and Therapy*, 24(1), 1–8.
- Rescorla, R. A., Wagner, A. R., et al. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2, 64–99.
- Revelle, W. (2017). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from <https://CRAN.R-project.org/package=psych> (R package version 1.7.3)
- Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence*, 161, 230–237.
- Rinaldi, S., & Scheffer, M. (2000). Geometric analysis of ecological models with slow and fast processes. *Ecosystems*, 3(6), 507–521.
- Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., ... others (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, 18(4), 497.
- Roberson-Nay, R., Gorlin, E. I., Beadel, J. R., Cash, T., Vrana, S., & Teachman, B. A. (2017). Temporal stability of multiple response systems to 7.5% carbon dioxide challenge. *Biological psychology*, 124, 111–118.
- Roberson-Nay, R., & Kendler, K. (2011). Panic disorder and its subtypes: a comprehensive analysis of panic symptom heterogeneity using epidemiological and treatment seeking samples. *Psychological Medicine*, 41(11), 2411–2421.
- Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., ... Borsboom, D. (2019). Advancing the network theory of mental disorders: a computational model of panic disorder.
- Robinaugh, D. J., Hoekstra, R. H., Toner, E. R., & Borsboom, D. (2019). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, 1–14.
- Robinaugh, D. J., Hoekstra, R. H. A., & Borsboom, D. (2019). The network approach to psychopathology: A review of the literature 2008–2018.
- Robinaugh, D. J., LeBlanc, N. J., Vuletic, H. A., & McNally, R. J. (2014). Network analysis of persistent complex bereavement disorder in conjugally bereaved adults. *Journal of Abnormal Psychology*, 123(3), 510.
- Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, 125(6), 747.
- Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical analysis and forecasting of economic structural change* (pp. 253–264). Springer.
- Ross, T. A. (1937). *The common neuroses, their treatment by psychotherapy. an introduction to psychological treatment for students and practitioners*.

- 
- Roth, W. T., Wilhelm, F. H., & Pettit, D. (2005). Are current theories of panic falsifiable? *Psychological Bulletin*, 131(2), 171.
- Ruzzano, L., Borsboom, D., & Geurts, H. M. (2015). Repetitive behaviors in autism and obsessive-compulsive disorder: New perspectives from a network analysis. *Journal of Autism and Developmental Disorders*, 45(1), 192–202.
- Ryan, O. (2018). *Interventions in dynamic systems: A causal approach to continuous-time mediation analysis*. PsyArXiv. Retrieved from psyarxiv.com/n2fwt doi: 10.31234/osf.io/n2fwt
- Ryan, O., Bringmann, L. F., & Schuurman, N. K. (2019). *The challenge of generating causal hypotheses using network models*. PsyArXiv. Retrieved from psyarxiv.com/ryg69 doi: 10.31234/osf.io/ryg69
- Salkovskis, P. M. (1988). Phenomenology, assessment, and the cognitive model of panic. *Panic: psychological perspectives*, 111–136.
- Salkovskis, P. M. (1991). The importance of behaviour in the maintenance of anxiety and panic: A cognitive account. *Behavioural and Cognitive Psychotherapy*, 19(1), 6–19.
- Salkovskis, P. M., Clark, D. M., Hackmann, A., Wells, A., & Gelder, M. G. (1999). An experimental investigation of the role of safety-seeking behaviours in the maintenance of panic disorder with agoraphobia. *Behaviour Research and Therapy*, 37(6), 559–574.
- Salmon, W. (1989). Scientific explanation. In P. Kitcher (Ed.), *Minnesota studies in the philosophy of science, vol 13*. Minneapolis: University of Minnesota Press.
- Sandin, B., Sánchez-Arribas, C., Chorot, P., & Valiente, R. M. (2015). Anxiety sensitivity, catastrophic misinterpretations and panic self-efficacy in the prediction of panic disorder severity: Towards a tripartite cognitive model of panic disorder. *Behaviour Research and Therapy*, 67, 30–40.
- Santos Jr, H., Fried, E. I., Asafu-Adjei, J., & Ruiz, R. J. (2017). Network structure of perinatal depressive symptoms in latinas: relationship to stress and reproductive biomarkers. *Research in Nursing & Health*, 40(3), 218–228.
- Savi, A., van der Maas, H., Maris, G., et al. (2018). The wiring of intelligence.
- Sayama, H. (2015). *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks.
- Schaffner, K., et al. (1994). Psychiatry and molecular biology: reductionistic approaches to schizophrenia. *Philosophical perspectives on psychiatric diagnostic classification*, 279–294.
- Scheffer, M. (2009). *Critical transitions in nature and society*. Princeton University Press. Who Are We.
- Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., ... Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260), 53. doi: 10.1038/nature08227
- Scheffer, M., Bolhuis, J. E., Borsboom, D., Buchman, T. G., Gijzel, S. M., Goulson, D., ... others (2018). Quantifying resilience of humans and other animals. *Proceedings of the National Academy of Sciences*, 115(47), 11883–11890.
- Schmidt, N. B., Lerew, D. R., & Jackson, R. J. (1997). The role of anxiety sensitivity

- in the pathogenesis of panic: Prospective evaluation of spontaneous panic attacks during acute stress. *Journal of Abnormal Psychology*, 106(3), 355.
- Schmidt, N. B., Lerew, D. R., & Jackson, R. J. (1999). Prospective evaluation of anxiety sensitivity in the pathogenesis of panic: replication and extension. *Journal of Abnormal Psychology*, 108(3), 532.
- Schmidt, N. B., Zvolensky, M. J., & Maner, J. K. (2006). Anxiety sensitivity: Prospective prediction of panic attacks and axis I pathology. *Journal of Psychiatric Research*, 40(8), 691–699.
- Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, 31(1), 43–53.
- Schmittmann, V. D., Jahfari, S., Borsboom, D., Savi, A. O., & Waldorp, L. J. (2015). Making large-scale networks from fmri data. *PLoS ONE*, 10(9), e0129074.
- Schulze, P., et al. (1996). *Engineering within ecological constraints*. National Academies Press.
- Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Segal, Z. V. (1988). Appraisal of the self-schema construct in cognitive models of depression. *Psychological Bulletin*, 103(2), 147.
- Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2, 395.
- Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, 52(2), 149–178. doi: 10.1016/j.jsp.2013.11.004
- Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'Er, I., ... others (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, 460(7256), 753.
- Shulgin, B., Stone, L., & Agur, Z. (1998). Pulse vaccination strategy in the sir epidemic model. *Bulletin of mathematical biology*, 60(6), 1123–1148.
- Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., & Thase, M. E. (2003). Do the seconds turn into hours? relationships between sustained pupil dilation in response to emotional information and self-reported rumination. *Cognitive Therapy and Research*, 27(3), 365–382.
- Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: event-related fmri assessment of sustained amygdala activity in response to emotional information in depressed individuals. *Biological Psychiatry*, 51(9), 693–707.
- Smaldino, P. E. (2017). Models are stupid, and we need more of them. In *Computational social psychology* (pp. 311–331). Routledge.
- Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11(1), 87–104.
- Snippe, E., Viechtbauer, W., Geschwind, N., Klippel, A., De Jonge, P., & Wichers, M. (2017). The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials. *Scientific Reports*, 7, 46523.

- 
- Song, L., Kolar, M., & Xing, E. P. (2009, June). KELLER: Estimating time-varying interactions between genes. *Bioinformatics*, 25(12), i128–i136. doi: 10.1093/bioinformatics/btp192
- Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, 11(May), 1643–1662.
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT Press.
- Spitzer, R. L., Md, K. K., & Williams, J. B. (1980). Diagnostic and statistical manual of mental disorders. In *American psychiatric association*.
- Steele, R. J., & Raftery, A. E. (2010). Performance of bayesian model selection criteria for gaussian mixture models. *Frontiers of Statistical Decision Making and Bayesian Analysis*, 2, 113–130.
- Stephan, K. E., & Mathys, C. (2014, April). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, 25, 85–92. Retrieved 2020-02-15, from <https://linkinghub.elsevier.com/retrieve/pii/S0959438813002316> doi: 10.1016/j.conb.2013.12.007
- Stephens, P. A., Sutherland, W. J., & Freckleton, R. P. (1999). What is the allee effect? *Oikos*, 185–190.
- Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A., ... Hardy, G. E. (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology*, 71(1), 14.
- Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of qualitative research*, 17, 273–85.
- Strogatz, S. H. (2015). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Reading, MA.
- Stutz, C., & Williams, B. (1999). Obituary: Ernst ising. *Physics Today*, 52, 106–108.
- Suárez, M., & Pero, F. (2019). The representational semantic conception. *Philosophy of Science*, 86(2), 344–365.
- Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Computational Biology*, 9(1).
- Suppes, P. (1962). *Models of data. logic in e. nagel, p. suppes & a. tarski (eds.) methodology and the philosophy of science: Proceedings of the 1960 international congress. palo alto*. CA, Stanford University Press.
- Swain, K. (2018). ‘extraordinarily arduous and fraught with danger’: syphilis, salvarsan, and general paresis of the insane. *The Lancet Psychiatry*, 5(9), 702–703.
- Swoyer, C. (1991). Structural representation and surrogate reasoning. *Synthese*, 87(3), 449–508.
- Szollosi, A., & Donkin, C. (2019). Neglected sources of flexibility in psychological theories: From replicability to good explanations. *Computational Brain & Behavior*, 1–3.
- T, B. A. (1985). Anxiety disorders and phobias. a cognitive perspective. *New York Basic Books*.

- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Allyn & Bacon/Pearson Education.
- Tao, Q., Huang, X., Wang, S., Xi, X., & Li, L. (2016, March). Multiple Gaussian Graphical Estimation with Jointly Sparse Penalty. *Signal Processing*. Retrieved 2016-04-06, from <http://linkinghub.elsevier.com/retrieve/pii/S0165168416300032> doi: 10.1016/j.sigpro.2016.03.009
- Tarvainen, M. P., Hiltunen, J. K., Ranta-aho, P. O., & Karjalainen, P. A. (2004). Estimation of nonstationary eeg with kalman smoother approach: an application to event-related synchronization (ers). *IEEE Transactions on Biomedical Engineering*, 51(3), 516–524. doi: 10.1109/TBME.2003.821029
- Taylor, C. B., King, R., Ehlers, A., Margraf, J., Clark, D., Hayward, C., ... Agras, S. (1987). Treadmill exercise test and ambulatory measures in panic attacks. *The American Journal of Cardiology*, 60(18), J48–J52.
- Taylor, C. B., Sheikh, J., Agras, W. S., Roth, W. T., Margraf, J., Ehlers, A., ... Gossard, D. (1986). Ambulatory heart rate changes in patients with panic attacks. *The American Journal of Psychiatry*.
- Taylor, S. (1994). Comment on otto et al.(1992): Hypochondriacal concerns, anxiety sensitivity, and panic disorder. *Journal of Anxiety Disorders*, 8(1), 97–99.
- Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, 75(2), 76–92.
- Thagard, P. R. (1989). Explanatory coherence. *Behavioral and brain sciences*, 12(3), 435–467.
- Tong, H., & Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3), 245–268.
- Trip, D. S. L., & van Wieringen, W. N. (2018). A parallel algorithm for penalized learning of the multivariate exponential family from data of mixed types. *arXiv preprint arXiv:1812.02401*.
- van Borkulo, C. D. (2016). Networkcomparisonstest: Statistical comparison of two networks based on three invariance measures [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=NetworkComparisonTest> (R package version 2.0.1)
- Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, 4, 5918.
- van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014, August). A new method for constructing networks from binary data. *Scientific Reports*, 4. Retrieved 2014-08-07, from <http://www.nature.com/doifinder/10.1038/srep05918> doi: 10.1038/srep05918
- van Borkulo, C. D., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA Psychiatry*, 72(12), 1219–1226.

- 
- van Borkulo, C. D., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R., Borsboom, D., & Waldorp, L. (2016). Comparing network structures on three aspects: A permutation test. *Manuscript submitted for publication*, 24. doi: 10.1561/2200000001
- van Borkulo, C. D., Epskamp, S., & Robitzsch, w. c. f. A. (2014, October). *Isingfit: Fitting Ising models using the eLasso method*. Retrieved 2015-09-16, from <https://cran.r-project.org/web/packages/IsingFit/index.html>
- van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., ... others (2014a). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1), 87–92.
- van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., ... Scheffer, M. (2014b, January). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1), 87–92. Retrieved 2015-10-29, from <http://www.pnas.org/cgi/doi/10.1073/pnas.1312114110> doi: 10.1073/pnas.1312114110
- van den Hout, M. (2014). Psychiatric symptoms as pathogens. *Clinical Neuropsychiatry: Journal of Treatment Evaluation*, 11(6), 153–159.
- van den Hout, M., & Griez, E. (1983). Some remarks on the nosology of anxiety states and panic disorders. *Acta Psychiatrica Belgica*, 83(1), 33–42.
- van der Krieke, L., Blaauw, F. J., Emerencia, A. C., Schenk, H. M., Slaets, J. P., Bos, E. H., ... Jeronimus, B. F. (2017). Temporal dynamics of health and well-being: A crowdsourcing approach to momentary assessments and automated generation of personalized feedback. *Psychosomatic medicine*, 79(2), 213–223. doi: 10.1097/PSY.0000000000000378
- Van Der Maas, H., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence*, 5(2), 16.
- van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842.
- van Nes, E. H., & Scheffer, M. (2004). Large species shifts triggered by small forces. *The American Naturalist*, 164(2), 255–266.
- Van Nes, E. H., & Scheffer, M. (2007). Slow recovery from perturbations as a generic indicator of a nearby catastrophic shift. *The American Naturalist*, 169(6), 738–747.
- Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner Jr, T. E. (2010). The interpersonal theory of suicide. *Psychological Review*, 117(2), 575.
- van Rooijen, G., Isvoranu, A.-M., Meijer, C. J., van Borkulo, C. D., Ruhé, H. G., de Haan, L., et al. (2017). A symptom network structure of the psychosis spectrum. *Schizophrenia Research*, 189, 75–83.
- Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, 36(7), 1–21. Retrieved from

- <http://www.jstatsoft.org/v36/i07/>
- Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of Management Journal*, 15(4), 407–426.
- von Kentzinsky, H., Wijtsma, S., & Treur, J. (2019). A temporal-causal modelling approach to analyse the dynamics of burnout and the effects of sleep.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638.
- Wainwright, M. J., & Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2), 1–305. doi: 10.1561/22000000001
- Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2), 1–305.
- Walker, M. (2017). *Why we sleep: Unlocking the power of sleep and dreams*. Simon and Schuster.
- Walters, J. M., & Gardner, H. (1986). The theory of multiple intelligences: Some issues and answers. *Practical intelligence: Nature and origins of competence in the everyday world*, 163–182.
- Wan, Y.-W., Allen, G. I., Baker, Y., Yang, E., Ravikumar, P., & Liu, Z. (2015). Xmrfr: Markov random fields for high-throughput genetics data [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=XMRFR> (R package version 1.0)
- Wang, X.-J., & Krystal, J. (2014, November). Computational Psychiatry. *Neuron*, 84(3), 638–654. Retrieved 2020-02-15, from <https://linkinghub.elsevier.com/retrieve/pii/S0896627314009167> doi: 10.1016/j.neuron.2014.10.018
- Warren, K. (2002). Thresholds and the abstinence violation effect: A nonlinear dynamical model of the behaviors of intellectually disabled sex offenders. *Journal of Interpersonal Violence*, 17(11), 1198–1217.
- Wegener, A. (1966). *The origin of continents and oceans*. Courier Corporation.
- White, K. S., Brown, T. A., Somers, T. J., & Barlow, D. H. (2006). Avoidance behavior in panic disorder: The moderating influence of perceived control. *Behaviour Research and Therapy*, 44(1), 147–157.
- Wichers, M., Groot, P. C., Psychosystems, E., Group, E., et al. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, 85(2), 114–116. doi: 10.1159/000441458
- Wichers, M., Schreuder, M. J., Goekoop, R., & Groen, R. N. (2019). Can we predict the direction of sudden shifts in symptoms? transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological Medicine*, 49(3), 380–387.
- Wichers, M., Wigman, J. T. W., & Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, 7(4), 362–367.
- Wigman, J. T. W., de Vos, S., Wichers, M., van Os, J., & Bartels-Velthuis, A. A. (2016). A transdiagnostic network approach to psychosis. *Schizophrenia*

- 
- Bulletin*, 43(1), 122–132.
- Wigman, J. T. W., van Os, J., Borsboom, D., Wardenaar, K., Epskamp, S., Klippel, A., ... Wichers, M. (2015). Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. *Psychological Medicine*, 45(11), 2375–2387. doi: 10.1017/S0033291715000331
- Wild, B., Eichler, M., Friederich, H.-C., Hartmann, M., Zipfel, S., & Herzog, W. (2010). A graphical vector autoregressive modelling approach to the analysis of electronic diary data. *BMC medical research methodology*, 10(1), 28.
- Williams, D. R., & Rast, P. (2018). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/bmsp.12173
- Wimsatt, W. C. (1987). False models as means to truer theories. *Neutral models in biology*, 23–55.
- Wittchen, H.-U., Gloster, A. T., Beesdo-Baum, K., Fava, G. A., & Craske, M. G. (2010). Agoraphobia: a review of the diagnostic classificatory position and criteria. *Depression and Anxiety*, 27(2), 113–133.
- Wittenborn, A., Rahmandad, H., Rick, J., & Hosseinichimeh, N. (2016). Depression as a systemic syndrome: mapping the feedback loops of major depressive disorder. *Psychological Medicine*, 46(3), 551–562.
- Wolfram Research, Inc. (n.d.). *Mathematica, Version 12.0*. (Champaign, IL, 2019)
- Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.
- Wood, S. N., & Augustin, N. H. (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157(2), 157–177. doi: 10.1016/S0304-3800(02)00193-X
- Woods, S. W., Charney, D. S., Goodman, W. K., & Heninger, G. R. (1988). Carbon dioxide—induced anxiety: behavioral, physiologic, and biochemical effects of carbon dioxide in patients with panic disorders and healthy subjects. *Archives of General Psychiatry*, 45(1), 43–52.
- Woodward, J. F. (2011). Data and phenomena: a restatement and defense. *Synthese*, 182(1), 165–179.
- Woodward, J. F. (2014). Scientific explanation. In E. Z. Zalda (Ed.), *Stanford encyclopedia of philosophy*. The Metaphysics Research Lab.
- Wray, N. R., Pergadia, M., Blackwood, D., Penninx, B., Gordon, S., Nyholt, D., ... others (2012). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Molecular Psychiatry*, 17(1), 36.
- Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., ... others (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5), 668.
- Yang, E., Baker, Y., Ravikumar, P., Allen, G., & Liu, Z. (2014a). Mixed graphical models via exponential families. In *Artificial intelligence and statistics* (pp. 1042–1050).



- Yang, E., Baker, Y., Ravikumar, P., Allen, G., & Liu, Z. (2014b). Mixed Graphical Models via Exponential Families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (pp. 1042–1050). Retrieved 2014-09-26, from <http://jmlr.org/proceedings/papers/v33/yang14a.pdf>
- Yang, E., Ravikumar, P., Allen, G. I., & Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1), 3813–3847.
- Yang, E., Ravikumar, P. K., Allen, G. I., & Liu, Z. (2013). On poisson graphical models. In *Advances in neural information processing systems* (pp. 1718–1726).
- Yarkoni, T. (2019). The generalizability crisis.
- Zhang, B., Horvath, S., et al. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 1128.
- Zhao, T., Li, X., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2015). huge: High-dimensional undirected graph estimation [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=huge>
- Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13(1), 1059–1062. Retrieved 2015-09-16, from <http://dl.acm.org/citation.cfm?id=2343681>
- Zhou, S., Lafferty, J., & Wasserman, L. (2010a). Time varying undirected graphs. *Machine Learning*, 80(2-3), 295–319.
- Zhou, S., Lafferty, J., & Wasserman, L. (2010b, September). Time varying undirected graphs. *Machine Learning*, 80(2-3), 295–319. Retrieved 2016-04-06, from <http://link.springer.com/10.1007/s10994-010-5180-0> doi: 10.1007/s10994-010-5180-0
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2), 301–320.

---

# NODEWISE PREDICTABILITY: REANALYSIS

---

## A.1 Details about Literature Review

We performed a literature search on the databases PsycNET, ISI Web of Science and GoogleScholar using “Network AND X” as a keyword, where we made 9 separate searches, where X was either “Psychopathology”, “Comorbidity”, “Post Traumatic Stress Disorder”, “De-pression”, “Anxiety”, “Schizophrenia”, “Psychosis”, “Personality Disorder”, or “Substance”. We constrained our search to the period 2010 – 2016 as we consider the paper of Cramer et al. (2010) as the first ‘network paper’ in the field of psychopathology. While we checked all search results for PsycNET and ISI Web of Science. For GoogleScholar we only went through the first 10 pages of results, because going through all results was not feasible (e.g. the search query “Network” + “Psychopathology” led to 187,000 results (10/7/2016)).

Table A.1 lists the 23 papers we found by combining papers the authors knew of with the additional papers found in the literature review.

Paper	Outcome of Data Request
Anderson et al. (2015)	Obtained
Armour et al. (2017)	Obtained
Beard et al. (2016)	Obtained
Borsboom and Cramer (2013)	Obtained
Boschloo et al. (2015)	New policy of U.S. National Institute on Alcohol Abuse and Alcoholism does not allow sharing data anymore (personal communication)
Boschloo, van Borkulo, et al. (2016)	Obtained
Boschloo, van Borkulo, et al. (2016)	Requirements to obtain data from NESDA for re-analysis unfeasible for this project)
Curtiss and Klemanski (2016)	Did not share their data
Cramer et al. (2010)	Data Identical to Borsboom and Cramer (2013)
Deserno et al. (2017)	Obtained
Fried et al. (2015)	Obtained
Fried, Epskamp, et al. (2016)	Obtained
Goekoop and Goekoop (2014)	Obtained
Hoorelbeke et al. (2016)	Obtained
Koenders et al. (2015)	Obtained
McNally et al. (2015)	Obtained
Rhemtulla et al. (2016)	Obtained
Robinaugh et al. (2014)	Obtained
Robinaugh et al. (2016)	Obtained
Ruzzano et al. (2015)	Obtained
Santos Jr et al. (2017)	Obtained
van Borkulo et al. (2015)	Requirements to obtain data from NESDA for re-analysis unfeasible for this project)
Wigman et al. (2016)	Obtained

**Table A.1:** All 23 papers retrieved from the literature review and the outcome of the data request.

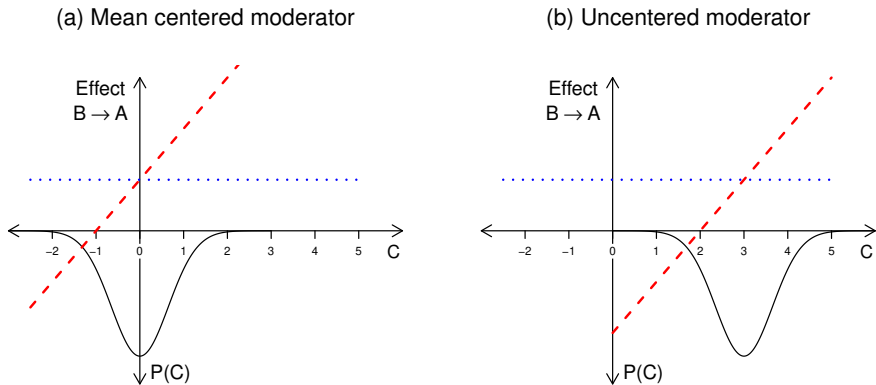
---

# MODERATED NETWORK MODELS

---

## B.1 Mean-centering in Moderation Analysis

In the procedure to estimate MNM described in Section 5.2.5 we mean-center all variables before estimation. The reason is that for centered variables, the interpretation of parameters is more meaningful. We illustrate this issue in Figure B.1:



**Figure B.1:** Illustration of the advantages of mean-centering predictors in moderated regression. For the mean-centered moderator in (a) the effect of  $B$  on  $A$  for  $C = 0$  is the same as the effect of  $B$  on  $A$  when ignoring  $C$  (dotted blue line), which allows to compare parameters in models with/without moderation. In (b) this is not the case. Indeed, the effect of  $B$  on  $A$  if  $C = 0$  is an effect that is hardly ever observed because the probability of observing values of  $C$  close to 0 is extremely small.

In Figure B.1 (a), the red dashed line represents the partially moderated effect which we already considered above in Figure 5.1 (d). In this case the moderator  $C$  was mean-centered and its mean is therefore equal to zero. Recall the interpretation of the parameter  $\beta_B$  in (5.3): it is the effect of  $B$  on  $A$  when  $C = 0$ . The blue dotted line represents the effect of  $B$  on  $A$  when averaging over all values of  $C$ , which is the effect of  $B$  on  $A$  one would obtain from a regression *without* moderation effects. We see that the red and blue line intersect at  $C = 0$ : this means that  $\beta_B$  in the moderated regression in (5.3) has the same value as  $\beta_B$  in the regression without moderation in (5.1). This is desirable because it allows to compare the parameters in models with/without moderation. Second, mean-centering moderators ensures that the parameter  $\beta_B$  is meaningful in the sense

that the probability of  $C$ -values close to zero is large. This is in contrast to Figure B.1 (b) in which the mean of the moderator  $C$  is equal to 3. Now the effect of  $B$  on  $A$  when  $C = 0$  does not intersect with the blue line at the appropriate level. Therefore,  $\beta_B$  captures the effect of  $B$  on  $A$  for values of  $C$  close to zero which occur with an extremely low probability and are therefore irrelevant.

For a detailed discussion of interaction/moderation effects in linear regression we refer the reader to Aiken et al. (1991) and Afshartous and Preston (2011).

## B.2 Joint distribution for $p = 3$

Here we show for the case of  $p = 3$  variables how to factorize  $p$  conditional distributions to a joint distribution.

We begin with the standard formulation of the conditional univariate Gaussian distribution

$$P(X_1|X_2 = x_2, X_3 = x_3) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_1 - \mu_1)^2}{2\sigma^2}\right\},$$

where the mean of  $X_1$ ,  $\mu_1$  is a function of  $X_2$  and  $X_3$ . If we let  $\sigma = 1$  and expand  $(X_1 - \mu_1)^2$  we get

$$P(X_1|X_2 = x_2, X_3 = x_3) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{X_1^2 + \mu_1^2 - 2X_1\mu_1}{2}\right\},$$

and can rearrange

$$P(X_1|X_2 = x_2, X_3 = x_3) = \frac{1}{\sqrt{2\pi}} \exp\left\{X_1\mu_1 - \frac{X_1^2}{2} - \frac{\mu_1^2}{2}\right\}.$$

Our focus is on  $\mu_1$ , so we absorb  $\frac{1}{\sqrt{2\pi}}$  and  $-\frac{\mu_1^2}{2}$  in the log-normalizing constant  $\Psi_1(\alpha, \beta, \omega)$  and let  $C_1 = \frac{X_1^2}{2}$

$$P(X_1|X_2 = x_2, X_3 = x_3) = \exp\{X_1\mu_1 - C_1 - \Psi_1(\alpha, \beta, \omega)\} \quad (\text{B.1})$$

where  $\alpha, \beta$ , and  $\omega$  are the parameter vectors defining the mean  $\mu_1$ , which is a linear combination of the other variables  $X_2, X_3$ :

$$\mu_1 = \alpha_1 + \beta_{2,1}X_2 + \beta_{3,1}X_3 + \omega_{2,3,1}X_2X_3 \quad (\text{B.2})$$

where  $\alpha_1$  is the intercept,  $\beta_{2,1}, \beta_{3,1}$  are the parameters for the pairwise interactions with  $X_2$  and  $X_3$ , and  $\omega_{2,3,1}$  is the parameter for the three-way interaction with  $X_2X_3$  (or equivalently, the pairwise interaction with  $X_2$  moderated by  $X_3$  or the pairwise interaction with  $X_3$  moderated by  $X_2$ ).

Plugging the mean (B.2) into the conditional distribution (B.1) gives us

$$P(X_1|X_2 = x_2, X_3 = x_3) = \exp\{X_1(\alpha_1 + \beta_{2,1}X_2 + \beta_{3,1}X_3 + \omega_{2,3,1}X_2X_3) - C_1 - \Psi_1(\alpha, \beta, \omega)\}.$$

Multiplying out and collecting terms of the same order gives

$$P(X_1|X_2 = x_2, X_3 = x_3) = \exp\{\alpha_1 X_1 + X_1 \sum_{j=1, j \neq 1}^3 \beta_{i,j} X_j + X_1 \omega_{1,2,3} X_2 X_3 - C_1 - \Psi_1(\alpha, \beta, \omega)\}.$$

Similarly define  $P(X_2|X_1 = x_1, X_3 = x_3)$  and  $P(X_3|X_2 = x_2, X_1 = x_1)$ . Then factorize the three conditional distributions to obtain a joint distribution. After rearranging terms we get

$$P(X_1, X_2, X_3) = \exp\left\{\sum_{i=1}^3 \alpha_i X_i + \sum_{i=1}^3 \sum_{j=1, j \neq i}^3 \beta_{i,j} X_i X_j + \omega_{1,2,3} X_1 X_2 X_3 - \sum_{i=1}^3 [C_i + \Psi_i(\alpha, \beta, \omega)]\right\}, \quad (\text{B.3})$$

where we combined the parameters  $\beta_{i,j}, \beta_{j,i}$  and  $\omega_{j,i,z}, \omega_{i,j,z}, \omega_{z,j,i}$  into single parameters by taking their average.

The joint distribution can be constructed analogously for any  $p$ .

A sufficient condition for (B.3) to be normalizable is that the sum over all terms in the exponential is negative (Yang et al., 2014a). However, the constraints  $\mathcal{C}(\alpha, \beta, \omega)$  on the parameter space  $\{\alpha, \beta, \omega\}$  to ensure this, are unknown. Note that these constraints are possibly very complicated since they depend on the variances of the conditional distributions and the structure of the factor graph defining this higher order model (Koller & Friedman, 2009). To be able to sample from (B.3), we use a rejection sampler as described in Appendix B.3.

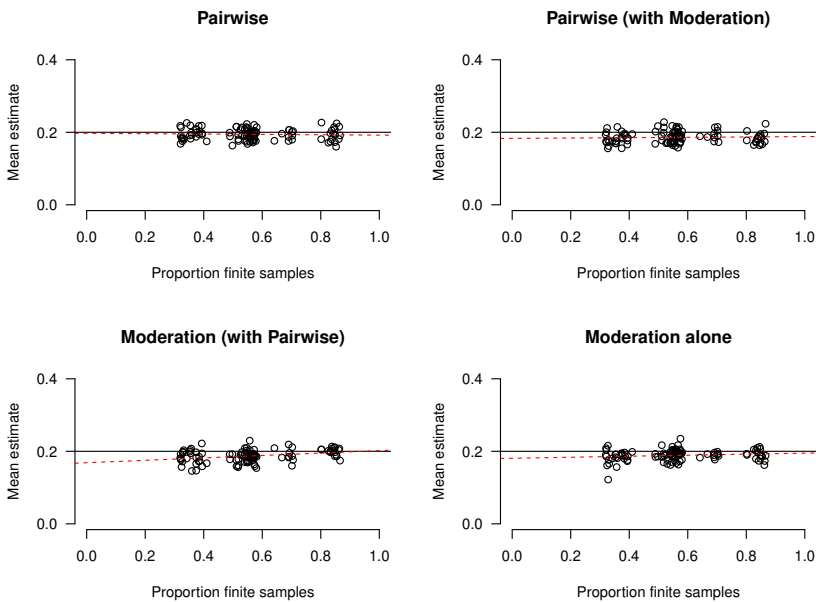
## B.3 Rejection Sampling

We do not know the constraints  $\mathcal{C}(\alpha, \beta, \omega)$  on the parameter space that ensure that the model in (B.3) (and it's generalization to  $p$  variables) is normalizable. To still be able to sample from this distribution we use a rejection sampler that rejects diverging chains in the Gibbs sampler Casella and George (1992). Specifically, we fix the MNM and sample cases using the Gibbs sampler, with a burn-in of 100 iterations. A chain is defined to diverge if  $|X_i| > \tau$  for at least one  $i \in \{1, 2, \dots, p\}$ , where we set  $\tau = 3.09$ , which is the 99.9% quantile of a standard normal distribution.

This way we define  $\mathcal{C}(\alpha, \beta, \omega)$  indirectly. If the chain remains within  $[-\tau, \tau]$  we assume that the constraints  $\mathcal{C}(\alpha, \beta, \omega)$  are satisfied. If the chain diverges, we know that  $\mathcal{C}(\alpha, \beta, \omega)$  is not satisfied. Note that diverging chains approach  $\pm\infty$  very quickly, therefore the exact value of  $\tau$  has only a small impact on the sampling procedure.

With this procedure it would be possible to sample cases for any fixed MNM generated with the procedure described in Section 5.3.1. However, for some MNM the proportion of rejected samples might be high, and consequently the running time until obtaining  $n = 1808$  cases would be very large. To keep the computational manageable, we sampled  $n = 10000$  cases from 130 models as specified in Section 5.3.1. We then ordered the 130 population models by increasing proportion of diverged samples, and selected the first 100. In those 100 iterations the proportion of rejected samples varied between 0.287 and 0.886. We take the  $n = 1808$  first observations in each of the 100 data sets. These data are used in the simulation study.

To check whether the rejection sampling introduced bias in the estimates, we recover every single parameter in the 100 data sets, using standard linear regressions in which we specified the correct moderator. Within data sets we take the average over pairs of parameter types (unmoderated, partially moderated, fully moderated) estimated from  $n = 1808$  observations, and plot them as a function of the proportion of finite samples in the given iteration in Figure B.2:



**Figure B.2:** Each point is the average over the two estimates of a given type in a given iteration, estimated with multiple regression with the correct moderator specified.

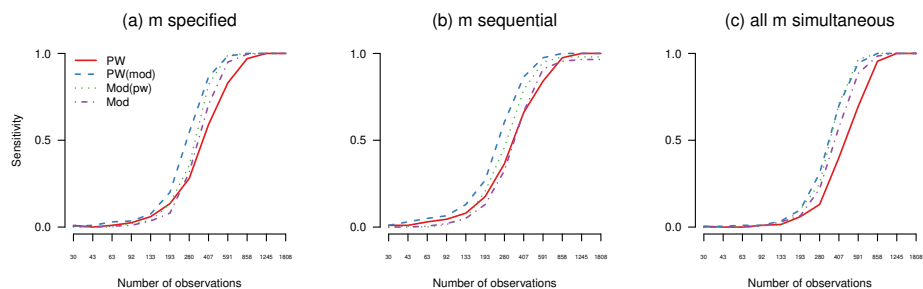
Figure B.2 shows that the unmoderated pairwise interactions are slightly biased downwards ( $\bar{\beta}_{pw} = 0.194$ ). The remaining three parameter types show a stronger downward bias ( $\bar{\beta}_{pw(mod)} = 0.186$ ,  $\bar{\omega}_{mod(pw)} = 0.187$  and  $\bar{\omega}_{mod} = 0.189$ ). The bias of the moderation effect seems to decrease with increasing proportion of finite samples. However, all parameter estimates are close to the value speci-



fied in the Gibbs sampler and hence the data can be used in the simulation. The result of this slight downward bias is that the performance across all conditions and estimators is also slightly biased downward.

## B.4 Sensitivity of Moderated Network Model across Parameter Types

In Section 5.3.3 we claimed that the sensitivity to detect unmoderated pairwise interactions (row 1 in Figure 5.4) and full moderations (row 2 in Figure 5.5) was lower than for detecting the pairwise interaction (row 2 in Figure 5.4) and moderation effect (row 2 Figure 5.5) in the partially moderated pairwise interaction. This is not easy to see in these figures and we therefore provide an additional figure with this comparison in Figure B.3:

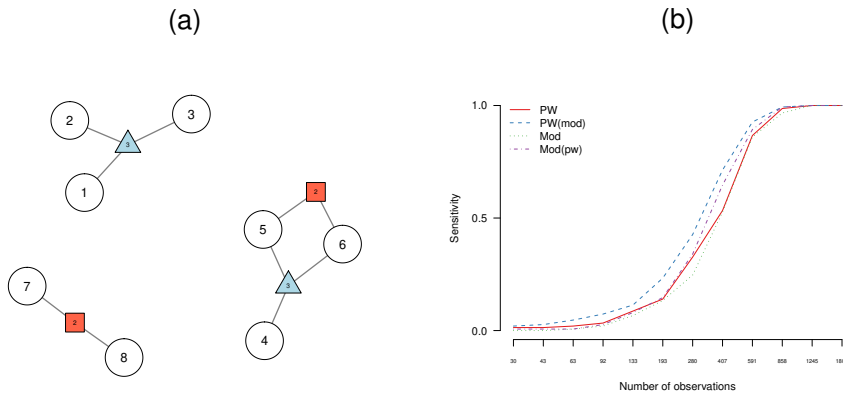


**Figure B.3:** Sensitivity to detect the four different parameter types, separately for the moderated network model with (a) correctly specified moderator, (b) sequentially searching all moderators and (c) specifying all moderators at once.

To explain this difference, we first run a simpler simulation in which we isolate all interaction types to exclude the possibility that the differences are explained by some characteristic of the graph (Appendix B.5). We will find that in such a "clean" setting the difference is even larger. Then we explain this difference in terms of the number of uncorrelated neighbors (Appendix B.6).

## B.5 Simulation with Isolated Interaction Types

Here we run a simplified version of the simulation reported in Section 6.3 to exclude the possibility that the sensitivity differences discussed in Appendix B.4 can be explained by some graph characteristic. We generate data from the graph shown in Figure B.4 panel (a):



**Figure B.4:** Panel (a): simplest possible graph including all four parameter types that separates all types as far as possible; panel (b): sensitivity to recover each parameter type for different numbers of observation for the moderated network model with specified moderator.

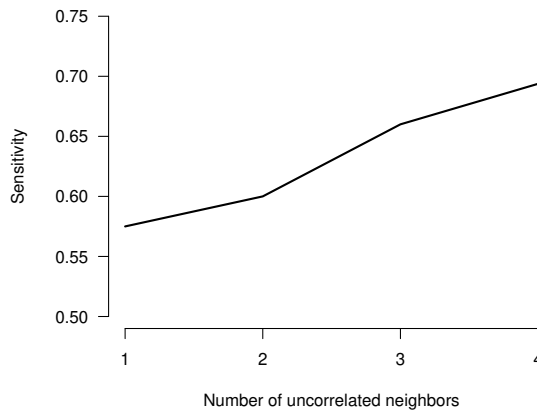
Except the graph structure, we use the same setup as reported in Section 5.3.1. In this graph the four parameter types are isolated as much as possible, so that the graph structure does not have any influence on estimation: 7-8 is isolated; the fully moderated pairwise interaction (or 3-way interaction) 1-2-3 is isolated; and the partially moderated pairwise interaction 4-5-6 is isolated. Note that the pairwise interaction 5-6 and the moderation effect 4-5-6 cannot be separated by definition.

We estimated the moderated network model with specified moderator (here variables 4 and 1). If we find the same sensitivity difference in this simulation as in Appendix B.4, we know that it is not a function of some unexpected graph characteristic. Figure B.4 panel (b) shows the sensitivity for the four parameter types as a function of the number of observations. We observe the same sensitivity difference as in Appendix B.4.

What is the difference between the pairwise interaction 7-8 and the moderation effect (or 3-way interaction) 1-2-3 on the one hand, and the pairwise interaction and the moderation effect in 4-5-6 on the other hand? The difference is that in the respective nodewise regressions, in the former case there is one nonzero predictor and in the latter case there are two nonzero predictors (see equation (5.2) in Section 5.2.1). The presence of two nonzero predictors results in that the EBIC selects a smaller (compared to the presence of one nonzero predictor) penalty parameter  $\lambda_s$ , which increases sensitivity. This reasoning is based on the assumption that the predictors are uncorrelated, which is the case if all variables and interaction effects are centered as in the present case. In Appendix B.7, we show that  $X$  is uncorrelated with  $XY$  if both  $X$  and  $Y$  are centered. In Appendix B.6, we test this explanation directly by investigating sensitivity as a function of the number of uncorrelated neighbors in a GGM.

## B.6 Sensitivity as Function of Number of Uncorrelated Neighbors

Here the goal is to directly test the hypothesis that the sensitivity to detect the edges between nodes  $X_1$  and  $X_2, \dots, X_p$  increases with the number of nonzero edges, if  $X_2, \dots, X_p$  are uncorrelated. To this end we generate a GGM with  $p = 20$ , which matches the maximum neighborhood size of each node to that of the simplified simulation in Appendix B.5. We compare four versions of this GGM, across which we vary the number of uncorrelated neighbors from 1-4. The 1-4 nonzero partial correlations have all the value  $\beta = 0.2$ , matching the setup of the simulations in Section 6.3 and Appendix B.5. We show the average (over 1-4 edges in the four conditions, respectively) sensitivity in Figure B.5:



**Figure B.5:** Sensitivity to detect a neighbor connected to node  $X_1$  as a function of the number of uncorrelated neighbors of  $X_1$ .

The results confirm the hypothesis that the sensitivity to detect neighbors increases as a function of the number of uncorrelated neighbors.

## B.7 $XY$ is uncorrelated with $X$ and $Y$ if the latter are mean-centered

For independent and identically distributed variables  $X, Y$  with finite variances, let  $Z = XY$ . We show  $\rho(X, Z) = \frac{\text{cov}[X, Z]}{\sigma_X \sigma_Z} = 0$ . Since  $\sigma_X, \sigma_Z > 0$ ,  $\rho(X, Z) = 0$  iff  $\text{cov}(X, Z) = 0$ . However,

$$\begin{aligned}
 \text{cov}(X, Z) &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] \\
 &= \mathbb{E}[X^2Y] - \mathbb{E}[X]\mathbb{E}[XY] \\
 &= \mathbb{E}[X^2]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[X]\mathbb{E}[Y] \\
 &= (\mathbb{E}[X^2] - (\mathbb{E}[X])^2)\mathbb{E}[Y] \\
 &= \sigma_X^2\mathbb{E}[Y],
 \end{aligned}$$

so  $\rho(X, Z) = 0$  if  $\mathbb{E}[Y] = 0$ , and similarly,  $\rho(Y, Z) = 0$  if  $\mathbb{E}[X] = 0$ .

## B.8 Additional Tutorial: Estimate MNM on Iteration 2 of Simulation Study

In this section, we show how to use the R-package *mgm* (Haslbeck & Waldorp, 2020) to fit a moderated network model to  $n = 858$  observations generated from the model shown in Figure 5.3. The *mgm* implements functions to estimate Mixed Graphical Models (MGMs), of which GGMs are a special case. The package can be installed and loaded in the following way:

```
install.packages("mgm")
library(mgm)
```

We will also present different options for visualizing the moderated network model using factor graphs.

### B.8.1 Fit Moderated Network Model to Data

The data set is automatically available in the list object `modnw` when loading the *mgm* package. As specified in Section 5.3.1, the data set contains 13 continuous variables and we sampled 858 observations:

```
> dim(modnw)
[1] 858 13
```

We provide the data in `modnw` to the estimation function `mgm()` of the *mgm* package. Next to the data we specify the types and levels for each variable. Since we model all variables as continuous Gaussian distributions, we specify "g" for each variable and the number of levels as 1 by convention for continuous variables. This specification is necessary in *mgm*, because the package also allows to model Poisson variables and categorical variables with  $k$  categories. Via the argument `moderator` one specifies the moderators to be included in the model. For instance, if we select `moderator = c(3, 7)` all moderation effects of variables 3 and 7 are included in the model. Here we pretend not to know that variable 13 is the only moderator in the model and therefore include all variables as moderators by setting `moderator = 1:13`.

The estimation algorithm uses  $p$  nodewise penalized regressions, for each of which an appropriate regularization parameter  $\lambda_s$  has to be selected (see Section 5.2.5). We select the  $\lambda_s$  that minimizes the EBIC with the hyperparameter  $\gamma = 0.5$  by setting `lambdaSel = "EBIC"` and `lambdaGam = .5`. Alternatively one could select  $\lambda_s$  using cross-validation (`lambdaSel = "CV"`). With `scale = TRUE` we specify that all predictors are scaled to mean zero and SD = 1. This is a standard procedure in regularized regression and avoids that the penalization of a given parameter depends on the standard deviation of the associated variable. With `ruleReg = "AND"` we specify that the nodewise regressions are combined with the AND-rule (see Section 5.2.5).

```
mgm_mod <- mgm(data = modnw,
               type = rep("g", 13),
               level = rep(1, 13),
               moderator = 1:13,
               lambdaSel = "EBIC",
               lambdaGam = 0.5,
               ruleReg = "AND",
               scale = TRUE)
```

The main output is stored in `mgm_mod$interactions`. For a detailed description of the output see the helpfile `?mgm` and the *mgm* paper (Haslbeck & Waldorp, 2020). The list entry `mgm_mod$interactions$indicator` contains a list of all estimated parameters separately for each order (2-way, 3-way, etc.):

```
> mgm_mod$interactions$indicator
[[1]]
[,1] [,2]
[1,]  1  12
[2,]  2   4
[3,]  4  11
[4,]  8  11

[[2]]
[,1] [,2] [,3]
[1,]  1  12  13
[2,]  6   7  13
[3,]  8  10  13
[4,]  8  11  13
```

The first level contains pairwise (2-way) interactions and the second entry contains moderation effects (or 3-way interactions). Thus, in the above output the entry `mgm_mod$interactions$indicator[[1]][2, ]` indicates that there is a nonzero pairwise interaction between variables 2-4. And the entry `mgm_mod$interactions$indicator[[2]][4, ]` indicates that there is a nonzero moderation effect (or 3-way interaction) between variables 8-11-13. In the present model we estimated four pairwise interactions and four moderation effects. To obtain more information about a given interaction we use the function `showInteraction()`. Here is how to obtain the parameter for the pairwise interaction 4-11:

```
> showInteraction(object = mgm_mod, int = c(4,11))
Interaction: 4-11
Weight: 0.103782
Sign: 1 (Positive)
```

The pairwise interaction can be interpreted as in a linear regression: when increasing  $X_4$  by one unit,  $X_{11}$  increases by  $\approx 0.104$  units, when keeping all other variables constant. The parameters for the moderation effects can be obtained similarly: let's say we are interested in the moderation effect 6-7-13. Then we obtain the absolute value and the sign of the parameter via:

```
> showInteraction(object = mgm_mod, int = c(6,7,13))
Interaction: 6-7-13
Weight: 0.1174669
Sign: 1 (Positive)
```

We can interpret this moderation effect in the following way: the pairwise interaction between  $X_6$  and  $X_7$  is zero when  $X_{13}$  is equal to zero. When increasing  $X_{13}$  by one unit, the pairwise interaction between  $X_6$  and  $X_7$  is equal to  $\approx 0.117$ . Similarly, this parameter can be interpreted the moderation effect of  $X_6$  on the pairwise interaction between  $X_7$  and  $X_{13}$ , or the moderation effect of  $X_7$  on the pairwise interaction between  $X_6$  and  $X_{13}$ .

The interpretation is slightly different if a variable is involved in a partially moderated pairwise interaction (or equivalently, in both a 2-way and 3-way interaction). We take variable 12 as an example. We have a pairwise interaction

```
> showInteraction(object = mgm_mod, int = c(1,12))
Interaction: 1-12
Weight: 0.1269736
Sign: 1 (Positive)
```

and a moderation effect:

```
> showInteraction(object = mgm_mod, int = c(1,12,13))
Interaction: 1-12-13
Weight: 0.1467114
Sign: 1 (Positive)
```

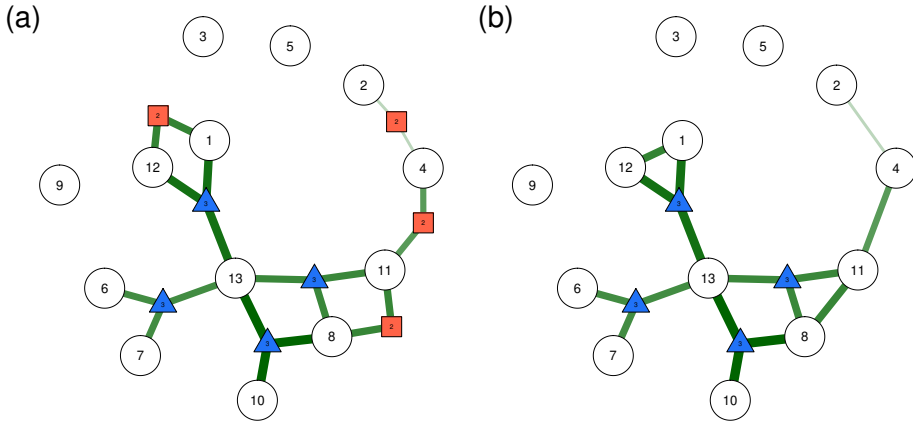
In this case the pairwise interaction between  $X_1$  and  $X_{12}$  is equal to  $\beta_{1,12} \approx 0.130$  if  $X_{13} = 0$ . If  $X_{13}$  increases one unit, then the pairwise interaction between  $X_1$  and  $X_{12}$  increases by  $\approx 0.147$ , so  $\approx 0.130 + 1 \cdot 0.147$ .

## B.8.2 Moderated Network Model as Factor Graph

Moderation effects (3-way interactions) cannot be visualized in a standard graph. However, they can be visualized in a factor graph, which introduces a new node for each interaction parameter. Such a factor graph can be drawn using the `FactorGraph()` function that takes the output of `mgm()` as input:

```
FactorGraph(object = mgm_mod,
PairwiseAsEdge = FALSE)
```

The `FactorGraph()` function plots the graph visualization in Figure B.6 (a):



**Figure B.6:** Two different factor graph visualizations: (a) variable-nodes are displayed as circle nodes, pairwise interactions are displayed as square nodes, and moderation effects (3-way interactions) are displayed as triangles; (b) Only moderation effects (3-way interactions) are displayed as triangle nodes, pairwise interactions are displayed as simple edges. Green edges indicate parameters with positive sign. The widths of edges is proportional to the absolute value of the parameter.

The green (red) edges indicate parameters with positive (negative) sign, and the width of edges is proportional to the absolute value of the parameter. `FactorGraph()` is a wrapper around the `qgraph()` function from the `qgraph` package `epskamp2012qgraph` and all `qgraph()` arguments can be passed to customize the visualization.

In larger graphs with many pairwise interactions this visualization may become unclear. For these situations, the factors representing pairwise interactions can be replaced by simple edges by setting `PairwiseAsEdge = TRUE`. The resulting visualization is shown in Figure B.6 (b). While this graph is not a typical factor graph anymore, the visualization contains the same information as the visualization in (a).

## B.9 Varying Moderation Effects across Nodewise Regressions: A closer Look

In Section 5.4.2.2 we used a data set of three mood variables to illustrate that moderation effects can differ across nodewise regressions, if the data are skewed. Here we provide some intuition for how this is possible, by conditioning on different values of one of the three variables, and show the resulting conditional scatter plots and linear relationships of the remaining two variables.

We first estimate the three conditional distributions of the moderated network model using unregularized linear regression:

```
> lm(afraid~ashamed*distressed, data = msq_p3) # On afraid
Call:
lm(formula = afraid ~ ashamed * distressed, data = msq_p3)
Coefficients:
(Intercept)          ashamed          distressed
-0.0481            0.1194            0.3210
ashamed:distressed
0.1142

> lm(ashamed ~ afraid*distressed, data = msq_p3) # On Ashamed
Call:
lm(formula = ashamed ~ afraid * distressed, data = msq_p3)
Coefficients:
(Intercept)          afraid          distressed
-0.04104            0.16279            0.25054
afraid:distressed
0.08581

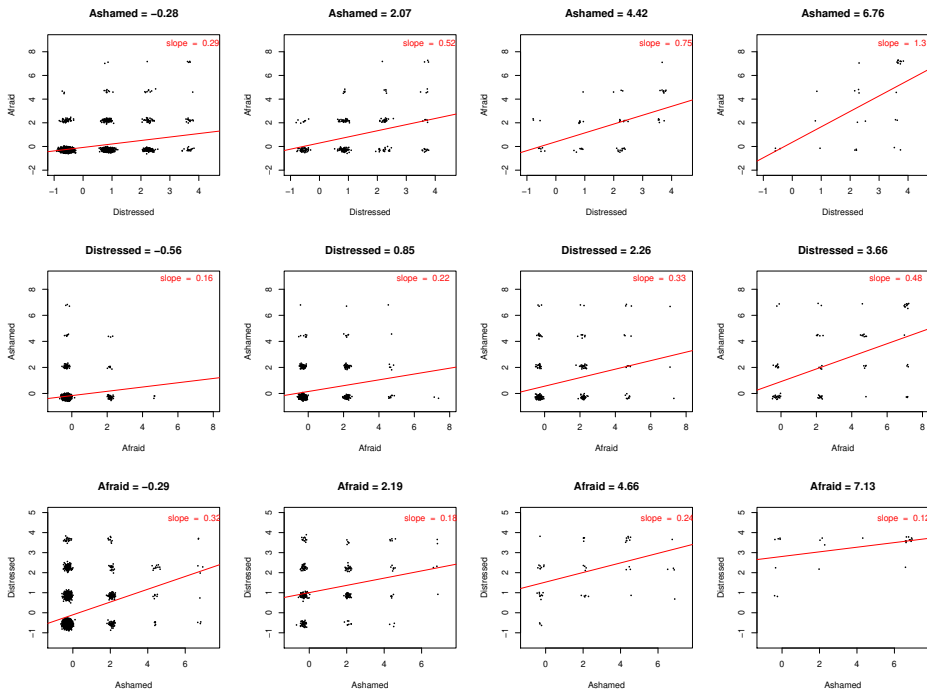
> lm(distressed ~ afraid*ashamed, data = msq_p3) # On Distressed
Call:
lm(formula = distressed ~ afraid * ashamed, data = msq_p3)
Coefficients:
(Intercept)          afraid          ashamed  afraid:ashamed
0.01339            0.40273            0.30110           -0.02984
```

We estimated moderation effects with positive sign in the regressions on *afraid* and *ashamed* and a moderation effect with negative sign in the regression on *distressed*. This is reflecting the results obtained from  $\ell_1$ -regularized regression shown in Figure 5.8 in Section 5.4.2.2.

We visualize the phenomena of different moderation effects across nodewise regressions in Figure B.7 by conditioning on different values of one of the predictors and inspect the linear relationship between the response variable and the remaining predictor. We do this for each of the three regressions:



## B.9. Varying Moderation Effects across Nodewise Regressions: A closer Look



**Figure B.7:** Row 1: The linear relationship between *afraid* and *distressed* depicted by a scatter plot and the best linear fit, for the four different values of *ashamed*. We added some noise in the visualization to capture the amount of data at each combination of *afraid* and *distressed*; row 2: the same visualization as in row 1, however, for the regression on *ashamed*, conditioning on the values of *distressed*; and row 3: the same visualization as in row 1, however, for the regression on *distressed*, conditioning on values of *afraid*.

The first row of Figure B.7 shows the scatter plot of variables *afraid* and *distressed* together with the best fitting regression line (red line) and its slope, for the different values of *ashamed*. To make the density of the data visible we added a small amount of noise to each data point for the visualization. The best regression line was calculated on the original data. We see that the positive linear relationship between *afraid* and *distressed* becomes stronger for larger values of *ashamed*. Similarly in row 2, the positive linear relationship between *ashamed* and *afraid* becomes stronger for larger values of *distressed*. The linear relationship between *afraid* and *distressed* increases more as a function of *ashamed*, than does the linear relationship between *ashamed* and *afraid* as a function of *distressed*. This reflects the fact that the moderation/interaction parameter in the regression on *afraid* is larger than in the regression on *ashamed*.

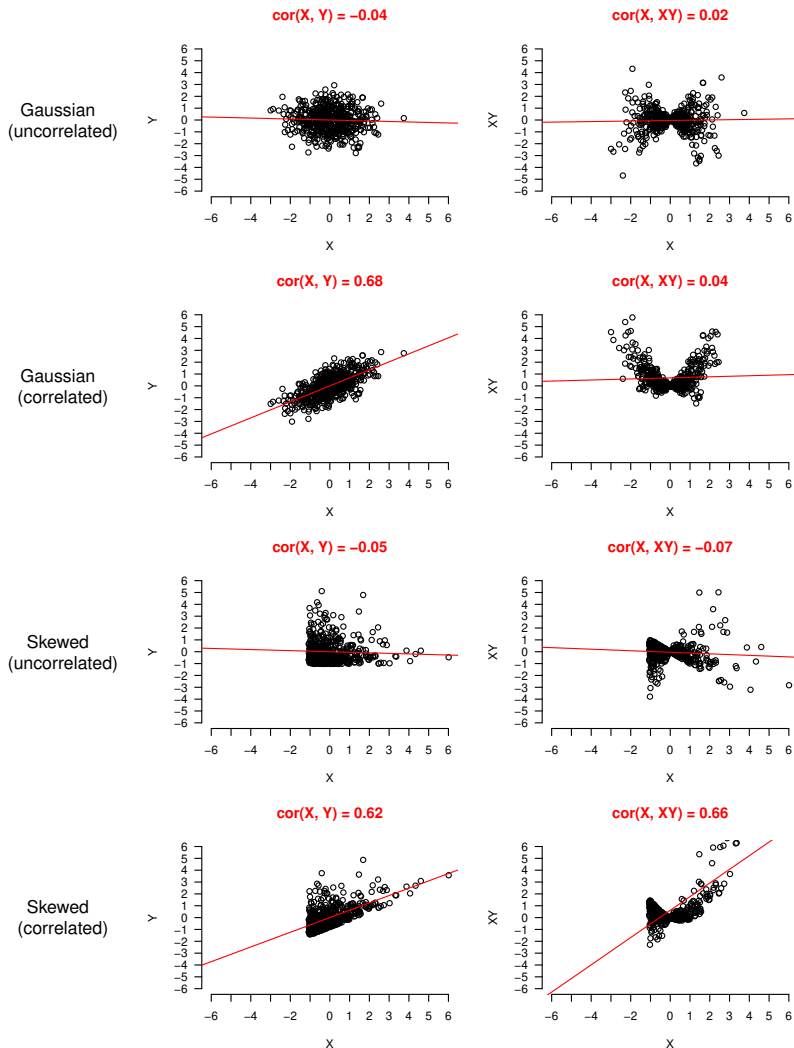
Finally, row 3 shows that the linear relationship between *ashamed* and *distressed* first decreases, then increases, and then decreases again as a function of *afraid*. Thus, we see a non-linear moderation effect of *afraid* on the linear relationship between *ashamed* and *distressed*. It happens to be the case that this non-linear moderation effect is not canceled out exactly, but is best approximated by a small

negative linear moderation effect.

## B.10 Correlations between lower- and higher-order terms

If predictors  $X$  and  $Y$  are centered and independent distributions of any kind, we have  $\text{cor}(X, XY) = 0$ . That is, the lower order terms (singleton predictors such as  $X$ ) and higher order terms (product term such as  $XY$ ) are uncorrelated. We prove this in Appendix B.7. Also, if  $X, Y$  are Gaussian distributions and linearly dependent,  $\text{cor}(X, XY) = 0$ . However, if  $X, Y$  are skewed *and* linearly dependent, we have  $\text{cor}(X, XY) \neq 0$ . We illustrate these four cases (distribution symmetric vs. non-symmetric and dependent vs. independent  $X$  and  $Y$ ) in Figure B.8.

In the first row we generated  $n = 500$  observations from two uncorrelated Gaussian distributions  $X = \mathcal{N}(\mu = 0, \sigma = 1)$  and  $Y = \mathcal{N}(\mu = 0, \sigma = 1)$ . We then mean-centered both distributions. The first column shows the scatter plot, the best fitting regression line (red line) and the correlation  $\text{cor}(X, Y)$ . As we would expect the correlation is close to zero. We then plot  $X$  against  $XY$ . And we see that the correlation  $\text{cor}(X, XY)$  remains close to zero as claimed. In the second row we sampled from two correlated Gaussian distributions  $X = \mathcal{N}(\mu = 0, \sigma = 1)$  and  $Y = X + \mathcal{N}(\mu = 0, \sigma = 1)$  and mean-centered both distributions. As expected, we find a linear relationship between  $X$  and  $Y$ . However, the correlation  $\text{cor}(X, XY)$  remains close to zero. In the third row we sampled from two uncorrelated exponential distributions with rate parameter  $\lambda = 1$ , and mean-centered both distributions. We see that  $\text{cor}(X, Y)$  is close to zero as expected. And also  $\text{cor}(X, XY)$  is zero, as expected. Row four shows the problematic case. We correlated the exponential distributions by adding  $X$  to  $Y$ , and mean-centered both distributions. Now we get a linear relationship between  $X$  and  $Y$  as expected. However, in contrast to the symmetric Gaussian case, this correlation now also implies a nonzero correlation  $\text{cor}(X, XY)$  between  $X$  and  $XY$ .



**Figure B.8:** The correlation between lower- and higher-order terms  $\text{cor}(X, XY)$  for the four combinations of symmetric (Gaussian) / skewed (Exponential) and correlated/uncorrelated (see text for details).

Uncorrelated predictors are a desirable property for two reasons: if correlations between predictors (also called collinearity) are very high, the model becomes unstable in the sense that changing only a few data points can substantially change the parameter estimates. Second,  $\ell_1$ -regularized regression has the property to select only one of several highly correlated predictors. If  $\text{cor}(X, XY)$  is large the  $\ell_1$ -regularized estimator is likely to estimate a nonzero value for one of the two corresponding parameters, even though both parameters are nonzero in the true model. This problem is more severe for large regularization parameters

$\lambda$  (which imply strong regularization) which tend to be selected if the number of observations  $n$  is small.

The problem of correlated predictors can be diagnosed by correlating all lower-order terms with all higher-order terms. If they are highly correlated one solution is to transform the skewed variables towards a symmetric distribution, which reduces the correlation. Popular transformations are taking the log, the square-root or computing the nonparanormal transform (Liu et al., 2009). However, transforming variables always has the disadvantage that a meaningful interpretation of relationships with transformed variables becomes difficult. For example, many people will have little intuition for relationships like "increasing *Feelings of Guilt* by one unit increases *log Mood* by 0.15 units". A solution for the problem of  $\ell_1$ -regularized (LASSO) estimators is to use an estimator that does not use an  $\ell_1$ -penalty. This could for instance be a regularized estimator with an  $\ell_2$ -penalty (Ridge regression). But then we loose the convenient LASSO property of setting small parameters to zero with the result that every network (or factor graph) will be fully connected. Another option would be to use significance-test based estimators. However, then one would have to deal with the problem of multiple testing, which is a serious issues in moderated network models due to the potentially large number of estimated parameters. A detailed investigation of solutions to this particular problem of model misspecifications is beyond the scope though and we leave it for future research.

## B.11 Simulation Results in Tables

In Section 6.3 we reported simulation results in Figures 5.4 and 5.5. Here, we report the same findings in tables. Table B.1 shows the performance in recovering unmoderated pairwise interactions (first row Figure 5.4):

		n											
		30	43	63	92	133	193	280	407	591	858	1245	1808
MNM (1)	SE	0.01	0.00	0.01	0.02	0.06	0.14	0.28	0.59	0.83	0.97	1.00	1.00
MNM (2)	SE	0.01	0.01	0.03	0.04	0.08	0.18	0.36	0.66	0.84	0.98	1.00	1.00
MNM (3)	SE	0.00	0.00	0.00	0.01	0.02	0.06	0.13	0.40	0.70	0.96	1.00	1.00
MNM (1)	PR	0.50	0.33	0.80	0.85	0.92	0.97	0.98	0.98	0.98	0.98	0.98	0.98
MNM (2)	PR	0.50	0.57	0.82	0.90	0.90	0.97	0.98	0.97	0.97	0.95	0.95	0.94
MNM (3)	PR				0.80	0.89	0.97	0.98	1.00	1.00	1.00	0.99	1.00

**Table B.1:** Sensitivity and precision of all compared methods as a function of sample size  $n$ . The missing values for precision indicate that no edges were estimated in 95 or more iterations.

For the different methods, (1) indicates that the correct moderator was specified, (2) that moderators were specified in  $p = 13$  sequential models, and estimates were combined, and (3) indicates that all moderators are specified in a single model (see also Figures 5.4 and 5.5).

Table B.2 shows the results on moderated pairwise interactions shown in the second row of Figure 5.4:

		n											
		30	43	63	92	133	193	280	407	591	858	1245	1808
MNM (1)	SE	0.00	0.01	0.03	0.04	0.08	0.20	0.55	0.86	0.98	1.00	1.00	1.00
MNM (2)	SE	0.01	0.03	0.05	0.06	0.13	0.27	0.60	0.86	0.98	1.00	1.00	1.00
MNM (3)	SE	0.00	0.00	0.01	0.01	0.04	0.10	0.31	0.70	0.94	1.00	1.00	1.00
MNM (1)	PR	0.50	0.33	0.80	0.85	0.92	0.97	0.98	0.98	0.98	0.98	0.98	0.98
MNM (2)	PR	0.50	0.57	0.82	0.90	0.90	0.97	0.98	0.97	0.97	0.95	0.95	0.94
MNM (3)	PR				0.80	0.89	0.97	0.98	1.00	1.00	1.00	0.99	1.00

**Table B.2:** Sensitivity and precision of all compared methods as a function of sample size  $n$ . The missing values for precision indicate that no edges were estimated in 95 or more iterations.

Table B.3 shows the results on moderation effects with pairwise part shown in the first row of Figure 5.5:

		n											
		30	43	63	92	133	193	280	407	591	858	1245	1808
MNM (1)	SE	0.00	0.00	0.00	0.02	0.04	0.14	0.36	0.82	0.99	1.00	1.00	1.00
MNM (2)	SE	0.00	0.00	0.00	0.02	0.05	0.20	0.46	0.79	0.95	0.98	0.98	0.98
MNM (3)	SE	0.00	0.00	0.00	0.01	0.01	0.10	0.25	0.71	0.96	1.00	1.00	1.00
NCT (1)	SE	0.00	0.00	0.00	0.01	0.03	0.10	0.22	0.40	0.67	0.90	1.00	1.00
NCT (2)	SE	0.00	0.00	0.01	0.01	0.02	0.10	0.22	0.41	0.65	0.88	0.99	1.00
FGL (1)	SE	0.00	0.00	0.02	0.06	0.18	0.33	0.71	0.92	0.99	1.00	1.00	1.00
FGL (2)	SE	0.02	0.02	0.06	0.12	0.24	0.43	0.76	0.92	0.98	1.00	1.00	1.00
MNM (1)	PR				1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MNM (2)	PR				1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.98
MNM (3)	PR					1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
NCT (1)	PR					0.71	0.92	0.96	0.94	0.97	0.98	0.98	0.98
NCT (2)	PR			0.25	0.08	0.13	0.32	0.45	0.58	0.49	0.38	0.34	0.33
FGL (1)	PR		0.00	0.43	0.54	0.79	0.85	0.83	0.82	0.74	0.66	0.57	0.59
FGL (2)	PR	0.08	0.15	0.19	0.19	0.32	0.37	0.46	0.48	0.45	0.34	0.28	0.26

**Table B.3:** Sensitivity and precision of all compared methods as a function of sample size  $n$ . The missing values for precision indicate that no edges were estimated in 95 or more iterations.

Table B.4 shows the results on moderation effects without pairwise part shown in the second row of Figure 5.5:

		n											
		30	43	63	92	133	193	280	407	591	858	1245	1808
MNM (1)	SE	0.00	0.00	0.00	0.01	0.04	0.08	0.32	0.70	0.95	1.00	1.00	1.00
MNM (2)	SE	0.00	0.00	0.00	0.02	0.05	0.13	0.32	0.66	0.91	0.96	0.96	0.96
MNM (3)	SE	0.00	0.00	0.00	0.00	0.03	0.06	0.22	0.57	0.88	0.98	1.00	1.00
NCT (1)	SE	0.00	0.00	0.00	0.00	0.00	0.04	0.11	0.35	0.65	0.90	0.98	1.00
NCT (2)	SE	0.00	0.00	0.00	0.00	0.01	0.04	0.12	0.34	0.60	0.90	0.98	0.99
FGL (1)	SE	0.00	0.00	0.00	0.00	0.03	0.02	0.10	0.23	0.53	0.86	1.00	1.00
FGL (2)	SE	0.00	0.01	0.00	0.00	0.03	0.02	0.10	0.24	0.54	0.86	1.00	1.00
MNM (1)	PR				1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MNM (2)	PR				1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.98
MNM (3)	PR					1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
NCT (1)	PR					0.71	0.92	0.96	0.94	0.97	0.98	0.98	0.98
NCT (2)	PR			0.25	0.08	0.13	0.32	0.45	0.58	0.49	0.38	0.34	0.33
FGL (1)	PR		0.00	0.43	0.54	0.79	0.85	0.83	0.82	0.74	0.66	0.57	0.59
FGL (2)	PR	0.08	0.15	0.19	0.19	0.32	0.37	0.46	0.48	0.45	0.34	0.28	0.26

**Table B.4:** Sensitivity and precision of all compared methods as a function of sample size  $n$ . The missing values for precision indicate that no edges were estimated in 95 or more iterations.



---

# TIME-VARYING VAR MODELS

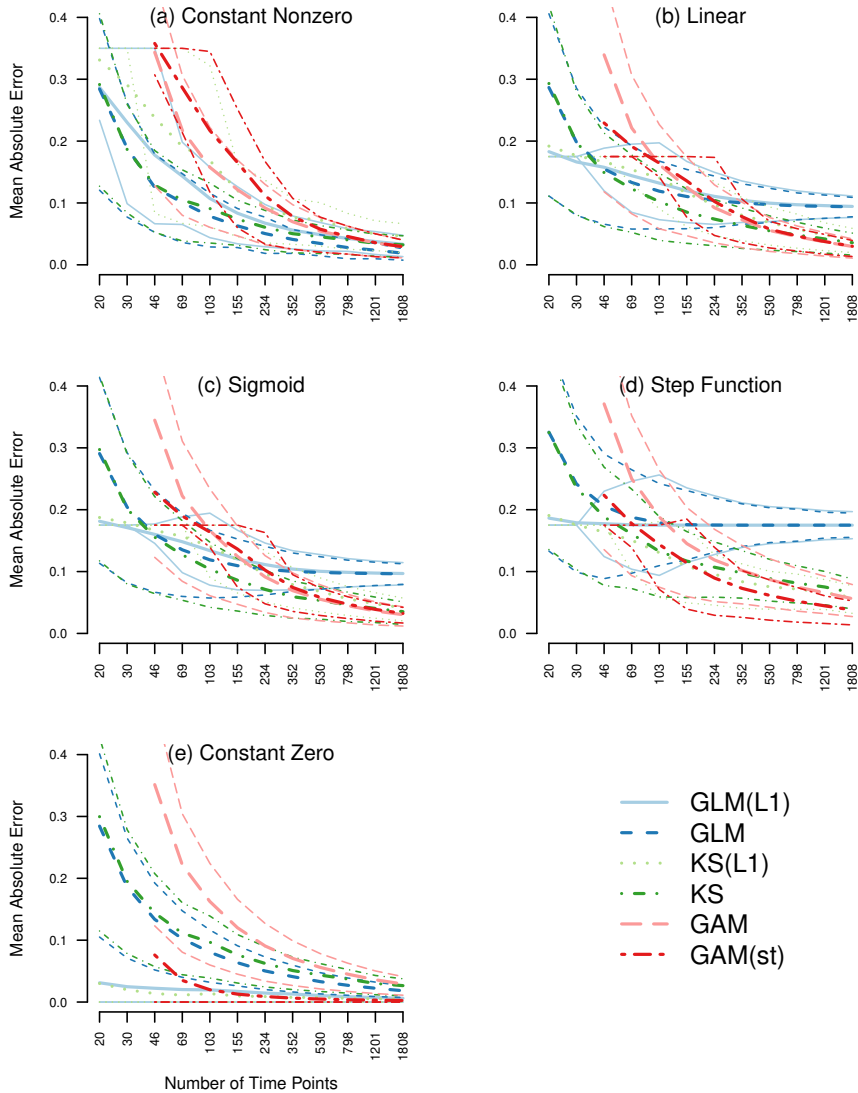
---

## C.1 Sampling Variation around Aggregated Absolute Errors

In Figure 6.5 we reported the mean absolute error, averaged over time points and iterations. These population level mean errors indicate which method has the lowest *expected* error in a given scenario. However, it is also interesting to evaluate how large the population sampling variance is around the mean errors. We therefore display a version of Figure 6.5 that includes the 25% and 75% quantiles of the population sampling distribution.

How can we interpret these quantiles? Let's take the performance of GAM and KS for  $n = 103$  in panel (b) as an example. The population mean error is larger for GAM than for KS in this scenario. Note that this difference in mean errors is on the population level and therefore no test is necessary to judge its significance. However, we see that the sampling distributions of the two errors are largely overlapping. This implies that also the difference of the two errors has a large variance, which means that if  $n = 103$ , it is difficult to predict for a specific sample whether GAM or KS has a larger error.

We see that for unregularized methods the confidence interval is large for small  $n$  and becomes smaller when increasing  $n$ . For the  $\ell_1$ -regularized methods, the quantiles are first small, then increase, and then decrease again as a function of  $n$ . The reason is that for small  $n$ , these methods set all most estimates to zero, and therefore the upper and lower quantiles have the same value. An extreme case is the true zero constant function in Figure C.1 panel (e). Here both quantiles are zero for all  $n$ , while the mean absolute error is larger than 0 and approaches 0 with increasing  $n$ .

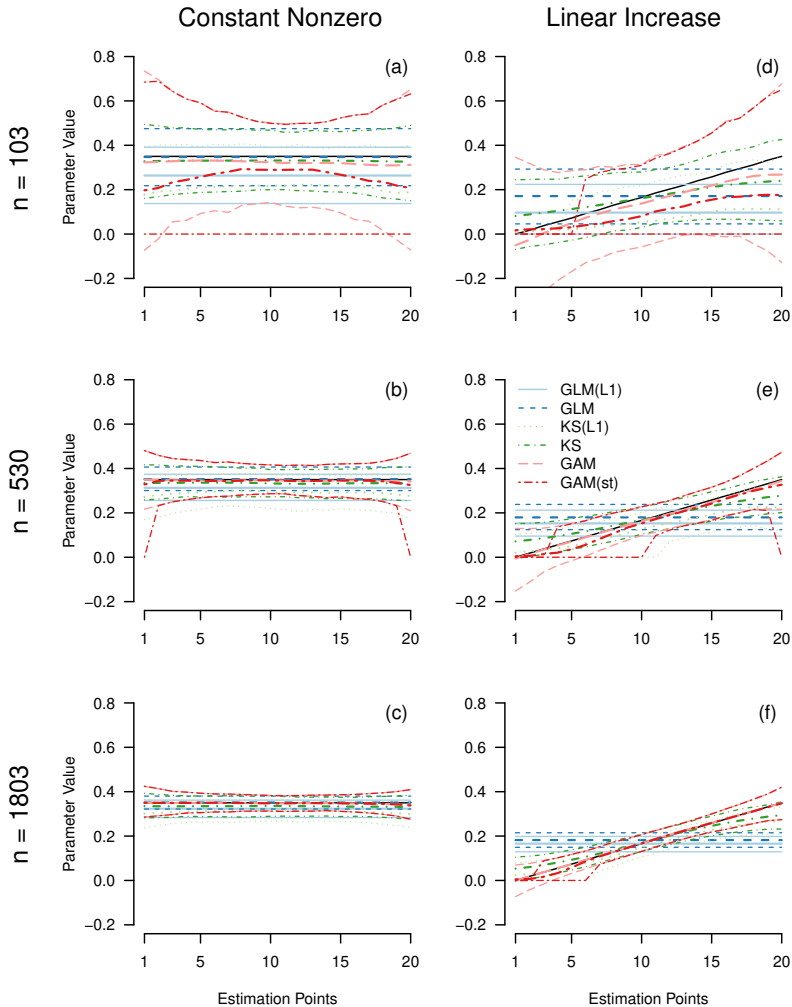


**Figure C.1:** The five panels show the mean absolute estimation error (solid lines) averaged over the same type, time points, and iterations as a function of the number of observations  $n$  on a log scale. We report the error of six estimation methods: stationary unregularized regression (blue), stationary  $\ell_1$ -regularized regression (red), time-varying regression via kernel-smoothing (yellow), time-varying  $\ell_1$ -regularized regression via kernel-smoothing (green), time-varying regression via GAM (pink), and time-varying regression via GAM with thresholding at 95% CI (orange). Some data points are missing because the respective models are not identified in that situation (see Section 6.3.1.2). The dashed lines indicate the 25% and 75% quantiles, averaged over time points.



## C.2 Sampling Variation around Absolute Errors over Time

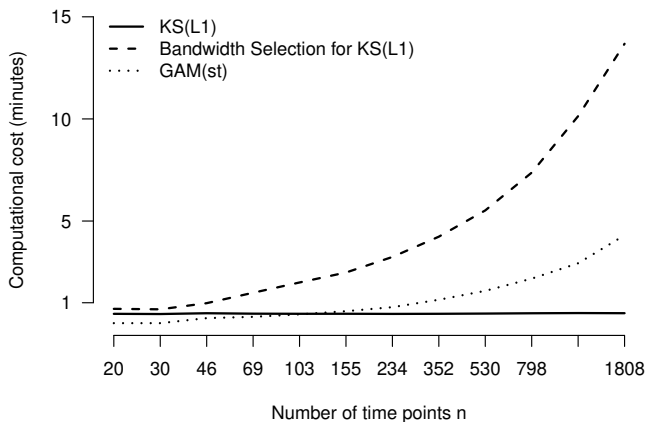
Figure C.2 displays the mean estimates also shown in Figure 6.6 in Section 6.3.1.3, but in addition displays the 10% and 90% quantiles of the estimates. The sampling variance is small for  $n = 103$ , but approaches zero as  $n$  becomes large.



**Figure C.2:** Mean (tick line) and standard deviations (thin line) of estimates for the constant parameter (left column), and the linear increasing parameter (right column), for  $n = 103$  (top row),  $n = 530$  (second row) and  $n = 1803$  (bottom row) averaged over iterations, separately for the five estimation methods: stationary  $\ell_1$ -regularized regression (red), unregularized regression (blue), time-varying  $\ell_1$ -regularized regression via kernel-smoothing (green), time-varying regression via GAM (pink), and time-varying regression via GAM with thresholding at 95% CI (orange).

### C.3 Computational Cost

In Figure C.3 we depict the computational cost of the KS(L1) method versus the GAM(st) method. The computational complexity of the KS(L1) method is  $\mathcal{O}(|E|p \log p|L|)$ , where  $p$  is the number of variables,  $|E|$  is the number of estimation points and  $|L|$  is the number of lags included in the VAR model. The computational complexity for the bandwidth selection is  $\mathcal{O}(|F||F_s|p \log p|L|)$ , where  $|F|$  is the number of folds and  $|F_s|$  the number of time points in the leave-out set of each fold. For details see (Haslbeck & Waldorp, 2020). For the standard GAM function from the *R* package *mgcv* the computational complexity is  $\mathcal{O}(nq^2)$ , where  $n$  is the number of time points modelled, and  $q$  is the total number of coefficients, which increases if the number of basis functions increases (Wood & Augustin, 2002). Note that the credible intervals necessary for thresholding require additional computational cost. Figure C.3 shows the average running time (in minutes) of the two methods as a function of  $n$  in the simulation reported above on a 2.60 GHz processor.



**Figure C.3:** Computational cost in minutes to fit time-varying VAR models with the KS(L1) method (solid line) and the GAM(st) method (dotted line) as a function of observations  $n$ . The dashed line indicates the computational cost for selecting an appropriate bandwidth for the KS(L1) method.

As expected, the computational cost of KS(L1) hardly increases as a function of  $n$ . The computational cost of GAM(st) increases roughly linearly as a function of  $n$ . Also, the cost of the bandwidth selection scheme increases roughly linearly as a function of  $n$ . When considering that KS(L1) requires the data-driven selection of a bandwidth parameter, the computational cost of both method is larger for the KS(L1) method for the current setting of  $p = 10$  variables. However, since the computational complexity of the GAM method includes a quadratic term of the number of parameters, it is likely to perform worse when increasing the number of variables to  $p > 20$ . The KS(L1) method also works for huge number of variables, since its computational complexity only includes  $\log(p)$ .

## C.4 Code to select Appropriate Bandwidth in KS(L1) Method

The function `bwSelect()` fits time-varying VAR models with different bandwidth parameters to a set of training sets and computes the out-of-sample prediction error in the hold-out sets. We then select the bandwidth that minimizes this prediction error across variables and hold-out sets. For details about how these training/test sets are chosen exactly see `?bwSelect` or (Haslbeck & Waldorp, 2020).

Since we fit the time-varying VAR model of our choice repeatedly, we provide all parameters we specified to the estimation function `tvmvar()` as described in Section 6.4.3. In addition, we specify via `bwFolds` the number of training set vs. test set splits, via `bwFoldsize` the size of the test sets, and via `bwSeq` the sequence of candidate bandwidth-values. Here, we chose ten equally spaced values in  $[0.01, 1]$ .

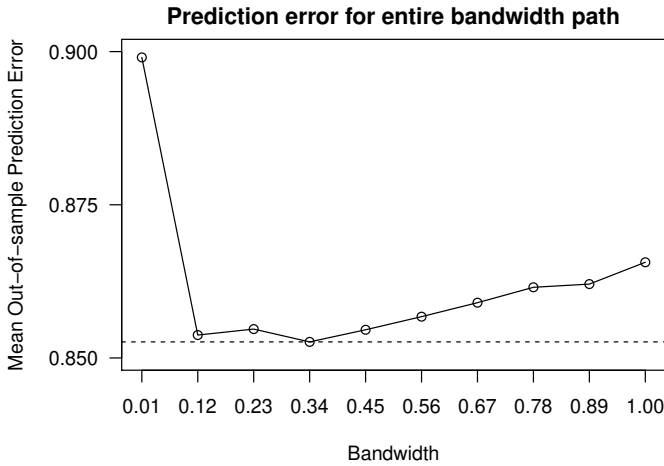
```
bwSeq <- seq(0.01, 1, length = 10)

set.seed(1)
bw_object <- bwSelect(data = mood_data,
                      type = rep("g", 12),
                      level = rep(1, 12),
                      bwSeq = bwSeq,
                      bwFolds = 1,
                      bwFoldsize = 20,
                      modeltype = "mvar",
                      lags = 1,
                      scale = TRUE,
                      timepoints = time_data$time_norm,
                      beepvar = time_data$beepno,
                      dayvar = time_data$dayno,
                      pbar = TRUE)

bandwidth <- bwSeq[which.min(bw_object$meanError)]

[1] 0.34
```

The output object `bw_object` contains all fitted models and unaggregated prediction errors. We see that the bandwidth 0.34 minimized the average out-of-sample prediction error. The full bandwidth path is shown in Figure C.4.



**Figure C.4:** Average out-of-sample prediction error for different bandwidth values obtained from the function. The bandwidth value 0.34 returns the smallest error, indicated by the dashed line.

The bandwidth value of 0.01 is clearly too small, indicated by a large prediction error. The error then tends to become smaller as a function of  $b$  until its minimum at 0.34 and then increases again. Note that if the smallest/largest considered bandwidth value minimizes the error, another search should be conducted with smaller/larger bandwidth values.

## C.5 Estimating time-varying VAR model via GAM(st)

Here we show how to estimate a time-varying VAR model via the GAM(st) method. All analyses are performed using the R-package *tvvarGAM* (Bringmann & Haslbeck, 2017) and the shown code is fully reproducible, which means that the reader can execute the code while reading. The code below can also be found in an R-file on Github: <https://github.com/jmbh/tvvar-paper>.

### C.5.1 Load R-packages and dataset

Similar to Section 6.4.2 we load the dataset from the *mgm* package, and subset the 12 mood related variables. In addition, we load the *tvvarGAM* package (version 0.1.1).

```
library(mgm) # Version 1.2-8

mood_data <- as.matrix(symptom_data$data[, 1:12]) # Subset variables
mood_labels <- symptom_data$colnames[1:12] # Subset variable labels
colnames(mood_data) <- mood_labels
time_data <- symptom_data$data_time
```

```
# Install from Github:
library(devtools)
install_github("LauraBringmann/tvvarGAM")
library(tvvarGAM)
```

## C.5.2 Estimating time-varying VAR model

We use the function `tvvarGAM()` to estimate the time-varying VAR model. We provide the data via the `data` argument and provide an integer vector of length  $n$  indicating the successiveness of measurements by specifying the number of the recorded notification and the day number via the arguments `beepvar` and `dayvar`. The latter is used similarly as in the `mgm` package to compute the VAR design matrix. Via the argument `nb` we specify the number of desired basis functions (see Section 6.2.2). First, we estimated the model with 10 basis functions. However, because some of the edf of the smooth terms were close to 10, we doubled the number of basis functions (see discussion in Section 6.2.2).

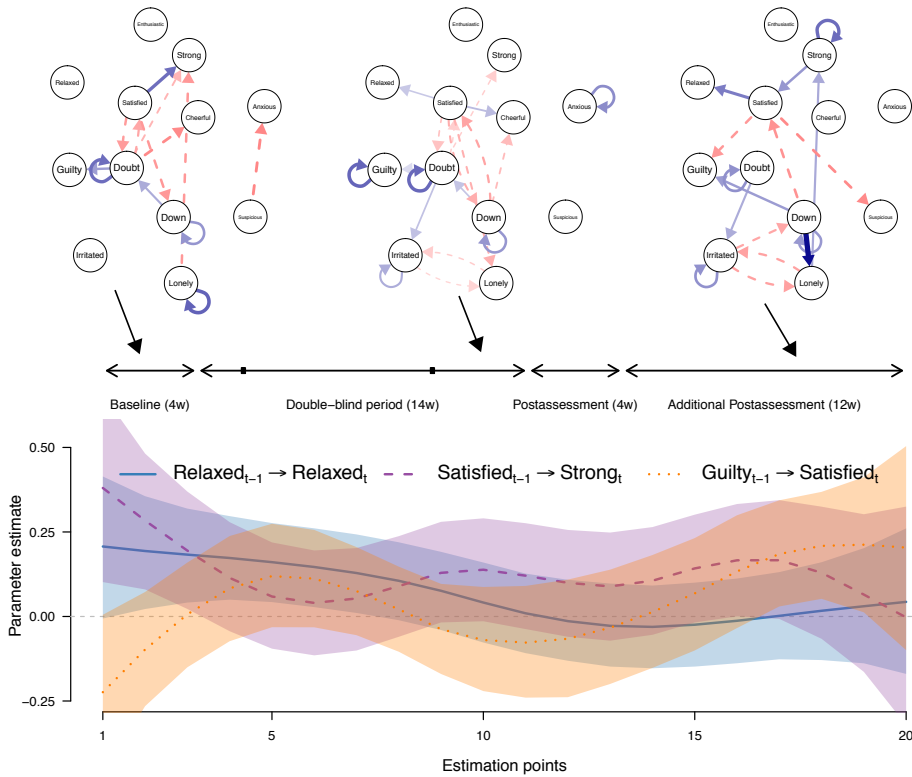
```
tvvargam_obj <- tvvarGAM(data = mood_data,
                        beepvar = time_data$beepno,
                        dayvar = time_data$dayno,
                        nb = 20,
                        scale = TRUE)
```

The output object consists of a list with three entries:

`tvvargam_obj$Results_GAM$Estimate` is a  $(p + 1) \times p \times \text{timepoints}$  array that contains the parameter estimate at each time point. The first row contains the estimated intercepts. The two other list entries have the same dimensions and contain the 5% and 95% confidence intervals for the estimates in `tvvargam_obj$Results_GAM$Estimate`. Thus, in case of the `tvvarGAM` package no separate resampling scheme is necessary in order to get a measure for the reliability of parameters.

## C.5.3 Visualize time-varying VAR model

Figure C.5 visualizes the part of the time-varying VAR like Figure 6.10 above, however, now with the estimates from the `tvvarGAM` package. Notice that for visualization purposes we used the thresholded version of the time-varying VAR, thus showing only the arrows that are significant ( $p\text{-value} < 0.05$ ).



**Figure C.5:** Top row: visualization of thresholded VAR models at estimation points 2, 10 and 18, estimated with the spline-based method. Blue arrows indicate positive relationships, red arrows indicate negative relationships, and the width of the arrows is proportional to the absolute value of the corresponding parameter. The self-loops indicate autocorrelations. Bottom row: three parameters plotted as a function of time; the points are unthresholded point estimates, the shading indicates the 5% and 95% credible intervals at each estimation point.

Similarly to the analysis performed with the KS(L1) method we visualize the VAR parameters at estimation points 2, 10 and 18 (top row Figure C.5). We see that less edges are present than in the results of the KS(L1) method, which indicates that the GAM(ks) method is more conservative. The bottom row of Figure C.5 shows a line plot of the same three parameters as in the analysis with the KS(L1) method. We see that the effect of Relaxed on itself tends to decrease over the measured time interval, which is consistent with the results of the KS(L1) method. However, results of the cross-lagged effects of Satisfied on Strong, and of Guilty on Satisfied are only consistent with the results of the KS(1) method in the middle of the time series. The largest difference between the two methods is the increase of the effect of Guilty on Satisfied is noteworthy, while the KS(L1) method estimates a decrease. It seems that the GAM(st) estimates in the second half of the time series are incorrect, because because if one splits the time series

in half and estimates two unregularized stationary VAR models, then the effect of Guilty on Satisfied is clearly negative in the second half of the time series. In general, the large changes and the much larger credible intervals at the beginning and the end of the time series indicate that the estimates are very unstable in those regions. This is consistent with the high standard deviation of estimates of the GAM and GAM(st) method shown in Figure C.2. The code to fully reproduce Figure C.5 is not shown here due to its length, but can be obtained from Github [https://github.com/jmbh/tvvar\\_paper](https://github.com/jmbh/tvvar_paper).

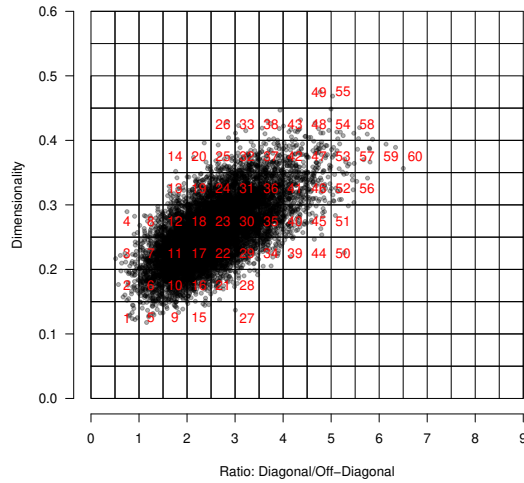




# SELECTING BETWEEN AR AND VAR MODELS

## D.1 Sampling cells on the $R \times D$ grid

Figure D.1 shows the  $R$  and  $D$  values of the 10000 VAR models sampled from the mixed VAR model estimated from the “MindMaastricht” data (Geschwind et al., 2011):



**Figure D.1:**  $R$  and  $D$  values for the initially sampled 10000 VAR models.

Each point in Figure D.1 represents one of the 10000 VAR models we initially sampled from the mixed model. We see that there are 60 cells in which at least one model has been sampled. We then discarded these initial models and sampled from the mixed model until each of the 60 cells was filled with 100 models. We used these  $60 \times 100 = 6000$  in the simulation study reported in the main text.



---

# THE INPUT MATTERS: INTERPRETING THE ISING MODEL

---

## E.1 Statistical Equivalence worked out for two variable example

Here we show that the two models shown in Figure 8.1 are statistically equivalent. Two models statistically equivalent if they output the same probability for any of states on which the models are defined.

We begin with the model estimated on the domain  $\{-1, -1\}$ . We first compute the potentials for the four states  $\{(-1, -1), (-1, 1), (1, -1), (1, 1)\}$ :

$$\exp\{0.318(-1) + 0.318(-1) + 0.193(-1)(-1)\} = 0.6415304$$

$$\exp\{0.318(-1) + 0.318(1) + 0.193(-1)(1)\} = 0.8248249$$

$$\exp\{0.318(1) + 0.318(-1) + 0.193(1)(-1)\} = 0.8248249$$

$$\exp\{0.318(1) + 0.318(1) + 0.193(1)(1)\} = 2.29118$$

and then the normalization constant

$$Z = 0.6415304 + 0.8248249 + 0.8248249 + 2.29118 = 4.58236$$

We divide the potentials by  $Z$  and obtain the probabilities

$$P(Y_1 = -1, Y_2 = -1) = \frac{0.6415304}{Z} = 0.14$$

$$P(Y_1 = -1, Y_2 = 1) = \frac{0.8248249}{Z} = 0.18$$

$$P(Y_1 = 1, Y_2 = -1) = \frac{0.8248249}{Z} = 0.18$$

$$P(Y_1 = 1, Y_2 = 1) = \frac{2.29118}{Z} = 0.5$$

We now repeat the same with domain  $\{0, 1\}$  and first compute the potentials for the states  $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ :

$$\exp\{0.251(0) + 0.251(0) + 0.77(0)(0)\} = 1$$

$$\exp\{0.251(0) + 0.251(1) + 0.77(0)(1)\} = 1.285714$$

$$\exp\{0.251(1) + 0.251(0) + 0.77(1)(0)\} = 1.285714$$

$$\exp\{0.251(1) + 0.251(1) + 0.77(1)(1)\} = 3.571429$$

and then the normalization constant

$$Z = 1 + 1.285714 + 1.285714 + 3.571429 = 7.142857$$

We divide the potentials by  $Z$  and obtain the probabilities

$$P(X_1 = 0, X_2 = 0) = \frac{1}{Z} = 0.14$$

$$P(X_1 = 0, X_2 = 1) = \frac{1.285714}{Z} = 0.18$$

$$P(X_1 = 1, X_2 = 0) = \frac{1.285714}{Z} = 0.18$$

$$P(X_1 = 1, X_2 = 1) = \frac{3.571429}{Z} = 0.5$$

We see that both models predict the same probabilities and are therefore statistically equivalent.

## **E.2 Increasing interaction parameters only changes the marginal probabilities domain in $\{0, 1\}$**

Here we show that for an Ising model with  $p = 2$  variables with  $\alpha_1, \alpha_2 = 0$  and  $\beta_{12} > 0$  it holds that

$$P(X_1 = -1) = P(X_2 = -1) = P(X_2 = 1) = P(X_1 = 1) \tag{E.1}$$

for the domain  $\{-1, 1\}$ , and that

$$P(X_1 = 0) = P(X_2 = 0) < P(X_1 = 1) = P(X_2 = 1) \tag{E.2}$$

for the domain  $\{0, 1\}$ .

We first show (E.1). We assume  $\alpha_1, \alpha_2 = 0$  and  $\beta_{12} > 0$ . Then the Ising model is given by

$$\begin{aligned} P(X_1, X_2) &= \frac{1}{Z} \exp\{\alpha_1 X_1 + \alpha_2 X_2 + \beta_{12} X_2 X_1\} \\ &= \frac{1}{Z} \exp\{\beta_{12} X_2 X_1\}, \end{aligned}$$

where  $Z$  is the normalizing constant summing over all  $2^p = 4$  states. We calculate the probability of the four possible states:

$$\begin{aligned} P(X_1 = 1, X_2 = -1) &= \frac{1}{Z} \exp\{-\beta_{12}\}, \\ P(X_1 = 1, X_2 = 1) &= \frac{1}{Z} \exp\{\beta_{12}\}, \\ P(X_1 = -1, X_2 = -1) &= \frac{1}{Z} \exp\{\beta_{12}\}, \\ P(X_1 = -1, X_2 = 1) &= \frac{1}{Z} \exp\{-\beta_{12}\}. \end{aligned}$$

And average over the state of  $X_2$  to obtain the marginals probabilities  $P(X_1)$ :

$$P(X_1 = 1) = P(X_1 = 1, X_2 = -1) + P(X_1 = 1, X_2 = 1) = \frac{1}{Z} \exp\{-\beta_{12}\} + \frac{1}{Z} \exp\{\beta_{12}\}$$

$$P(X_1 = -1) = P(X_1 = -1, X_2 = -1) + P(X_1 = -1, X_2 = 1) = \frac{1}{Z} \exp\{\beta_{12}\} + \frac{1}{Z} \exp\{-\beta_{12}\}$$

We see that  $P(X_1 = 1) = P(X_1 = -1)$ . By symmetry the same is true for  $X_2$ , which proves our claim.

We next prove (E.2). We again assume  $\alpha_1, \alpha_2 = 0$  and  $\beta_{12} > 0$  and calculate the probabilities of the four possible states:

$$\begin{aligned} P(X_1 = 1, X_2 = 0) &= \frac{1}{Z} \exp\{0\}, \\ P(X_1 = 1, X_2 = 1) &= \frac{1}{Z} \exp\{\beta_{12}\}, \\ P(X_1 = 0, X_2 = 0) &= \frac{1}{Z} \exp\{0\}, \\ P(X_1 = 0, X_2 = 1) &= \frac{1}{Z} \exp\{0\}. \end{aligned}$$

The marginal probabilities  $P(X_1)$  are:

$$P(X_1 = 1) = P(X_1 = 1, X_2 = 0) + P(X_1 = 1, X_2 = 1) = \frac{1}{Z} \exp\{0\} + \frac{1}{Z} \exp\{\beta_{12}\}$$

$$P(X_1 = 0) = P(X_1 = 0, X_2 = 0) + P(X_1 = 0, X_2 = 1) = 2 \frac{1}{Z} \exp\{0\}$$

Since  $\exp\{\beta_{12}\} > \exp\{0\}$ , we have  $P(X_1 = 0) > P(X_1 = 1)$ , if  $\beta_{12} > 0$ . By symmetry the same is true for  $X_2$ , which proves our claim.

Note that if we assume  $\beta_{12} < 0$ , (E.1) holds again for  $\{-1, 1\}$ , while for  $\{0, 1\}$  we have

$$P(X_1 = 0) = P(X_2 = 0) > P(X_2 = 1) = P(X_1 = 1)$$

instead.

### E.3 Derivation of Transformation from $\{0, 1\}$ to $\{-1, 1\}$ and vice versa

In this section, we first introduce the Ising model for  $p$  variables with domain  $\{-1, 1\}$ , which is the domain used in physics applications. Next, we introduce the Ising model for  $p$  variables with domain  $\{0, 1\}$ , which is mostly used in the statistics literature. We connect both models by deriving a formula of the parameters of one parameterization as a function of the parameters of the other parameterization. This allows us to transform the parameterization based on domain  $\{-1, 1\}$  into the parameterization of domain  $\{0, 1\}$  and vice versa.

In the physics domain, variables can take on values in  $\{-1, 1\}$ . The probability distribution of the Ising model for  $p$  such random variables is specified by

$$p(y) = \frac{\exp\left(\sum_{i=1}^p \alpha_i y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij} y_i y_j\right)}{\sum_y \exp\left(\sum_{i=1}^p \alpha_i y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij} y_i y_j\right)}, \quad (\text{E.3})$$

where  $y, y \in \{-1, 1\}^p$ , denotes a configuration of the  $p$  random variables, and the sum  $\sum_y$  in the denominator denotes a sum that ranges over all  $2^p$  possible configurations or realizations of  $y$ .

From a statistical perspective, the Ising model is a model that is completely determined by the spin variables' main effects and their pairwise interactions. A spin variable in the network tends to have a positive value ( $y_i = 1$ ) when its main effect is positively valued ( $\alpha_i > 0$ ), and tends to have a negative value ( $y_i = -1$ ) when its main effect is negatively valued ( $\alpha_i < 0$ ). Furthermore, any two variables  $y_i$  and  $y_j$  in the network tend to align their values when their interaction effect is positive ( $\beta_{ij} > 0$ ), and tend to be in different states when their interaction effect is negative ( $\beta_{ij} < 0$ ).

In statistical applications, the Ising model is typically used to describe the probability distribution of  $p$  binary random variables,

$$p(x) = \frac{\exp\left(\sum_{i=1}^p \alpha_i^* x_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* x_i x_j\right)}{\sum_x \exp\left(\sum_{i=1}^p \alpha_i^* x_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* x_i x_j\right)}, \quad (\text{E.4})$$

where  $x, x \in \{0, 1\}^p$ , denotes a configuration of the  $p$  binary random variables, and again we use  $\sum_x$  to denote the sum that ranges over all  $2^p$  possible configurations or realizations of  $x$ .

Even though the model is again completely determined by main effects and pairwise interactions, its interaction parameters  $\beta^*$  carry a different interpretation than the interaction parameters of the Ising model for variables  $Y$  in the  $\{-1, 1\}$  domain. Here, two binary variables  $x_i$  and  $x_j$  in the network tend to both equal one ( $x_i x_j = 1$ ) when their interaction effect is positive ( $\beta_{ij}^* > 0$ ), but their product tends to equal zero ( $x_i x_j = 0$ ) when their interaction effect is negative ( $\beta_{ij}^* < 0$ ). That is, whenever the interaction between two binary variables  $x_i$  and  $x_j$  in the network is negative ( $\beta_{ij} < 0$ ), they tend to be in one of the states  $\{0, 0\}$ ,  $\{0, 1\}$  or  $\{1, 0\}$ .

Despite the different interpretations of the two Ising model formulations, one can traverse the two specifications by a simple change of variables. To wit, assume that we have obtained an Ising model for  $p$  binary variables  $p(x)$  and wish to express its solution in terms of the variables in the  $\{-1, 1\}$  domain, then we require the change of variables

$$x_i = \frac{1}{2}(y_i + 1) \text{ with inverse relation } y_i = 2x_i - 1. \quad (\text{E.5})$$

We use this transformation in the distribution of the binary random variables,

$$\begin{aligned} p(x) &= \frac{\exp\left(\sum_{i=1}^p \alpha_i^* x_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* x_i x_j\right)}{\sum_x \exp\left(\sum_{i=1}^p \alpha_i^* x_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* x_i x_j\right)} \\ &= \frac{\exp\left(\sum_{i=1}^p \alpha_i^* \frac{1}{2}(y_i + 1) + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* \frac{1}{2}(y_i + 1) \frac{1}{2}(y_j + 1)\right)}{\sum_y \exp\left(\sum_{i=1}^p \alpha_i^* \frac{1}{2}(y_i + 1) + \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* \frac{1}{2}(y_i + 1) \frac{1}{2}(y_j + 1)\right)} = p(y), \end{aligned} \quad (\text{E.6})$$

and observe that this transformation affects both main effects and pairwise interactions. Working out the sum over pairs of variables, we find

$$\begin{aligned}
 \sum_{i=1}^{p-1} \sum_{j>i}^p \beta_{ij}^* \frac{1}{2}(y_i + 1) \frac{1}{2}(y_j + 1) &= \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* (y_i y_j + y_i + y_j + 1) \\
 &= \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* y_i y_j + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* y_j \\
 &\quad + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* \\
 &= \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* y_i y_j + \sum_{i=1}^p \sum_{\substack{j=1 \\ j \neq i}}^p \frac{1}{4} \beta_{ij}^* y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* \\
 &= \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^* y_i y_j + \sum_{i=1}^p \frac{1}{4} \beta_{i+}^* y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4} \beta_{ij}^*, \quad (E.7)
 \end{aligned}$$

where the first term reflects pairwise interactions between the variables  $y$ , the second term reflects main effects of the variables with main effect  $\beta_{i+}^* = \sum_{j=1}^p \beta_{ij}^*$ , and the last term is constant with respect to (w.r.t.) the variables  $y$ . Similarly, we can express the sum over the main effects as

$$\sum_{i=1}^p \alpha_i^* \frac{1}{2}(y_i + 1) = \sum_{i=1}^p \alpha_i^* \frac{1}{2} y_i + \sum_{i=1}^p \alpha_i^* \frac{1}{2}, \quad (E.8)$$

where the last term is again constant w.r.t. the variables  $y$ . Collecting the main effects,

$$\sum_{i=1}^p \frac{1}{2} \alpha_i^* y_i + \sum_{i=1}^p \frac{1}{4} \beta_{i+}^* y_i = \sum_{i=1}^p \left( \frac{1}{2} \alpha_i^* + \frac{1}{4} \beta_{i+}^* \right) y_i, \quad (E.9)$$

and constant terms,

$$C = \sum_{i=1}^p \frac{1}{2} \alpha_i^* + \sum_{i=1}^p \sum_{j=1}^p \frac{1}{4} \beta_{ij}^*, \quad (E.10)$$

we obtain:



$$\begin{aligned}
 p(y) &= \frac{\exp\left(\sum_{i=1}^p \left(\frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*\right)y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4}\beta_{ij}^*y_iy_j + C\right)}{\sum_y \exp\left(\sum_{i=1}^p \left(\frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*\right)y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4}\beta_{ij}^*y_iy_j + C\right)} \\
 &= \frac{\exp\left(\sum_{i=1}^p \left(\frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*\right)y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4}\beta_{ij}^*y_iy_j\right)}{\sum_y \exp\left(\sum_{i=1}^p \left(\frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*\right)y_i + \sum_{i=1}^{p-1} \sum_{j>i}^p \frac{1}{4}\beta_{ij}^*y_iy_j\right)}, \tag{E.11}
 \end{aligned}$$

which is equal to the Ising model for variables in the  $\{-1, 1\}$  domain when we write  $\alpha_i = \frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*$  and  $\beta_{ij} = \frac{1}{4}\beta_{ij}^*$ . In a similar way, one can obtain the parameter values of the binary case from a solution of the Ising model for variables in the  $\{-1, 1\}$  domain using  $\alpha_i^* = 2\alpha_i - 2\beta_{i+}$  and  $\beta_{ij}^* = 4\beta_{ij}$ . Thus, we can obtain the binary Ising model parameters  $\alpha^*$  and  $\beta^*$  from a simple transformation of the  $\{-1, 1\}$  coded Ising model parameters  $\alpha$  and  $\beta$ , and vice versa. Table E.1 summarizes these transformations:

Transformation	$\alpha$	$\beta$
$\{0, 1\} \Rightarrow \{-1, 1\}$	$\alpha_i = \frac{1}{2}\alpha_i^* + \frac{1}{4}\beta_{i+}^*$	$\beta_{ij} = \frac{1}{4}\beta_{ij}^*$
$\{-1, 1\} \Rightarrow \{0, 1\}$	$\alpha_i^* = 2\alpha_i - 2\beta_{i+}$	$\beta_{ij}^* = 4\beta_{ij}$

**Table E.1:** Transformation functions to obtain the threshold and interaction parameters in one parameterization from the threshold and interaction parameters in the other parameterization. Parameters with asterisk indicate parameters in the  $\{0, 1\}$  domain.

## E.4 Model equivalence across domains with penalized estimation

If one estimates the Ising model with an unbiased estimator, one can estimate with domain  $\{0, 1\}$  and obtain by transformation the estimates one would have obtained by estimating with domain  $\{-1, 1\}$  (and vice versa). In this section we ask whether this is also the case for penalized estimation, which is a popular way to estimate the Ising model (e.g., Van Borkulo et al., 2014; Ravikumar, Wainwright, Lafferty, et al., 2010).

In penalized estimation, the likelihood is maximized with respect to a constraint  $c$ , typically on the  $\ell_1$ -norm of the vector of interaction parameters  $\beta_{ij}$

$$\sum_{i=1}^p \sum_{\substack{j=1 \\ j \neq i}}^p |\beta_{ij}| < c.$$

Estimation with an  $\ell_1$ -penalty is attractive because it sets small parameter estimates to zero, which makes it easier to interpret the model. The key problem in this setting is selecting an appropriate constraint  $c$ . A popular approach is to consider a sequence of candidate constraints  $C = \{c_1, \dots, c_k\}$  and select the  $c_i$  that minimizes the Extended Bayesian Information Criterion (EBIC) (Foygel & Drton, 2010), which extends the BIC (Schwarz et al., 1978) by an additional penalty (weighted by  $\gamma$ ) for the number of nonzero interaction parameters

$$\text{EBIC}_{c_i} = -2LL_{c_i} + s_0 \log n + 4s_0\gamma \log p,$$

where  $LL_{c_i}$  is the maximized log-likelihood under constraint  $c_i$ ,  $s_0$  is the number of nonzero interaction parameters,  $n$  is the number of observations and  $p$  the number of estimated interaction parameters.

We are interested in whether selecting models with this procedure in the two domains,  $\{0, 1\}$  and  $\{-1, 1\}$ , leads to statistically equivalent models. This is indeed the case for the following reason: assume that  $c^*$  minimizes the EBIC for domain  $\{0, 1\}$ , then from the transformation in Table 8.2,  $\frac{c^*}{4}$  should give the lowest EBIC in domain  $\{-1, 1\}$ , because the constraint  $\|\beta^*\|_1 < c^*$  on  $\{0, 1\}$  is equivalent to the constraint  $\|\beta^*\|_1 < \frac{c^*}{4}$  on  $Y$ . Thus, if  $\frac{c^*}{4}$  is included in the candidate set  $C$ , when estimating in domain  $\{-1, 1\}$ , two statistically equivalent models should be selected. Note that exactly  $\frac{c^*}{4}$  has to be included, because a slightly larger/smaller constraint can lead to a very different model, if the number of nonzero parameter changes. This nonlinearity arises from the EBIC, in which  $s_0$  decreases by 1 (large change) if some parameter with a tiny value (e.g. 0.0001) is set to zero (small change). Therefore, in order to ensure statistically equivalent models one would need to search a dense sequence  $C$ . Clearly, this is unfeasible in practice. This means that, in practice  $\ell_1$ -regularized estimation can return models from domains  $\{0, 1\}$  and  $\{-1, 1\}$  that are not statistically equivalent. We leave the task of investigating this issue for different estimation algorithms for future research. In what follows we provide an extended version of this argument.

We define:

$$\begin{aligned} c^* &= \arg_{c \in C} \min \text{EBIC}_c \\ &= \arg_{c \in C} \min -2 \log \left[ \frac{1}{Z} \prod_{m=1}^n \exp \left\{ \sum_{i=1}^p \alpha_i^* X_i + \sum_{\substack{i=1 \\ j \neq i}}^p \sum_{\substack{j=1 \\ j \neq i}}^p \beta_{ij}^* X_i X_j \right\} \right] \\ &\quad + s_0 \log n + 4s_0\gamma \log[p(p-1)/2], \end{aligned}$$

with constraint

$$\sum_{i=1}^p \sum_{\substack{j=1 \\ j \neq i}}^p |\beta_{ij}^*| < c,$$

where  $s_0$  is the number of nonzero interaction parameters,  $n$  is the sample size,  $p$  is the number of variables  $\frac{p(p-1)}{2}$  is the total number of interaction parameters, and  $\gamma$  is a tuning parameter.

Now, we would like to show that if  $c^*$  minimizes the EBIC in domain  $\{0, 1\}$ , then  $4c^*$  minimizes the EBIC in  $\{-1, 1\}$ .

We use the transformation in Table 8.2 to rewrite the EBIC into the parameterization implied by  $\{-1, 1\}$ :

$$\begin{aligned}
 c^* &= \arg_{c \in C} \min \text{EBIC}_c \\
 &= \arg_{c \in C} \min -2 \log \left[ \frac{1}{Z} \prod_{m=1}^n \exp \left\{ \sum_{i=1}^p \left( \frac{1}{2} \alpha_i^* + \frac{1}{4} \sum_{\substack{j=1 \\ j \neq i}}^p \beta_{ij}^* \right) X_i + \sum_{i=1}^p \sum_{\substack{j=1 \\ j \neq i}}^p \frac{1}{4} \beta_{ij}^* X_i X_j \right\} \right] \\
 &\quad + s_0 \log n + 4s_0 \gamma \log[p(p-1)/2],
 \end{aligned}$$

with constraint

$$\sum_{i=1}^p \sum_{\substack{j=1 \\ j \neq i}}^p \frac{1}{4} |\beta_{ij}^*| < c^*.$$

We can rewrite the constraint into

$$\sum_{i=1}^p \sum_{\substack{j=1 \\ j \neq i}}^p |\beta_{ij}^*| < 4c^*.$$

The last inequality shows that the constraint is 4 times larger for the parameterization in domain  $\{0, 1\}$ . Or the other way around, the constraint is  $\frac{1}{4}$  times smaller in  $\{-1, 1\}$  compared to  $\{0, 1\}$ .

We know that the models are statistically equivalent across domains. Therefore, the likelihood of the model with constraint  $c$  in domain  $\{0, 1\}$  is equal to the likelihood of the model with constraint  $\frac{c}{4}$  in domain  $\{-1, 1\}$ . Now, since the transformation never changes a zero estimate in a nonzero estimate or vice versa with probability 1, also the terms  $s_0 \log n + 4s_0 \gamma \log[p(p-1)/2]$  in the EBIC remain constant across domains. It follows that, if  $c^* = \arg_{c \in C} \min \text{EBIC}_c$  in domain  $\{0, 1\}$ , then  $\frac{c^*}{4} = \arg_{c \in C} \min \text{EBIC}_c$  in domain  $\{-1, 1\}$ .



---

# RECOVERING BISTABLE SYSTEMS FROM TIME SERIES DATA

---

## F.1 Determine Fixed Points of Bistable System

In this section we show how to compute the fixed points of the deterministic part of our model, which we report in Section 9.2.1. The fixed points of a set of differential equations is found by setting all equations to zero and solving that system. In our case this means solving the nonlinear system of equations:

$$\begin{aligned}
 0 &= r_1 x_1 + \sum_{j=1}^4 C_{1,j} x_j x_1 + a_1 \\
 0 &= r_2 x_2 + \sum_{j=1}^4 C_{2,j} x_j x_2 + a_2 \\
 0 &= r_3 x_3 + \sum_{j=1}^4 C_{3,j} x_j x_3 + a_3 \\
 0 &= r_4 x_4 + \sum_{j=1}^4 C_{4,j} x_j x_4 + a_4
 \end{aligned}$$

Since we have  $r_1, r_2 = 1$  and  $a = [1.6, 1.6, 1.6, 1.6]$  in all studied situations, we fill in those values and write out the summation:

$$\begin{aligned}
 0 &= x_1 + C_{1,1} x_1 x_1 + C_{1,2} x_1 x_2 + C_{3,1} x_1 x_3 + C_{4,1} x_1 x_4 + 1.6 \\
 0 &= x_2 + C_{2,1} x_2 x_1 + C_{2,2} x_2 x_2 + C_{2,3} x_2 x_3 + C_{2,4} x_2 x_4 + 1.6 \\
 0 &= r_3 x_3 + C_{3,1} x_3 x_1 + C_{3,2} x_3 x_2 + C_{3,3} x_3 x_3 + C_{3,4} x_3 x_4 + 1.6 \\
 0 &= r_4 x_4 + C_{4,1} x_4 x_1 + C_{4,2} x_4 x_2 + C_{4,3} x_4 x_3 + C_{4,4} x_4 x_4 + 1.6
 \end{aligned}$$

We can exploit the symmetries in  $r$  and  $C$  to simplify finding the fixed points. The derivatives of  $x_1$  and  $x_2$  are actually identical, and the derivatives of  $x_3$  and  $x_4$  are identical. Thus also their integrals are identical. Thus, we can substitute  $x_1$  into  $x_2$ , and  $x_3$  into  $x_4$  to arrive at a simpler 2-dimensional system. Making the

substitutions, and filling in the parameter values, the differential equations then reduce to

$$0 = 1x_1 - 0.2x_1^2 + 0.04x_1^2 - 0.4x_1x_2 + 1.6$$

$$0 = r_3x_3 - 0.2x_3^2 - 0.4x_3x_1 + 0.04x_3^2 + 1.6$$

where  $r_3$  is the stress level for which the fixed points should be computed.

We now solve these systems for a number of stress values ( $r_3$ ) using Mathematica (Wolfram Research, Inc., n.d.). This way, we computed the following fixed points shown in Table F.1, which are displayed in panel (a) of Figure 9.1 in Section 9.2.1:

Stress	Healthy:PE	Healthy:NE	Unhealthy:PE	Unhealthy:NE	Unstable:PE	Unstable:NE
0.90	5.28	1.15				
0.91	5.26	1.16				
0.91	5.24	1.17				
0.92	5.22	1.18				
0.93	5.19	1.19				
0.93	5.17	1.20				
0.94	5.15	1.22				
0.95	5.12	1.23				
0.95	5.10	1.24				
0.96	5.08	1.26				
0.90	5.28	1.15				
0.91	5.26	1.16				
0.91	5.24	1.17				
0.92	5.22	1.18				
0.93	5.19	1.19				
0.93	5.17	1.20				
0.94	5.15	1.22				
0.95	5.12	1.23				
0.95	5.10	1.24				
0.96	5.07	1.26	1.83	3.96	2.03	3.66
0.97	5.05	1.27	1.66	4.25	2.25	3.30
0.97	5.02	1.29	1.57	4.42	2.39	3.22
0.98	4.99	1.31	1.50	4.56	2.50	3.10
0.99	4.96	1.33	1.45	4.69	2.61	2.99
0.99	4.92	1.34	1.40	4.79	2.71	2.89
1.00	4.89	1.36	1.36	4.89	2.80	2.80
1.01	4.85	1.39	1.33	4.98	2.90	2.72
1.01	4.80	1.41	1.30	5.06	2.99	2.64
1.02	4.76	1.44	1.27	5.15	3.09	2.56
1.03	4.71	1.47	1.24	5.23	3.19	2.48
1.03	4.65	1.50	1.22	5.30	3.29	2.40
1.04	4.59	1.54	1.19	5.38	3.40	2.32
1.05	4.51	1.58	1.17	5.45	3.52	2.23
1.05	4.41	1.65	1.15	5.52	3.67	2.12
1.06	4.24	1.75	1.13	5.59	3.87	1.98
1.06			1.12	5.63		
1.07			1.11	5.66		
1.07			1.09	5.72		
1.08			1.08	5.79		
1.09			1.06	5.85		
1.09			1.04	5.91		
1.10			1.03	5.98		

**Table F.1:** Fixed points of the emotion model for different values of stress (rows), rounded to two decimals. The 2nd and 3rd columns refer to the fixed points of the healthy fixed points for positive and negative emotions; the 4th and 5th columns refer to the unhealthy fixed points; and the last two columns refer to the unstable fixed point.

## F.2 Mean-Switching Hidden Markov Model

In this appendix we provide additional details with respect to the specification of the mean-switching Hidden Markov Model, and using model selection to obtain the number of components, described in Section 9.3.2

### F.2.1 Model Specification

The mean-switching Hidden Markov Model is denoted

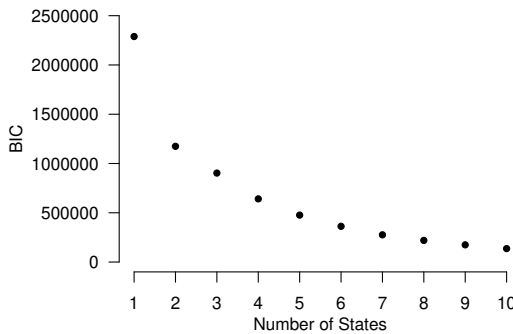
$$P(X, S | \mu, \sigma) = \pi_i \mathcal{N}(X_1) \prod_{t=1}^{T-1} A_{ji} \mathcal{N}(X_{t+1}),$$

where  $X = \{X_1, \dots, X_T\}$  is a matrix of  $p$ -variate elements  $X_j$ ,  $S \in \{1, \dots, K\}^T$  is a vector of length  $T$  indicating the state at each time point,  $\pi_i$  is the probability of being in state  $i \in \{1, \dots, K\}$ ,  $A_{i,j}$  is the probability of transitioning from state  $i$  to state  $j$ , and  $\mu, \sigma$  parameterize the multivariate Gaussian distribution  $\mathcal{N}$  with zero covariances.

In Section 9.3.2 we chose  $K = 2$  components, and fix the covariances of the Gaussian distribution to zero. Since we model four variables, this gives us  $2 \times 4$  means and  $2 \times 4$  standard deviations. The transition matrix  $A$  has three parameters since the last one is determined by the remaining three. Similarly, the marginal probabilities  $\pi_1, \pi_2$  are determined by  $A$  and therefore do not count as additional parameters. We therefore fit a model with 19 freely estimated parameters.

### F.2.2 Model Selection for Mean-Switching HMM

In Section 9.3.2 we inserted bistability as an assumption in the model by specifying that the HMM exhibits two states, and therefore the HMM does not provide us any evidence with respect to which number of states represents the data best. This can be done by performing model selection between HMMs with different numbers of states. A popular way to select between mean-switching HMMs / Gaussian mixtures is the Bayesian Information Criterion (BIC) (Schwarz et al., 1978), because it has been shown to be consistent in estimating Gaussian mixtures (Leroux, 1992), and has outperformed other information criteria (including the AIC) in simulations (Steele & Raftery, 2010). Here we fit HMMs with  $K \in \{1, \dots, 10\}$  and report the BIC values in Figure F.1:



**Figure F.1:** The figure depicts BIC values as a function of the number of states  $K$ , for HMMs fitted to the ideal data.



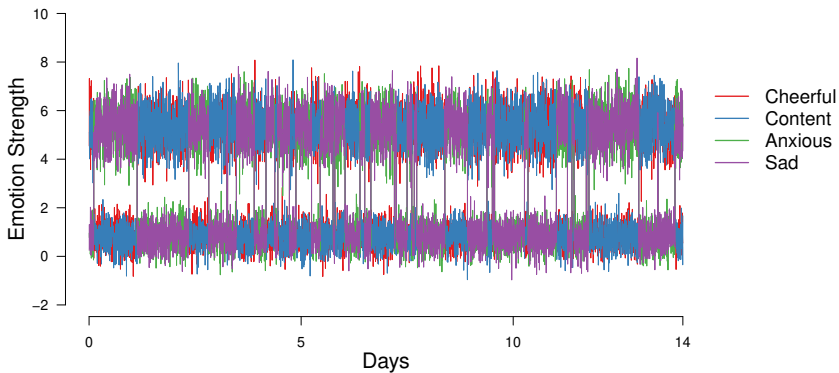
We see that the BIC is highest for  $K = 1$  and then decreases for larger  $K$ , however the the change in BIC becomes less and less when adding additional states. Since we know from the true bistable system that the number of states is  $K = 2$ , we see that the BIC does not select the true number of states. The reason is that the BIC has been shown to be a consistent estimator of  $K$  if the data is generated from a Gaussian mixture. However, in the present case the data is generated from a bistable dynamical system. This failed attempt at model selection based on statistical models again highlights the problems of using misspecified statistical models to make inferences about dynamical systems models.

### F.3 Data Generated from Estimated Models

In this Appendix we show data generated from estimated models for the time period of two weeks of the original time series.

#### F.3.1 Mean Switching Hidden Markov Model

Figure F.2 displays a time series of two weeks generated from the Mean switching HMM estimated in Section 9.3.2:

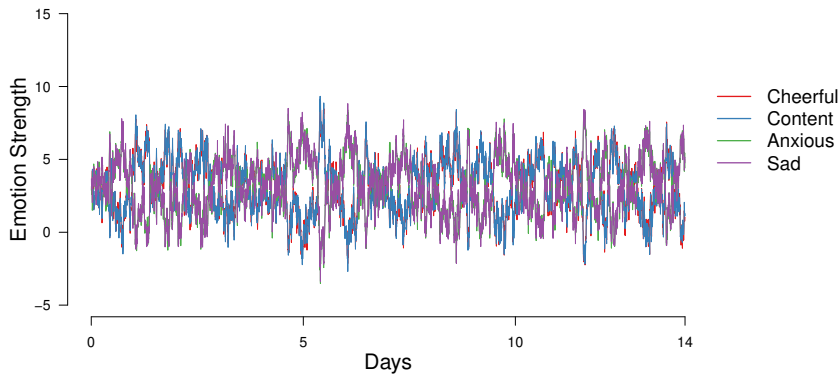


**Figure F.2:** A time series of two weeks generated from the HMM estimated in Section Section 9.3.2.

The generated time series looks similar to the original data in that it switches between the two fixed points at around (1,6) and (6,1). However, there are also differences. In the original data there are less switches that lead to a long-lived change in fixed point, but more switches that are very short-lived. Second, due to the form of the Mean-Switching HMM, there are no “intermediate” observations leading from one fixed point to the other. These observations exist in the original time series (see panel (b) in Figure 9.2).

### F.3.2 First-order Vector Autoregressive (VAR(1)) model

Figure F.3 displays a time series of two weeks generated from the VAR(1) model in Section 9.3.4:

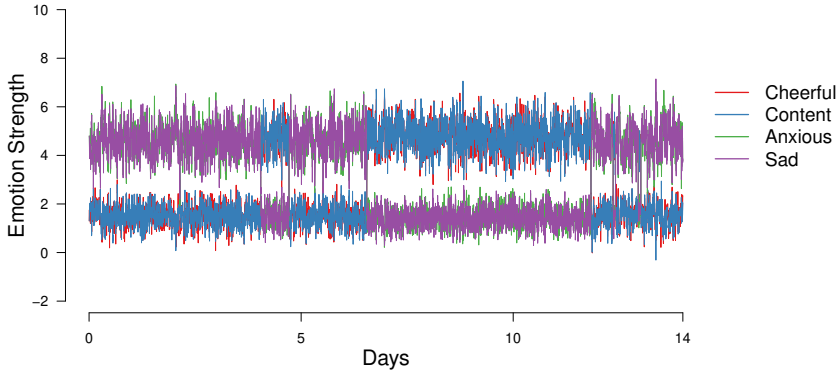


**Figure F.3:** A time series of two weeks generated from the VAR(1) model estimated in Section Section 9.3.4.

The generated data does not show bistability, which is expected because the VAR(1) model exhibits only a single fixed point. What looks approximately like oscillating behaviour is a result of the high auto-regressive effects present in the estimated VAR(1) model: given a stochastic input, the high auto-regressive effects ensure that the system is slow to eventually return to equilibrium. This oscillating behaviour is also evident in the eigenvalues of  $\Phi$ , which consist of one complex conjugate pair (Strogatz, 2015).

### F.3.3 Threshold VAR(1) Model

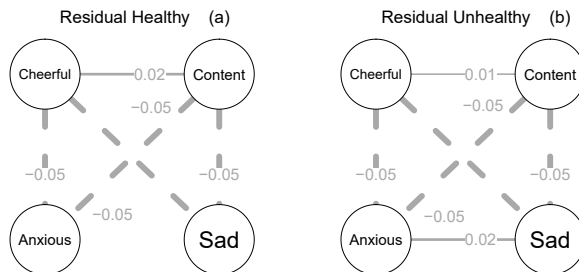
Figure F.4 displays a time series of two weeks generated from the TVAR(1) model in Section 9.3.5:



**Figure F.4:** A time series of two weeks generated from the TVAR(1) model estimated in Section Section 9.3.5.

The data generated from the TVAR(1) model looks similar to the original time series in that the position of the fixed points and the variance around them is very similar. However, the system seems to switch less often between states, and similarly to the data generated from the HMM above, there are much fewer observations on the transitions between states.

## F.4 Residual Partial Correlations of TVAR(1) Model



**Figure F.5:** Residual partial correlation networks for both regimes in the TVAR model described in Section 9.3.5 in the main text.

## F.5 Differential Equation Model Building Details

In this appendix we present additional information relating to the two-step DE model building procedure utilized in Sections 9.3 and 9.4. This includes details on how model fit is computed, as well as full model fit results and parameter estimates for each of the models described in the main text.

### F.5.1 Evaluating Model Fit

The fit of each model is evaluated with the mean out-of-bag explained variance, referred to throughout as  $R^2$ . This metric is calculated using 10-fold cross-validation. First, the given dataset is randomly partitioned into ten mutually exclusive training and test sets. Second, for each partitioned dataset, regression models A through G, (defined by the expression in the second column of Table F.2) are fit to the training set four times, once each of the four outcome variables  $dx_i/dt, \forall i \in \{1, 2, 3, 4\}$ . Third, the resulting parameters are then used to predict the values of the outcome variable in the test set  $dx_i/dt$ . The variance of the resulting residuals  $VAR(dx_i/dt - \hat{dx}_i/dt)$  is then divided by the variance of the outcome variable in the test set,  $VAR(dx_i/dt)$  yielding an out-of-bag variance explained for variable  $i$  based on model  $m$  in partition  $k$ ,  $R_{i,k,m}^2$ . Averaging the explained variance across each of the partitions yields an average explained variance for variable  $i$  in model  $m$ ,  $R_{i,m}^2$ , and averaging this number across all four outcome variables yields the average out-of-bag explained variance for model  $m$ .

### F.5.2 Ideal Data

In Table F.2 we show the fit of models A through G for the ideal dataset analysis in Section 9.3.6. In Table F.3 we show the full parameter estimates, standard errors and  $p$ -values for the selected model, Model C.

Model	$\frac{dx_{i,t}}{dt} \sim a + r_i x_i + \dots$	$q$	$R^2$
A	$\sum_{j \neq i} r_j x_j$	5	.04464
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	.06874
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	.06870
D	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3$	19	.06871
E	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{j \neq k \neq l}^p \zeta_j(x_j x_k x_l)$	23	.06870
F	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j(x_j x_k x_l)$	35	.06860
G	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j(x_j x_k x_l) + \sum_{(j,k,l,m)}^p \eta_j(x_j x_k x_l x_m)$	70	.06846

**Table F.2:** Model fit results for each of the seven models described in text in Section 9.3.6 for the ideal dataset. The second column gives the model equation for each variable,  $q$  denotes the number of parameters estimated per univariate regression model, and the final column indicates  $R^2$ , the explained variance, as calculated based on the prediction error on a hold-out set, using 10-fold cross-validation.

## F.5. Differential Equation Model Building Details

	$dx_1/dt$			$dx_2/dt$			$dx_3/dt$			$dx_4/dt$		
	Est	SE	$p$	Est	SE	$p$	Est	SE	$p$	Est	SE	$p$
$a$	1.40	0.13	<.01	1.37	0.12	<.01	1.25	0.12	<.01	1.27	0.12	<.01
$x_1$	0.88	0.05	<.01	0.03	0.02	0.19	-0.02	0.02	0.33	0.04	0.02	0.05
$x_2$	0.02	0.02	0.34	0.95	0.05	<.01	0.05	0.02	0.02	-0.01	0.02	0.57
$x_3$	-0.01	0.02	0.72	0.01	0.02	0.68	0.96	0.05	<.01	0.08	0.02	<.01
$x_4$	0.01	0.02	0.51	<.01	0.02	0.80	0.04	0.02	0.10	0.91	0.05	<.01
$x_1 \times x_1$	-0.18	0.01	<.01	-	-	-	-	-	-	-	-	-
$x_1 \times x_2$	0.04	0.01	<.01	0.03	0.01	<.01	-	-	-	-	-	-
$x_1 \times x_3$	-0.17	0.01	<.01	-	-	-	-0.18	0.01	<.01	-	-	-
$x_1 \times x_4$	-0.18	0.01	<.01	-	-	-	-	-	-	-0.19	0.01	<.01
$x_2 \times x_2$	-	-	-	-0.19	0.01	<.01	-	-	-	-	-	-
$x_2 \times x_3$	-	-	-	-0.19	0.01	<.01	-0.19	0.01	<.01	-	-	-
$x_2 \times x_4$	-	-	-	-0.19	0.01	<.01	-	-	-	-0.18	0.01	<.01
$x_3 \times x_3$	-	-	-	-	-	-	-0.19	0.01	<.01	-	-	-
$x_3 \times x_4$	-	-	-	-	-	-	0.03	0.01	<.01	0.02	0.01	<.01
$x_4 \times x_4$	-	-	-	-	-	-	-	-	-	-0.18	0.01	<.01

**Table F.3:** Full parameter estimates, standard errors and  $p$ -values for Model B in Section 9.3.6, for the DE model fit to ideal data.

### F.5.3 ESM Data

In Table F.4 we show the fit of models A through G for the emulated ESM dataset analysis, from Section 9.4.4. In Table F.5 we show the full parameter estimates, standard errors and  $p$ -values for the selected model, Model G.

Model	$\frac{dx_{i,t}}{dt} \sim a + r_i x_i + \dots$	$q$	$R^2$
A	$\sum_{j \neq i} r_j x_j$	5	0.13991
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	0.16827
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	0.16928
D	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3$	19	0.19455
E	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{j \neq k \neq l}^p \zeta_j (x_j x_k x_l)$	23	0.19801
F	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l)$	35	0.19940
G	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l) + \sum_{(j,k,l,m)}^p \eta_j (x_j x_k x_l x_m)$	70	0.20420

**Table F.4:** Model fit results for each of the seven models described in text, for the ESM time series, described in Section 9.4. The second column gives the model equation for each variable,  $q$  denotes the number of parameters estimated per univariate regression model. The final two columns indicate  $R^2$ , the explained variance, as calculated based on the prediction error on a hold-out set, using 10-fold cross-validation, for the snapshot ESM data and the mean-aggregated ESM data, respectively.  $R^2$  for Model G in the mean-aggregated ESM data case was not available due to multicollinearity problems encountered when fitting the model.

F. Recovering Bistable Systems from Time Series Data

	$dx_1/dt$			$dx_2/dt$			$dx_3/dt$			$dx_4/dt$		
	Est	SE	p	Est	SE	p	Est	SE	p	Est	SE	p
(Intercept)	-0.07	0.73	0.92	-0.16	0.73	0.82	0.31	0.74	0.68	0.20	0.74	0.78
x1	0.19	0.31	0.53	0.16	0.31	0.61	-0.09	0.31	0.78	-0.12	0.31	0.71
x2	0.07	0.30	0.80	0.18	0.30	0.54	-0.29	0.30	0.34	-0.21	0.30	0.48
x3	-0.06	0.31	0.86	0.01	0.31	0.96	-0.05	0.31	0.88	-0.07	0.31	0.82
x4	-0.01	0.31	0.97	-0.04	0.31	0.89	0.01	0.31	0.98	0.10	0.31	0.74
x1 × x1	-0.02	0.06	0.78	-0.01	0.06	0.87	-0.03	0.06	0.65	-0.01	0.06	0.81
x1 × x2	-0.11	0.09	0.21	-0.10	0.09	0.22	0.14	0.09	0.10	0.13	0.09	0.13
x1 × x3	-0.02	0.10	0.86	-0.04	0.10	0.66	<.01	0.10	0.99	0.02	0.10	0.80
x1 × x4	-0.05	0.10	0.59	<.01	0.10	0.99	0.01	0.10	0.94	<.01	0.10	0.98
x2 × x2	0.01	0.06	0.83	-0.02	0.06	0.76	0.02	0.06	0.72	0.01	0.06	0.82
x2 × x3	<.01	0.10	0.98	-0.01	0.10	0.94	0.05	0.10	0.58	0.06	0.10	0.57
x2 × x4	-0.01	0.10	0.95	-0.05	0.10	0.60	0.06	0.10	0.57	0.02	0.10	0.86
x3 × x3	0.02	0.06	0.68	0.01	0.06	0.87	-0.01	0.06	0.87	0.02	0.06	0.73
x3 × x4	0.01	0.09	0.92	0.01	0.09	0.92	0.01	0.09	0.94	-0.04	0.09	0.62
x4 × x4	0.02	0.06	0.80	0.03	0.06	0.68	-0.03	0.06	0.69	-0.03	0.06	0.68
x1 × x1 × x1	<.01	0.01	0.71	<.01	0.01	0.92	0.01	0.01	0.38	0.01	0.01	0.33
x1 × x1 × x2	0.02	0.01	0.19	0.01	0.01	0.48	-0.01	0.01	0.33	-0.02	0.01	0.15
x1 × x1 × x3	<.01	0.01	0.95	0.01	0.01	0.61	0.01	0.01	0.64	<.01	0.01	0.82
x1 × x1 × x4	<.01	0.01	0.78	<.01	0.01	0.71	<.01	0.01	0.73	<.01	0.01	0.81
x1 × x2 × x2	<.01	0.01	0.87	0.01	0.01	0.49	-0.01	0.01	0.39	<.01	0.01	0.76
x1 × x2 × x3	0.01	0.02	0.58	0.01	0.02	0.69	-0.02	0.02	0.32	-0.02	0.02	0.27
x1 × x2 × x4	0.02	0.02	0.25	0.02	0.02	0.22	-0.03	0.02	0.21	-0.02	0.02	0.34
x1 × x3 × x3	<.01	0.01	0.79	<.01	0.01	0.96	<.01	0.01	0.81	<.01	0.01	0.98
x1 × x3 × x4	0.01	0.02	0.73	<.01	0.02	0.80	<.01	0.02	0.99	<.01	0.02	0.99
x1 × x4 × x4	<.01	0.01	0.98	-0.01	0.01	0.62	<.01	0.01	0.80	<.01	0.01	0.75
x2 × x2 × x2	<.01	0.01	0.92	<.01	0.01	0.96	<.01	0.01	0.98	<.01	0.01	0.82
x2 × x2 × x3	<.01	0.01	1.00	<.01	0.01	0.87	<.01	0.01	0.79	<.01	0.01	0.79
x2 × x2 × x4	-0.01	0.01	0.70	<.01	0.01	0.90	<.01	0.01	0.89	<.01	0.01	0.96
x2 × x3 × x3	<.01	0.01	0.92	<.01	0.01	0.93	<.01	0.01	0.89	-0.01	0.01	0.50
x2 × x3 × x4	-0.01	0.02	0.70	<.01	0.02	0.79	<.01	0.02	0.83	0.01	0.02	0.59
x2 × x4 × x4	<.01	0.01	0.91	0.01	0.01	0.63	<.01	0.01	0.83	<.01	0.01	0.78
x3 × x3 × x3	<.01	0.01	0.95	<.01	0.01	0.93	<.01	0.01	0.97	<.01	0.01	0.51
x3 × x3 × x4	-0.01	0.01	0.40	-0.01	0.01	0.68	0.01	0.01	0.63	0.01	0.01	0.39
x3 × x4 × x4	0.01	0.01	0.44	<.01	0.01	0.76	-0.01	0.01	0.56	<.01	0.01	0.73
x4 × x4 × x4	-0.01	0.01	0.36	-0.01	0.01	0.42	0.01	0.01	0.33	0.01	0.01	0.40

**Table F.5:** First part of full parameter estimates, standard errors and p-values for Model G in Section 9.4.4, for the DE model fit to the emulated ESM data. See Table F.6 below for the remaining estimates.

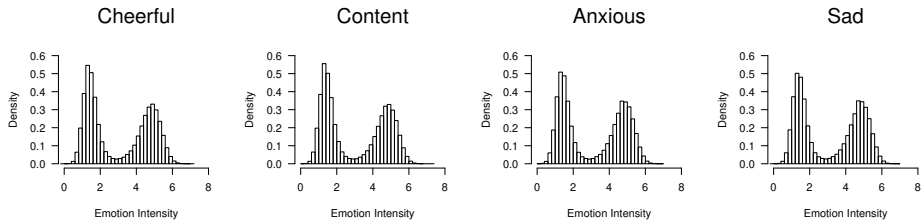
## F.6. Additional Model Results from ESM Time Series

	$dx_1/dt$			$dx_2/dt$			$dx_3/dt$			$dx_4/dt$		
	Est	SE	<i>p</i>	Est	SE	<i>p</i>	Est	SE	<i>p</i>	Est	SE	<i>p</i>
$x_1 \times x_1 \times x_1 \times x_1$	<.01	<.01	0.94	<.01	<.01	0.71	<.01	<.01	0.91	<.01	<.01	0.48
$x_1 \times x_1 \times x_1 \times x_2$	<.01	<.01	0.82	<.01	<.01	0.55	<.01	<.01	0.52	<.01	<.01	0.92
$x_1 \times x_1 \times x_1 \times x_3$	<.01	<.01	0.99	<.01	<.01	0.74	<.01	<.01	0.46	<.01	<.01	0.40
$x_1 \times x_1 \times x_1 \times x_4$	<.01	<.01	0.80	<.01	<.01	0.78	<.01	<.01	0.83	<.01	<.01	0.88
$x_1 \times x_1 \times x_2 \times x_2$	<.01	<.01	0.25	<.01	<.01	0.23	<.01	<.01	0.15	<.01	<.01	0.36
$x_1 \times x_1 \times x_2 \times x_3$	<.01	<.01	0.70	<.01	<.01	0.66	<.01	<.01	0.43	<.01	<.01	0.22
$x_1 \times x_1 \times x_2 \times x_4$	<.01	<.01	0.33	<.01	<.01	0.76	<.01	<.01	0.84	<.01	<.01	0.81
$x_1 \times x_1 \times x_3 \times x_3$	<.01	<.01	0.37	<.01	<.01	0.44	<.01	<.01	0.19	<.01	<.01	0.17
$x_1 \times x_1 \times x_3 \times x_4$	<.01	<.01	0.27	<.01	<.01	0.16	<.01	<.01	0.36	<.01	<.01	0.19
$x_1 \times x_1 \times x_4 \times x_4$	<.01	<.01	0.47	<.01	<.01	0.12	<.01	<.01	0.23	<.01	<.01	0.17
$x_1 \times x_2 \times x_2 \times x_2$	<.01	<.01	0.35	<.01	<.01	0.51	<.01	<.01	0.43	<.01	<.01	0.52
$x_1 \times x_2 \times x_2 \times x_3$	<.01	<.01	0.96	<.01	<.01	0.92	<.01	<.01	0.97	<.01	<.01	0.80
$x_1 \times x_2 \times x_2 \times x_4$	<.01	<.01	0.82	<.01	<.01	0.35	<.01	<.01	0.28	<.01	<.01	0.44
$x_1 \times x_2 \times x_3 \times x_3$	<.01	<.01	0.28	<.01	<.01	0.18	<.01	<.01	0.10	<.01	<.01	0.04
$x_1 \times x_2 \times x_3 \times x_4$	<.01	<.01	0.39	<.01	<.01	0.16	<.01	<.01	0.27	<.01	<.01	0.12
$x_1 \times x_2 \times x_4 \times x_4$	<.01	<.01	0.13	<.01	<.01	0.05	<.01	<.01	0.07	<.01	<.01	0.06
$x_1 \times x_3 \times x_3 \times x_3$	<.01	<.01	0.76	<.01	<.01	0.90	<.01	<.01	0.99	<.01	<.01	0.98
$x_1 \times x_3 \times x_3 \times x_4$	<.01	<.01	0.96	<.01	<.01	0.99	<.01	<.01	0.66	<.01	<.01	0.80
$x_1 \times x_3 \times x_4 \times x_4$	<.01	<.01	0.82	<.01	<.01	0.94	<.01	<.01	0.74	<.01	<.01	0.88
$x_1 \times x_4 \times x_4 \times x_4$	<.01	<.01	0.75	<.01	<.01	0.53	<.01	<.01	0.52	<.01	<.01	0.59
$x_2 \times x_2 \times x_2 \times x_2$	<.01	<.01	0.52	<.01	<.01	0.57	<.01	<.01	0.56	<.01	<.01	0.50
$x_2 \times x_2 \times x_2 \times x_3$	<.01	<.01	0.76	<.01	<.01	0.60	<.01	<.01	0.64	<.01	<.01	0.49
$x_2 \times x_2 \times x_2 \times x_4$	<.01	<.01	0.71	<.01	<.01	0.68	<.01	<.01	0.63	<.01	<.01	0.69
$x_2 \times x_2 \times x_3 \times x_3$	<.01	<.01	0.47	<.01	<.01	0.32	<.01	<.01	0.30	<.01	<.01	0.44
$x_2 \times x_2 \times x_3 \times x_4$	<.01	<.01	0.57	<.01	<.01	0.26	<.01	<.01	0.18	<.01	<.01	0.29
$x_2 \times x_2 \times x_4 \times x_4$	<.01	<.01	0.41	<.01	<.01	0.43	<.01	<.01	0.37	<.01	<.01	0.36
$x_2 \times x_3 \times x_3 \times x_3$	<.01	<.01	0.64	<.01	<.01	0.80	<.01	<.01	0.89	<.01	<.01	0.31
$x_2 \times x_3 \times x_3 \times x_4$	<.01	<.01	0.28	<.01	<.01	0.48	<.01	<.01	0.76	<.01	<.01	0.36
$x_2 \times x_3 \times x_4 \times x_4$	<.01	<.01	0.55	<.01	<.01	0.77	<.01	<.01	0.74	<.01	<.01	0.77
$x_2 \times x_4 \times x_4 \times x_4$	<.01	<.01	0.68	<.01	<.01	0.95	<.01	<.01	0.86	<.01	<.01	0.91
$x_3 \times x_3 \times x_3 \times x_3$	<.01	<.01	0.70	<.01	<.01	0.99	<.01	<.01	0.95	<.01	<.01	0.64
$x_3 \times x_3 \times x_3 \times x_4$	<.01	<.01	0.44	<.01	<.01	0.83	<.01	<.01	0.96	<.01	<.01	0.83
$x_3 \times x_3 \times x_4 \times x_4$	<.01	<.01	0.79	<.01	<.01	0.96	<.01	<.01	0.77	<.01	<.01	0.72
$x_3 \times x_4 \times x_4 \times x_4$	<.01	<.01	0.75	<.01	<.01	0.76	<.01	<.01	0.48	<.01	<.01	0.54
$x_4 \times x_4 \times x_4 \times x_4$	<.01	<.01	0.30	<.01	<.01	0.33	<.01	<.01	0.18	<.01	<.01	0.23

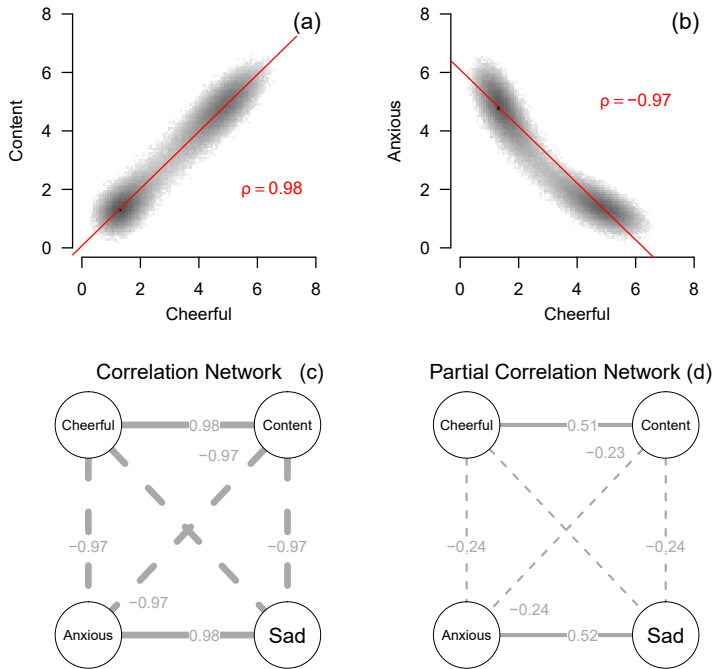
**Table F.6:** Remaining full parameter estimates, standard errors and p-values for Model G in Section 9.4.4, for the DE model fit to the emulated ESM data. For the other estimates see Table F.6 above.

## F.6 Additional Model Results from ESM Time Series

In this appendix, we provide additional figures to visualize the results of the statistical models fit to the ESM time series in Section 9.4.

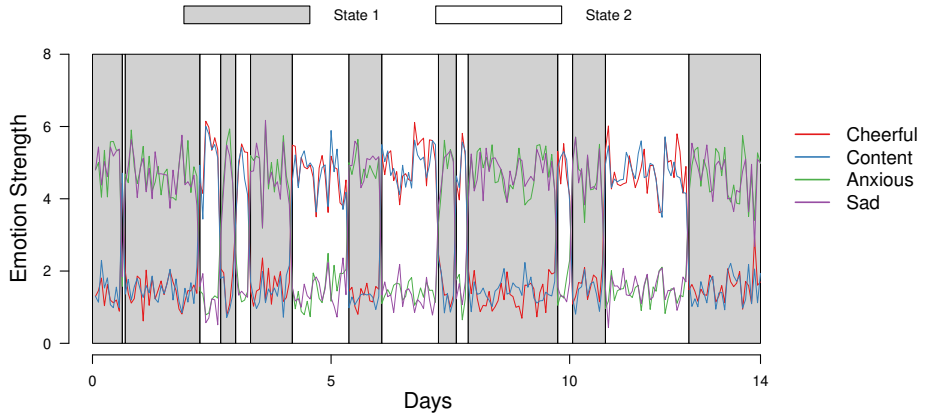


**Figure F.6:** The histograms of the emotion intensity of the four modeled emotions Cheerful, Content, Anxious and Sad, for the ESM data



**Figure F.7:** Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, at the same time point. The red line indicates the best fitting regression model, for ESM time series. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence. Panel (c) displays the correlation matrix as a network, and panel (d) displays the partial correlation matrix as a network.





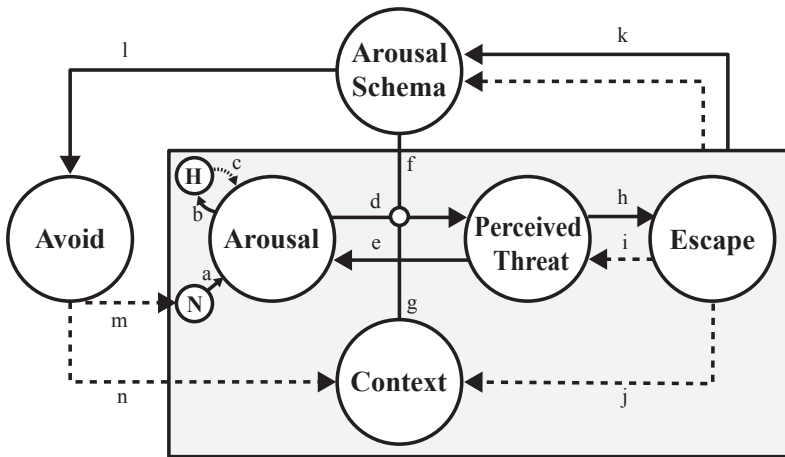
**Figure F.8:** Time series of the four emotion variables, also shown in panel (a) of Figure 9.2, with background color indicating whether a given time point is assigned to the first or second component of the mean-switching HMM estimated from the ESM dataset.



# A FORMAL THEORY OF PANIC DISORDER

## G.1 Overview of Mathematical and Computational Models of Panic Disorder

In this section we provide the exact specification of the model presented in the main chapter. Figure G.1 provides an overview of the specified relationships between variables; Table G.1 provides an overview of the most important difference equations; Table G.2 provides an overview of the variables and their notation in difference equations and the R-code; and Table G.3 provides an overview of the parameters used in the model.



**Figure G.1:** An annotated conceptual model of Panic Disorder. The model includes the six state variables focused on in the main text as well as two additional variables that play an important role in the model behavior: homeostatic feedback (H) and a noise function (N) that induces stochastic variation in arousal. The lowercase English letters provide labels for the relations among the state variables. Below, we use these annotations to further detail how we implemented this conceptual model in the statistical software environment R.

The computational model developed here implements the mathematical model of Panic Disorder using difference equations: iterative functions that use

the value of the state variables at time  $t$  to calculate the value of those same variables at time  $t + 1$ . Below, we present and describe the equations used to calculate each of our state variables. In addition, we present the code used to implement the equations in R. Throughout this overview, we will refer to state variables (the variables the causal diagram in Figure G.1) and the parameters that shape the behavior of either the state variables themselves (e.g., the rate at which they can change) or the relationships between them (e.g., the slope of a linear relationship).

In our model equations, state variables will be represented by uppercase English letters (e.g., A for arousal) and parameters will be represented by lowercase Greek letters (e.g.,  $\kappa$  for the slope of the effect of homeostasis on arousal).

In our R code, uppercase English letters will again be used to represent state variables (e.g., A for arousal). However, here, lowercase English letters will be used to refer to the type of parameter variable (e.g., s for the slope of a linear relationship) and specific parameter variables will be further identified by the state variable(s) related to that parameter (e.g., s\_H\_A for the slope of the linear effect of homeostasis on arousal).

Variable	Difference Equation
Arousal	$A_{t+1} = A_t + \alpha((\nu T_t - A_t) - \kappa_t H_t + N_t)$
Perceived Threat	$T_{t+1} = T_t + \gamma \left( \left( \frac{A_t^\mu}{A_t^\mu + \lambda(S,C)_t^\mu} - T_t \right) - \tau E_t \right)$
	$\lambda(S,C)_t = 1 - \frac{S_t}{S_t + \xi} - \pi C_t$
Escape Behavior	$E_{t+1} = E_t + \varepsilon \left( \frac{T_t^\sigma}{T_t^\sigma + \rho^\sigma} - E_t \right)$
Arousal Schema	$\frac{dS}{dt} = \begin{cases} S_t - \zeta_E S_t, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} < \psi \\ S_t + \zeta_A (\max\{T_{t-\Omega}, \dots, T_t\} - S_t), & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} > \omega \\ S_t - \zeta_E S_t, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} \leq \omega \end{cases}$
Avoidance	$V_{t+1} = V_t + \eta \left( \frac{S_t^\chi}{S_t^\chi + \phi^\chi} - V_t \right)$

**Table G.1:** Key Difference Equations for Computational Model. Our computational model uses difference equations rather than differential equations. This table introduces the difference equations corresponding to the differential equations introduced in the main text. Uppercase English letters represent state variables (e.g., A for arousal). Lowercase Greek letters represent parameters (e.g.,  $\alpha$  for the intrinsic rate of change of arousal).

## G.1. Overview of Mathematical and Computational Models of Panic Disorder

Variable	Time Scale	Notation	Difference Equation	Notation R Code
Arousal	Fast	A		A
Perceived Threat	Fast	T		PT
Escape Behavior	Fast	E		E
Homoestatic Feedback	Fast	H		H
Noise	Fast	N		N
Situational Context	Fast	C		sit
Arousal Schema	Slow	S		AS
Avoidance	Slow	V		AV

**Table G.2:** State Variables in Difference Equations and R-Implementation

Parameter Type	Notation	Difference Equation	Notation R-Code	Reference to Fig G.1	Default Value
Rate	$\alpha$		r_A	Arousal	0.50
Rate	$\beta$		r_H	Homoestasis	0.01
Rate	$\gamma$		r_PT	Perceived Threat	1.00
Rate	$\varepsilon$		r_E	Escape	0.50
Rate	$\zeta_A$		r_AS_a	Arousal Schema	0.25
Rate	$\zeta_B$		r_AS_e	Arousal Schema	0.10
Rate	$\eta$		r_AV	Avoidance	1.00
Relationship	$\theta$		h_A_H	b	0.75
Relationship	$\iota$		p_A_H	b	8.00
Relationship	$\kappa$		s_H_A	c	10.00
Relationship	$\lambda$		h_A_PT	d	0.50
Relationship	$\mu$		p_A_PT	d	5.00
Relationship	$\nu$		s_PT_A	e	1.00
Relationship	$\xi$		h_AS_APT	f	0.50
Relationship	$\pi$		s_sit_APT	g	0.05
Relationship	$\rho$		h_PT_E	h	0.50
Relationship	$\sigma$		p_PT_E	h	10.00
Relationship	$\tau$		s_E_PT	i	0.20
Relationship	$\varphi$		h_AS_AV	l	0.75
Relationship	$\chi$		p_AS_AV	l	7.00
Relationship	$\psi$		cr_F_AS	k	0.50
Relationship	$\omega$		cr_E_AS	k	0.25
Time	$\Omega$		daydef		1440

**Table G.3:** Model Parameters for Mathematical and Computational Model Equations. This table depicts the model parameters for key equations that define the computational model. The first column indicates the type of parameter. Rate parameters define the rate at which a variable can change. Relationship parameters define the relationships among state variables. For each parameter we provide the notation used in difference and differential equations (column 2), the notation used to implement the difference equations in R (column 3), the variables or relationships affected by the parameter (column 4), and the default value we used for the parameter in our implementation of the computational model (column 5).

In what follows, we provide the difference equation and its implementation in R for each of the variables in Table G.2.

### G.1.1 Arousal

As described in the main text, arousal (A) is a function of itself and three variables: perceived threat (T), homeostatic feedback (H), and a correlated noise variable (N) that creates variability in arousal. These effects are denoted as a, c, and e in Figure G.1, respectively.

In the main text, we described the differential equations that constitute our mathematical model. These differential equations provide the instantaneous rate of change in continuous time. The computational model uses difference equations which operate in discrete time. The differential and difference equations used to define arousal are shown in equations G.1 and G.2, respectively.

$$\frac{dA}{dt} = \alpha((\nu T - A) - \kappa H + N) \quad (G.1)$$

$$A_{t+1} = A_t + \alpha((\nu T_t - A_t) - \kappa_t H_t + N_t) \quad (G.2)$$

In the computational model we used Euler’s method to allow our difference equations to adequately approximate the differential equations. The basic idea of this approach is that we take only a very small discrete step in the direction dictated by the difference equation, thus allowing us to re-evaluate the direction we should move after this very small step. In doing so, we avoid under- or over-shooting the destination and better approximate the differential equation. In this case, we use a step size of .001, meaning we take 1000 “small steps” for each time step. The smaller the step size, the more precise the model. This same procedure is followed for all equations in the model.

We used the following code to implement the difference equation for arousal in R:

```
A_eq <- s_PT_A * PT[i]
A_eq2 <- - s_H_A * H[i]
A_eq3 <- s_N_A * N[i]

A[i + 1] <- A[i] + r_A * ((A_eq - A[i]) + A_eq2 + A_eq3) * stepsize
A[i + 1][A[i + 1] < 0] <- 0
```

Note that the subscript *i* is used to indicate each discrete time step (or iteration). Also note that arousal is restricted to being greater than or equal to zero.

### G.1.2 Perceived Threat

Perceived Threat (*T*) is a function of itself and two other state variables: arousal (*A*) and escape (*E*), depicted as effects *d* and *i* in Figure G.1, respectively. The differential and difference equations used to define perceived threat are shown in equations G.3 and G.4, respectively.

$$\frac{dT}{dt} = \gamma \left( \left( \frac{A^\mu}{A^\mu + \lambda^\mu} - T \right) - \tau E \right) \quad (G.3)$$

$$T_{t+1} = T_t + \gamma \left( \left( \frac{A_t^\mu}{A_t^\mu + \lambda(S, C)_t^\mu} - T_t \right) - \tau E_t \right) \quad (G.4)$$

We used the following code to implement the difference equation for arousal in R:

```

ifelse(is.na(sit[i] == 1), c <- 0, c <- sit[i])
h_A_PT[i] <- 1 - (AS[i] / (AS[i] + h_AS_APT)) - s_sit_APT * c
PT_eq <- (A[i]^p_A_PT) / ((A[i]^p_A_PT) + (h_A_PT[i]^p_A_PT))
PT_eq2 <- - s_E_PT * E[i]
PT[i + 1] <- PT[i] + r_PT * ((PT_eq-PT[i]) + PT_eq2) * stepsize
PT[i + 1][PT[i + 1] < 0] <- 0

```

Importantly, the value of the parameter  $\lambda$  is dependent on two other variables in the model: the current level of arousal schema ( $S$ ) and the current level of the situational context variable ( $C$ ). To signify this dependence, we will denote the parameter  $\lambda(S, C)$ . The difference equation for  $\lambda(S, C)$  is given by

$$\lambda(S, C)_t = 1 - \frac{S_t}{S_t + \xi} - \pi C_t \quad (\text{G.5})$$

and implemented in R as:

```

h_A_PT[i] <- 1 - (AS[i] / (AS[i] + h_AS_APT)) - s_sit_APT * c

```

### **G.1.3 Context**

Situational Context ( $C$ ) in our model is chosen probabilistically and remains fixed for a specified period of time. The specified period of time is drawn from a Gaussian distribution with specified mean ( $\mu_S$ ) and standard deviation ( $sd_S$ ). We used a mean of 30 and a standard deviation of 5, meaning that panic-predisposing contexts lasted, on average, 30 “minutes.” In our our code presented below, this randomly drawn duration time is labeled `dur_sit`.

When a new period begins (which is indicated by a missing value for the situation variable; `is.na(sit[i])==TRUE`), we assign to this new period a context that either predisposes one to experiencing a panic attack (i.e., assigning a value of 1) or does not predispose one to experiencing a panic attack (i.e., assigning a value of 0). This value remains (1 or 0) until the period is over (i.e., for the duration specified by `dur_sit`). When a period is over, the above steps are repeated.

We implement the above with the following R code:

```

if(is.na(sit[i]))
{
  dur_sit <- round(rnorm(n = 1, mean = mu_S, sd = sd_S))
  sit_steps <- dur_sit / stepsize
  max_step <- min((i + sit_steps-1), nIter)
  sit[i : max_step] <- rep(sample(0:1, size = 1,
                                prob = c(1 - v_prob_S[i], v_prob_S[i])),
                            times = max_step - i + 1)
}

```

Importantly, the probability of being in a panic predisposing situation is determined, in part, by avoidance behavior (AV). As avoidance increases, the probability of being in a panic predisposing situation decreases. We implemented this in R by making `v_prop_S` a function of avoidance behavior and an initial probability value (`prob_S`) which represents the probability of being in a panic predisposing context in the absence of avoidance. We implement the above with the following R code:

```
v_prob_S[i:nIter] <- prob_S * 2 / (1 + exp(6 * AV[i]))
```

We used an initial probability (`prob_S`) value of .10 (i.e., 10% probability of being in a panic predisposing context). With this equation, a mild level of avoidance ( $AV=.20$ ) reduces the probability of being in a panic-predisposing context to approximately 5%. A moderate level of avoidance ( $AV=.50$ ) reduces the probability to 1%. With more extreme avoidance, the probability approaches 0.

If escape behavior is sufficiently engaged at any point (i.e., exceeds a critical threshold `cr_E_sit`), then the situation variable is set to 0 for the length of the most recently selected “duration” variable (`sit_steps`). We implement this with the following R code:

```

if(E[i] > cr_E_sit) {
  max_steps<-min(i + sit_steps-1, nIter)
  sit[i:max_steps] <- 0
}

```

### G.1.4 Escape Behavior

Escape behavior (E) is a function of itself and perceived threat (T), as depicted in effect h in Figure G.1. The differential and difference equations are given in G.6 and G.7, respectively:

$$\frac{dE}{dt} = \varepsilon \left( \frac{T^\sigma}{T^\sigma + \rho^\sigma} \right) \quad (G.6)$$

$$E_{t+1} = E_t + \varepsilon \left( \frac{T_t^\sigma}{T_t^\sigma + \rho^\sigma} - E_t \right) \quad (G.7)$$

The difference equation is implemented with the following R-code:

```

E_eq <- PT[i]^p_PT_E / ((PT[i]^p_PT_E) + (h_PT_E^p_PT_E))
E[i+1] <- E[i] + r_E * (E_eq - E[i]) * stepsize

```



### G.1.5 Homeostatic Feedback

Homeostatic feedback (H) is a function of itself and arousal (A), as depicted in effect “b” in Figure G.1. The equation defining homeostatic feedback takes the same form as the equation defining escape behavior. The differential and difference equations are given by

$$\frac{dH}{dt} = \beta \left( \frac{A^t}{A^t + \theta^t} - H \right) \quad (\text{G.8})$$

and

$$H_{t+1} = H_t + \beta \left( \frac{A_t^t}{A_t^t + \theta^t} - H_t \right). \quad (\text{G.9})$$

The difference equation is implemented with the following R-code:

```
H_eq <- A[i]^p_A_H / ((A[i]^p_A_H) + (h_A_H^p_A_H))
H[i+1] <- H[i] + r_H * (H_eq - H[i]) * stepsize
```

### G.1.6 Noise

To incorporate stochastic variation in arousal arising from either physiological processes or from the environment, we added a noise variable into the model. Following prior models defined using difference equations, we used a difference equation for red noise (Hasselmann, 1976; van Nes & Scheffer, 2004). Conceptually, red noise at time point  $t$  is generated by adding white noise to the red noise at  $t-1$ . The red noise variable is therefore dependent on the previous time step. Noise that is dependent across time is more realistic than white noise, because it is unrealistic to assume that either arousal or the environmental effects on arousal would change erratically on a minute-to-minute basis. The red noise variable used here allows arousal to vary while also remaining similar from one minute to the next.

The difference equation we used to compute red noise at time step  $t$  is given below. In this equation,  $\lambda$  gives the approximate period of the noise in a given time step (we used a default of  $\lambda = 2$ ),  $N_0$  is the approximate mean of the red noise,  $\beta$  is a parameter that expresses the daily deviation, and  $\varepsilon$  (white\_noise in the R code below) drawn from a standard normal distribution.

$$N_t = \left( 1 - \frac{1}{\lambda} (N_{t-1} - N_0) + N_0 + \beta \varepsilon_t \right) \quad (\text{G.10})$$

The difference equation is implemented with the following R-code:

```
white_noise <- rnorm(n= nIter , mean = .1, sd = sd[i + 1])
red_noise[1] <- white_noise[1]
for(j in 2:nIter) red_noise[j] <- (1 - 1/lambda) * (red_noise[j-1])
+ beta * white_noise[j]
```

We calculated  $\beta$  as

$$\beta = \sigma \sqrt{\frac{2}{\lambda} - \frac{1}{\lambda^2}},$$

where  $\lambda$  is the same parameter as used in the calculation of  $N_t$  and  $\sigma$  is defined as a function of Avoidance. That is, just as Avoidance was part of the equation defining the probability of being in a panic-predisposing context, Avoidance is also part of the equation that defines variation in arousal (see effect a in Figure G.1). We implemented the calculation of  $\sigma$  and  $\beta$  in R as follows:

```
sd[(i+1):nIter] <- s_AV_sd * AV[i+1] + initial$sd
beta <- sd[i + 1] * sqrt(2/lambda - 1 / lambda^2)
```

We used a default value of `s_AV_sd = -1/3*initial$sd` and `initial$sd = .1`. As a result, as Avoidance increases from 0 to 1, `sd` decreases from .1 to approximately .066, a 1/3 reduction in the variability in arousal.

### G.1.7 Avoidance

Avoidance ( $V$ ) is a function of arousal schema (effect l in Figure G.1). Like arousal schema, avoidance updates every 1440 time steps and grows as a function of Arousal Schema. The differential and difference equations are given by:

$$\frac{dV}{dt} = \eta \left( \frac{S^x}{S^x + \varphi^x} - V \right) \quad (\text{G.11})$$

and

$$V_{t+1} = V_t + \eta \left( \frac{S_t^x}{S_t^x + \varphi^x} - V_t \right) \quad (\text{G.12})$$

We implemented the difference equation with the following R-code:

```
AV_eq <- AS[i]^p_AS_AV / ((AS[i]^p_AS_AV) + (h_AS_AV^p_AS_AV))
AV[(i+1):nIter] <- AV[i] + r_AV * (AV_eq - AV[i])
```

### G.1.8 Arousal Schema

Arousal Schema ( $S$ ) is affected by panic attacks (represented by effect k in Figure G.1). Importantly, the effect of panic attacks on arousal schema cannot be reduced to a single variable, but rather is determined jointly by arousal ( $A$ ), perceived threat ( $T$ ) and escape behavior ( $E$ ). Note that we define the new variable Fear ( $F$ ) as  $F = \sqrt{AT}$  to make  $S$  dependent on  $A$  and  $T$ .

As noted in the main text, arousal schema changes on a much slower time scale than the panic attack variables, operating on a time scale of days rather than minutes (i.e., 1,440th the rate of panic attack variables). To ease the computational burden of the model, we thus defined arousal schema to update only

every 1440 time steps, rather than every time step as is done for arousal, perceived threat, and escape behavior.

As described in greater detail in the main text, the calculation of arousal schema is based on two conditional statements. First, is there sufficient arousal and perceived threat to provide the opportunity for learning to occur? Second, is escape behavior present?

The differential equation is given by

$$\frac{dS}{dt} = \begin{cases} 0, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} < \psi \\ \zeta_A(\max\{T_{t-\Omega}, \dots, T_t\} - S), & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} > \omega \\ -\zeta_E S, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} \leq \omega \end{cases} \quad (\text{G.13})$$

and the difference equation is given by:

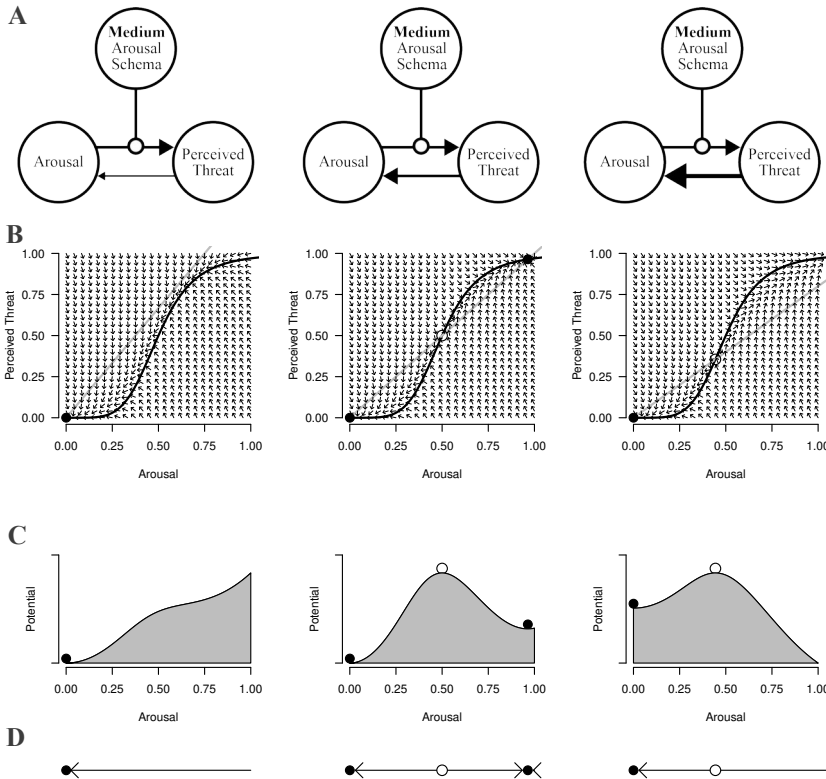
$$\frac{dS}{dt} = \begin{cases} S_t - \zeta_E S_t, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} < \psi \\ S_t + \zeta_A(\max\{T_{t-\Omega}, \dots, T_t\} - S_t), & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} > \omega \\ S_t - \zeta_E S_t, & \text{if } \max\{F_{t-\Omega}, \dots, F_t\} \geq \psi, \max\{E_{t-\Omega}, \dots, E_t\} \leq \omega \end{cases} \quad (\text{G.14})$$

And we implemented the difference equation with the following R-code:

```
if(max(fear[(i-daydef+1):i]) >= cr_F_AS)
{
  ifelse (max(E[(i-daydef+1):i]) > cr_E_AS,
    newAS <- AS[i] + r_AS_a * (max(PT[(i - daydef + 1):i]) - AS[i]),
    newAS <- AS[i] - r_AS_e * AS[i])
  AS[(i+1):nIter] <- newAS
}
```

Note that arousal schema is updated for all remaining iterations (i.e., the next iteration through the final iteration  $[i+1:nIter]$ ). Accordingly, if no further opportunities for learning occur, then arousal schema will remain the same for the remainder of the simulation.

## G.2 Further Examining the Vicious Cycle of Panic Attacks



**Figure G.2:** Like Figure 10.2 in the main chapter, this presents the vector field, stability landscape, and phase line depictions of the  $A \rightleftharpoons T$  feedback loop under varying conditions. In the main chapter, we varied arousal schema (thereby altering the  $A \rightarrow T$  effect) and examined the impact of these changes on the broader feedback loop. Here, arousal schema is kept constant ( $S = 0.50$ ) and the strength of  $T \rightarrow A$  effect is varied from 0.75 (left column) to 1.25 (right column). As seen in Panels B-D, when the strength of the  $T \rightarrow A$  effect is weak, there is no alternative stable state. Runaway positive feedback is not possible. As the strength of  $T \rightarrow A$  increases, an alternative stable state is formed (see middle column). As  $T \rightarrow A$  strengthens, the tipping point into runaway positive feedback shifts to lower levels of arousal. Note, in the right column, the alternative stable state has shifted beyond the bounds of the figure, indicating the system will settle at a higher state of arousal than observed when we varied arousal schema alone.

In main chapter, we analyzed the behavior of the positive feedback loop between arousal and perceived threat ( $A \rightleftharpoons T$ ) with a focus on the impact of arousal schema and, in turn, the strength of the  $A \rightarrow T$  effect, on the behavior of the broader positive feedback loop (see Figure 10.2). In Figure G.2, we further examine the feed-

back loop by keeping arousal schema constant and, instead, varying the strength of the  $T \rightarrow A$  effect (i.e., the extent to which a perception of threat elicits arousal). As seen in Panel B-D, the impact of varying the strength of  $T \rightarrow A$  on the vector field, stability landscape, and phase line diagrams is similar to what was reported in the main chapter when moderating the strength of the  $A \rightarrow T$  effect. When the effect of  $T \rightarrow A$  is weak, there is no alternative stable state and arousal and perceived threat always tend toward zero. However, as the  $T \rightarrow A$  effect strengthens, an alternative stable state is formed and the tipping point into that alternative stable state shifts to lower levels of arousal. Accordingly, individual differences in both the  $A \rightarrow T$  and  $T \rightarrow A$  effects can create vulnerability or resilience to panic attacks.

### G.3 Further Examining Cognitive Behavioral Therapy Intervention on the Model

In the main text, we simulated a brief cognitive behavioral therapy intervention (Otto et al., 2010). Here, we further describe that intervention. Mirroring the procedure we used to develop the model, we first identified individual treatment components and plausible causal effects of the treatment components on components of the model. We identified four treatment components: psychoeducation, cognitive restructuring, interoceptive exposure, and in vivo exposure. Figure G.3 presents the posited mechanisms by which they affect the model.

*Psychoeducation* conducted during Week 1 has its effect directly on the patient's Arousal Schema by presenting a model for understanding arousal-related bodily sensations and panic attacks themselves that explicitly identifies those experiences as benign. *Cognitive restructuring* during Week 2 similarly directly affects Arousal Schema by challenging one's thoughts about the danger of arousal and panic. Cognitive restructuring also equips patient's with a means of regulating their perception of threat when it arises, thereby adding a new negative feedback loop into the model between perceived threat and cognitive restructuring. In our implementation of this negative feedback loop, cognitive restructuring has modest affect on perceived threat relative to the effect of escape behavior.

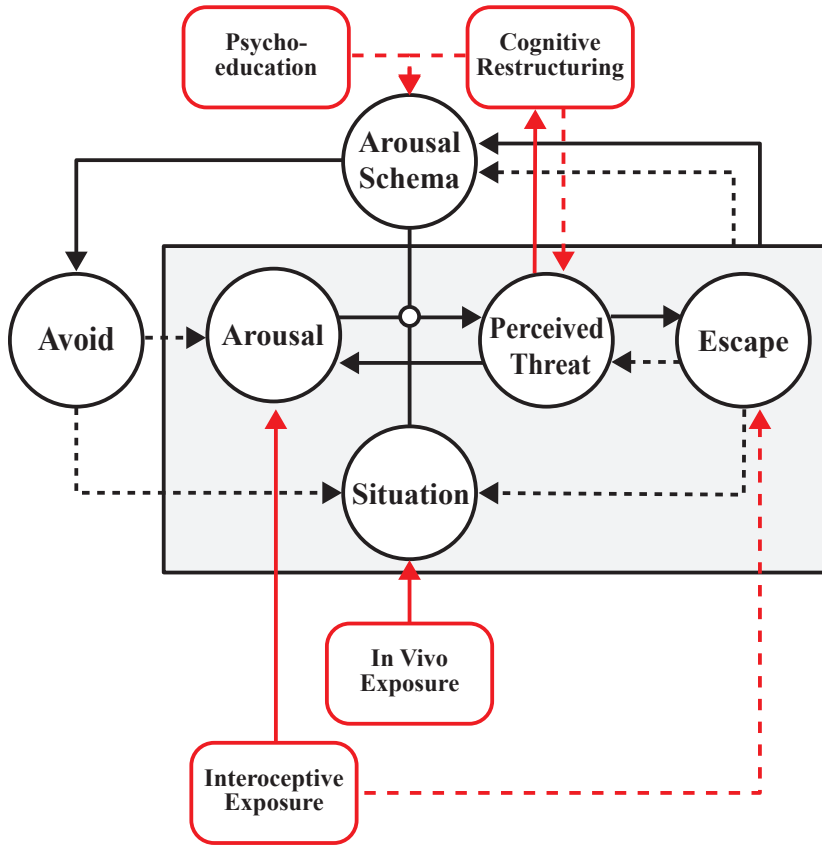


Figure G.3: Plausible relations between treatment components and model components

*Interoceptive Exposure* begins at the conclusion of Week 2 with a procedure (head rolling) intended to produce a modest level of arousal-related bodily sensations (e.g., dizziness). Daily interoceptive exposure continues for the remainder of the treatment, with exercises increasingly targeting those arousal-related sensations that most resemble panic attacks and, thus, elicit the highest level of perceived threat. These procedures are represented in our model as perturbations of arousal with increasing strength as treatment proceeds. *In Vivo Exposure* begins in Week 5. In this implementation, in vivo exposure is paired with ongoing interoceptive exposure, meaning that the perturbation to arousal occurs in a context in which the effect of arousal on perceived threat is strengthened. As a result, the interoceptive exposure is able to elicit greater perceived threat than it would without in vivo exposure, thus potentially allowing for more learning to occur.

## G.4 Theory Evaluation

The panic attack model proposed by Fukano and Gunji (2012) is, to our knowledge, the first effort to formalize a theory of the vicious cycle of panic attacks by implementing it as a mathematical model. Accordingly, we consider their model to be a significant contribution to the Panic Disorder literature. Fukano and Gunji (2012) implemented prior Panic Disorder theory by defining “physical symptoms” and “fear” with coupled logistic equations with Allee effects (see Fukano & Gunji, 2012, p. 461, Equation3). These variables correspond closely to the “arousal” and “perceived threat” variables used in the model proposed in the current chapter. The equations used to define “physical symptoms” (which we will denote as P) and “fear” (which we will denote as F) are:

$$\frac{dF}{dt} = F(F - a_1)1 - \frac{F}{b_1} + P \quad (\text{G.15})$$

$$\frac{dP}{dt} = P(P - a_1)1 - \frac{P}{b_1} + F \quad (\text{G.16})$$

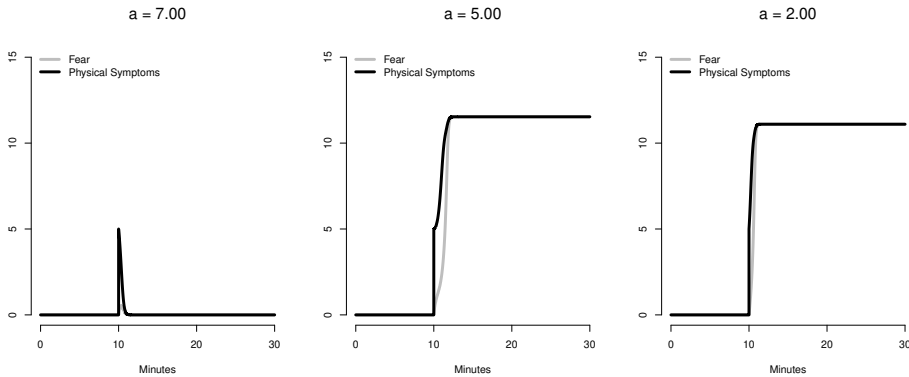
In these equations, the parameters  $a_1$  and  $a_2$  specify critical thresholds above which Fear and Physical Symptoms, respectively, will rise. Logistic equations with Allee effects such as these are often used to model population growth. In such models, the  $a$  parameters represent critical thresholds above which a population will grow. For example, if  $a = 100$ , then when the size of the population exceeds 100, the rate of change in the population will be positive and, thus, the population will grow. Analogously, in this panic attack model, the  $a$  parameters specify thresholds above which Fear and Physical symptoms will increase. These parameters play a critical role in the model as they are used to incorporate individual differences in the vulnerability to panic attacks. Fukano and Gunji (2012) defined three groups, which they refer to as “functional”, “acute phase” [of Panic Disorder], and “chronic phase” [of Panic Disorder], with  $a_1$  and  $a_2$  values of 7, 5, and 2, respectively. That is, in the “functional” parameter setting, the threshold at which Fear and Physical Symptoms will grow is higher than that of acute phase, which, in turn, is higher than that in the chronic phase.

The parameters  $b_1$  and  $b_2$  serve to regulate the severity of Fear and Physical Symptoms, respectively. When used to model population growth, these parameters define the “carrying capacity” of the population, representing constraints on the maximum population size by factors such as the depletion of available food. Analogously, the  $b$  parameters here represent constraints on the growth of Fear and Physical Symptoms. Fukano and Gunji (2012) use the same value for  $b_1$  and  $b_2$  (i.e., 10) across all models.

### G.4.1 What can the model explain?

By explicating their theory as a mathematical model, we were able to implement it as a computational model in R and examine the behavior implied by the theory. As described by Fukano and Gunji (2012) in their article, this model

can explain the most fundamental feature of panic attacks (the sudden rise of fear and arousal-related physical symptoms) as well as individual differences in the propensity to experience those attacks. For example, we ran a “biological challenge” simulation akin to that performed in our main text under the “functional,” “acute,” and “chronic” parameter settings. In this simulation, we perturbed Physical Symptoms by setting it to value of 5 for one time step beginning at time step 10.



**Figure G.4:** A “biological challenge” simulation in which we simulated a perturbation ( $P=5$ ) in three conditions: “functional,” “acute,” and “chronic” ( $a = 7, 5,$  and  $2,$  respectively).

As seen in Figure G.4, in the “functional” (i.e., non-Panic Disorder) parameter setting where the critical threshold value is high ( $a = 7$ ), the perturbation provokes only a transient increase in physical symptoms and a minimal increase in fear. However, for the “acute” and “chronic” conditions, this same perturbation is sufficient to send the system into an alternative stable state of elevated fear and physical symptoms. Note that the stable state in the “acute” phase is slightly higher ( $F = P = 11.5$ ) than that of the “chronic” phase ( $F = P = 11.1$ ), precisely as discussed by Fukano and Gunji (2012, p. 464).

## G.4.2 Limitations to the Model’s Accuracy and Consilience

As this simulation illustrates, the model is able to explain core features of panic attacks. However, the model also has limitations. For example, the model can account for the rapid rise of Fear and Physical Symptoms, but does not account for the subsequent termination of the panic attack. As depicted in Figure G.4, once the model enters an alternative stable state of a panic attack, it remains there indefinitely. Thus there are limitations to the model’s *consilience* (i.e., the amount the model can explain).

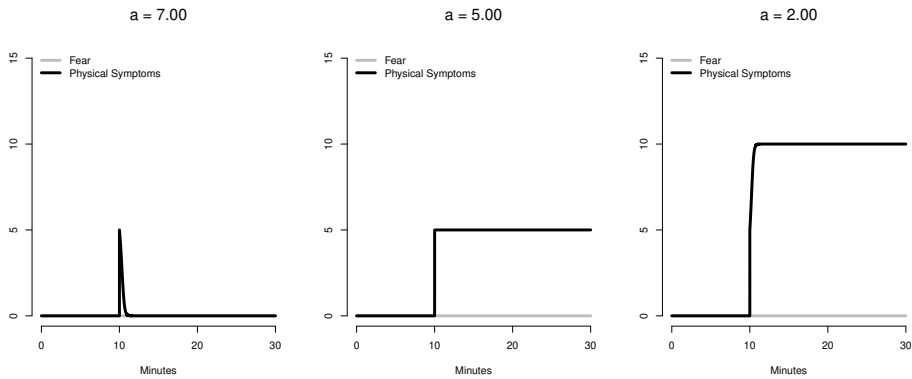
In addition, there are limitations to the model’s *accuracy* (i.e., its correspondence with established facts about panic attacks). For example, if we “successfully treat” the feedback between Physical Symptoms and Fear by eliminating the contribution of Physical Symptoms to Fear and vice versa



$$\frac{dF}{dt} = F(F - a_1)1 - \frac{F}{b_1}, \tag{G.17}$$

$$\frac{dP}{dt} = P(P - a_1)1 - \frac{P}{b_1}, \tag{G.18}$$

then the equations imply that above the critical threshold  $a_2$ , Physical Symptoms will still rise to an elevated stable equilibrium. In other words, even without the vicious cycle, Physical Symptoms will still flip into an alternative stable state of panic. To illustrate this behavior, we repeated the simulation presented in Figure G.4 under the same conditions, but with the feedback effects between Physical Symptoms on Fear removed.



**Figure G.5:** To determine the effect of eliminating the feedback between Fear and Physical symptoms, we removed the effect of Physical Symptoms on Fear and the effect of Physical Symptoms on Fear and repeated the “biological challenge” simulation presented in Figure G.4.

As seen in Figure G.5, in the “chronic” parameter settings, the perturbation to physical symptoms ( $P = 5$ ) remains sufficient to send the system into an alternative stable. Importantly, this occurs solely because of the intrinsic behavior of Physical Symptoms as we have removed the feedback between Physical Symptoms and Fear. This behavior is inconsistent with the simple observation that, absent external effects, Physical Symptoms tend toward a low stable equilibrium. More relevant to the context of Panic Disorder theory, it is inconsistent with the robust finding that intervening on the feedback relationship between Physical Symptoms and Fear is sufficient to prevent the onset of panic attacks in many patients. Accordingly, this simulation illustrates that the model also has limitations to its accuracy.

Note that in the “acute” parameter setting, the perturbation pushes the system into an alternative state that it remains in for the remainder of the simulation. This is neither the stable state of no physical symptoms nor the alternative stable state of elevated physical symptoms. Here, the system is precisely positioned at an “unstable” fixed point. It is “unstable” because any increase or decrease in Physical Symptoms would send the variable toward one of the two stable states.

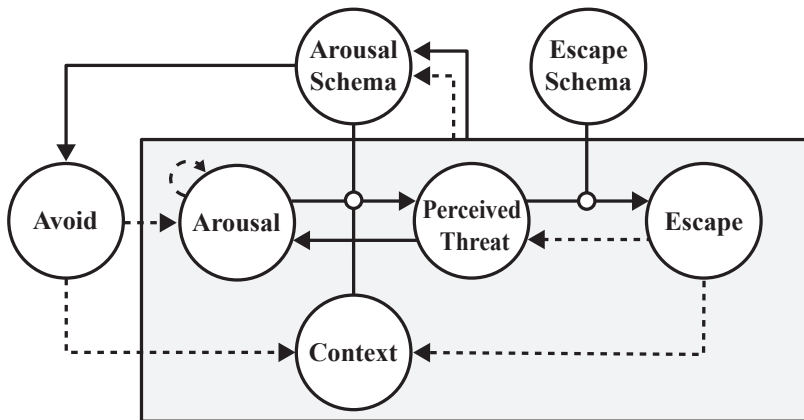
However, at this precise value ( $P = 5$ ), the rate of change is 0 and the system remains fixed. This can be readily seen by plugging in the relevant values ( $P = 5$ ,  $a_2 = 5$ ) and calculating the rate of change in  $P(\frac{dP}{dt} = 0)$

### G.4.3 Conclusion

By explicating their theory as a mathematical model, Fukano and Gunji (2012) produced a model that can be readily evaluated and further developed by other researchers. As we have shown in this brief examination of the system, the model is able to produce the most important features of a panic attack, but has significant limitations to what it can explain and its correspondence with what is known about panic attacks and Panic Disorder. This model thus identifies aspects of Panic Disorder theory in need of further development.

## G.5 Theory Development

The model developed in the main text fails to explain the phenomenon of non-clinical panic attacks, illustrating the need for further theory development. In this section, we propose the incorporation of an Escape Schema ( $S_E$ ) variable that moderates the effect of perceived threat on escape behavior, just as the arousal schema variable moderates the effect of arousal on perceived threat.

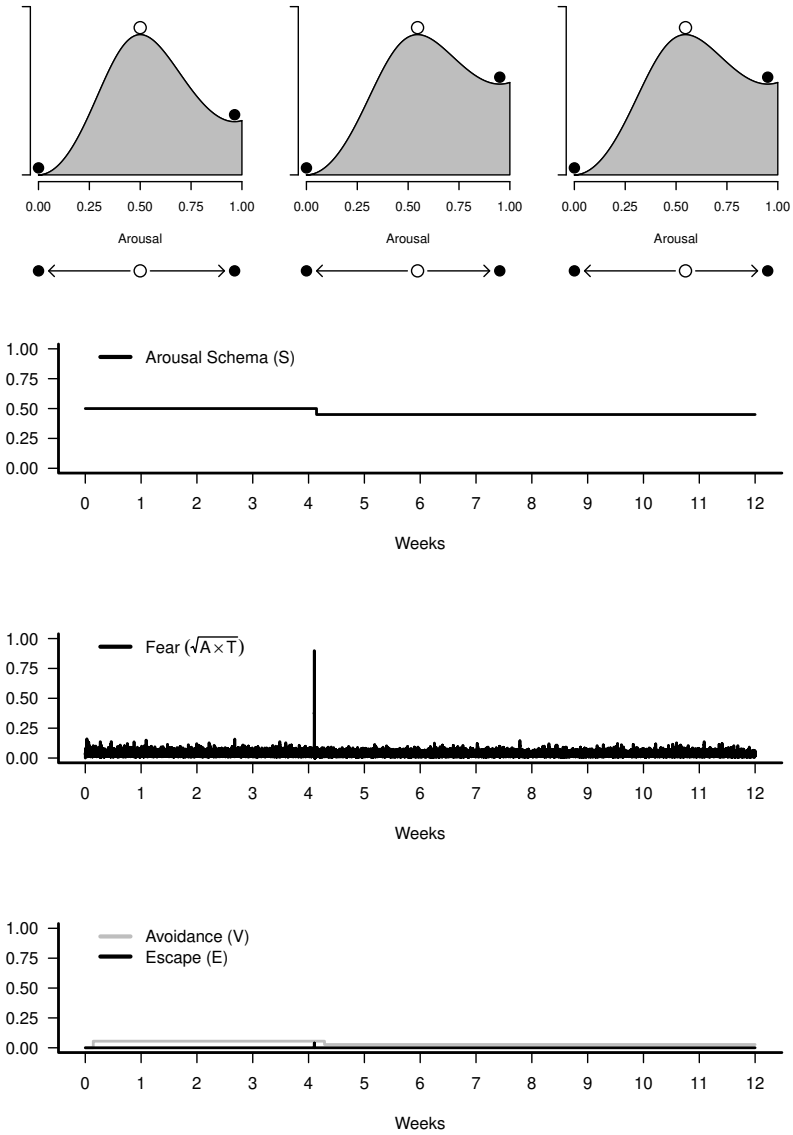


**Figure G.6:** An extended causal diagram of Panic Disorder Theory that incorporates an Escape Schema variable moderating the effect of Perceived Threat on Escape Behavior.

In order to incorporate the Escape Schema variable ( $S_E$ ), we followed the same approach used for Arousal Schema, making the parameter that defines the half saturation point of the  $T \rightarrow E$  effect ( $\rho$ ) dependent on the level of  $S_E$  as follows:

$$\rho(S_E) = 1.25 - \frac{S_E}{S_E + \xi_E} \tag{G.19}$$

As  $S_E$  increases (i.e., as one's beliefs about the value of escape behavior increase),  $\rho$  is decreased and less perceived threat is needed in order to elicit escape behavior. As  $S_E$  decreases (i.e., one doesn't believe escape behavior to be especially effective),  $\rho$  increases and more perceived threat is needed to elicit escape behavior. The precise form of this equation should not be overinterpreted. The equation is important only insofar as it allows us to constrain the range of  $\rho$  such that, at high  $S_E$ , the strength of the T→E effect is stronger than it was in our default implementation of the model as described in the main text whereas, at low  $S_E$ , very high levels of perceived threat are required to elicit escape behavior. We modeled the effect of  $S_E$  in this way to reflect the observation that, even in the context of a panic attack, some individuals do not engage in significant escape behavior (e.g., do not flee the situation).



**Figure G.7:** Twelve “week” simulation of model behavior beginning with moderately elevated arousal schema ( $S = 0.50$ ) and very low escape schema ( $S_E = 0$ ). The bottom row depicts the stability landscape and phase line for arousal at three points during the simulation: Week 0, Week 6, and Week 12.

### G.5.1 What can this modification to the model explain?

To examine whether this revision to the model allowed the model to explain the experience of panic attacks in the absence of Panic Disorder, we simulated 12-weeks of data from the model beginning with a moderate level of arousal schema ( $S = 0.50$ ) such that panic attacks were possible but not yet recurrent (cf. the simulations performed to examine Feature 3 (Section 10.4.3) in the main text). We specified a very low level of the newly incorporated escape schema variable ( $S_E = 0$ ), such that high levels of perceived threat were required to prompt escape behavior.

As seen in Figure G.7, in this simulation the model goes through an extended period of time (4 “weeks”) in which no panic attacks occur. Following the fourth week, a panic attack does occur, setting the stage for a vicious cycle of learning that would send the system into a state of recurrent panic attacks. However, because perceived threat led to minimal escape behavior over the course of the panic attack, the lesson imparted by this experience is that arousal did not prove to be dangerous, even in the absence of significant escape behavior. Thus, the arousal schema variable decreases and the system becomes less vulnerable to panic attacks.

With the incorporation of the escape schema variable, the model is able to explain the phenomenon of non-clinical panic attacks. A panic attack occurred without being followed by recurrent panic attacks, increased persistent cognitions about the danger of arousal, and avoidance behavior. Indeed, as a result of the panic attack in the absence of escape behavior, the system becomes less vulnerable to panic attacks, suggesting that even if another panic attack were to occur, it would only further disconfirm the belief that arousal is dangerous.

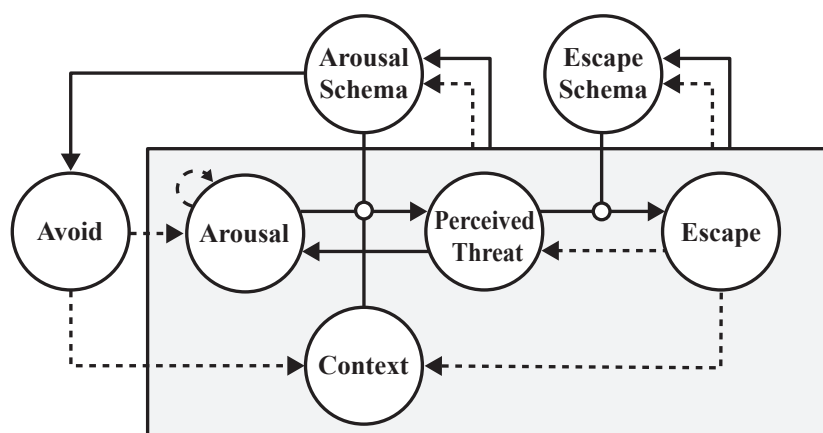
In addition, the model provides an explanation for why some individuals do not engage in escape behavior despite substantially elevated arousal and perceived threat. For these individuals, beliefs about the danger of arousal-related bodily sensations may be sufficient to create a system vulnerable to panic attacks, but escape behavior is not regarded as an especially valuable behavior in response to such attacks and, thus, is not engaged in despite the presence of elevated arousal and perceived threat.

### G.5.2 Future Research and Theory Development

Together, the added explanatory benefit of incorporating an escape schema variable suggests that this component is worth examining further as a possible development of Panic Disorder theory.

In the simulation depicted in Figure G.7, we modeled escape schema as a static variable that can represent individual differences in the propensity to experience panic attacks. However, it is almost certainly the case that experiences with elevated arousal and perceived threat provide opportunities to modify Escape Schema, just as they do Arousal Schema. For example, it seems likely that the relief in perceived threat achieved by engaging in escape behavior reinforces the value of such behavior, thereby increasing escape schema and making future escape behavior more likely. In contrast, refraining from escape behavior

may further reduce one's beliefs about the value of such behavior as it becomes clear that escape behavior is unnecessary to manage arousal and perceived threat (see Figure G.8). In future work examining the possible inclusion of an Escape Schema variable, Panic Disorder theorists should also consider the impact of panic attacks on the Escape Schema variable.



**Figure G.8:** An extended causal diagram of Panic Disorder Theory that incorporates a learning effect by which panic attacks impact not only Arousal Schema, but also Escape Schema, either increasing or decreasing beliefs about the value of escape behavior.

---

# FROM DATA MODELS TO FORMAL THEORIES

---

## H.1 Simulated Data from the Panic Model

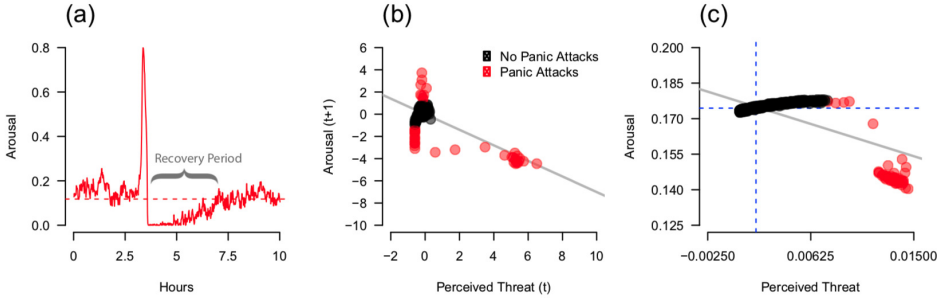
In this appendix we describe in more detail how the simulated dataset, presented in both Sections 11.3.2 and 11.3.3 are obtained.

Data is simulated from the Panic Model, the full specification of which is given by (Robinaugh, Haslbeck, et al., 2019), using the statistical programming language *R*. We use the Panic Model to generate time-series data of 1000 individuals, on a single minute time scale, for 12 weeks, using Euler’s method with a step size of .001. This yields a total of  $n_t = 12,0960$  repeated measurements per person. Each individual starts with a different initial value of arousal schema, drawn from a normal distribution with  $\mu = 0.25$  and  $\sigma = 0.0225$ . The parameters of this distribution were chosen to roughly generate a representative number of panic disorder sufferers (for more details see Robinaugh, Haslbeck, et al., 2019). Otherwise each individual obtains the same parameter values and the same starting values on all processes, with the stochastic noise terms drawn using a different random seed for each individual. The mapping from this raw data to the variables used in the network models of Section 11.3.2 is described in the main text. Code to reproduce this data-generation scheme can be found in the reproducibility archive of this paper.

## H.2 Panic Model and Statistical Dependencies

In this appendix we describe in more detail the patterns of statistical dependencies produced by the three data models fitted to data simulated from the Panic Model in Section 11.3.2. While in the main text we discuss the statistical dependencies between Arousal and Perceived Threat, and Arousal Schema and Avoidance, here we focus only on the former. Key to the Arousal-Perceived Threat dependencies is the positive feedback loop between Arousal and Perceived Threat in the Panic Model (as described in Section 11.2). If Arousal and Perceived Threat become sufficiently elevated, this “vicious cycle” leads to runaway positive feedback, with a pronounced spike in both Arousal and Perceived Threat (i.e., a panic attack). This spike initiates a process of homeostatic feedback that brings Arousal down and suppresses Arousal below its baseline for a period of time after this panic attack, a period which we will refer to as a *recovery period*.

The panic attack itself lasts about 30 minutes. However, the recovery period lasts for 2-3 hours (see Figure H.1 panel A).



**Figure H.1:** Panel (a) depicts Arousal during a panic attack, showing the short sharp peak of arousal levels, followed by a longer recovery period of low arousal, before the system returns to the usual resting state. The dotted line indicates the mean level of arousal over the observation window (0 - 10 hrs). Panel (b) depicts the state-space plot of Perceived Threat and Arousal at the next measurement occasion, as captured by the emulated ESM study and VAR model. Red points indicate an observation window of 90 min in which either part of a panic attack or the following recovery period is captured. The solid grey line reflects the marginal lagged relationship. Panel (c) depicts the cross-sectional marginal relationship between the mean of Arousal and mean of Perceived Threat, as analyzed in the GGM model. Red dots indicate individuals who suffer from panic attacks, and black dots represent “healthy” individuals. The solid grey line shows the negative marginal relationship. The dotted blue lines indicate the median of both variables, by which the binarized values used in the Ising model analysis are defined.

In the VAR model in Figure 11.5 (b) in the main text, we observed a strong negative conditional relationship between Perceived Threat at time  $t$  and Arousal at time  $t + 1$ , conditioning on all other variables at time  $t$ . The distribution of these lagged variables is shown in Figure H.1(b), with the grey line representing the also negative marginal relationship. This strong negative cross-lagged relationship is a direct consequence of the recovery period of Arousal: High values of Perceived Threat are closely followed by a long period of low Arousal values. This can be seen in Figure H.1(b), where observations over windows in which a panic attack and recovery period occur are shaded in red. By averaging arousal values over a window of 90 minutes, the strong positive causal effects operating locally in time (i.e. over a very short time-interval) are not directly captured, but instead the VAR(1) model describes correctly the negative relationship between the *means* of each variable over this window.

In the GGM in Figure 11.5 (c) in the main text, we saw a positive linear relationship between Arousal and Perceived Threat in the estimated GGM. This dependency indicates that high mean levels of Arousal are associated with high mean levels of Perceived Threat, *conditional* on all other variables. We stress the conditional nature of this relationship, because the *marginal* relationship between the two variables is in fact negative as can be seen in Figure H.1 (c). This negative marginal relationship comes about by combining two groups of individuals that have different mean values on both variables. Individuals who experience panic attacks (red points) have high average Perceived Threat, but low average



Arousal, due to the long recovery period of Arousal after a panic attack. On the other hand, individuals who do *not* experience panic attacks have higher average values of Arousal, and lower average values of Perceived Threat. When inspecting the two groups separately, we see that there is a positive linear relationship between mean Arousal and Perceived Threat in the group without panic attacks; The group with panic attacks is too small to determine a relationship. Since Escape and Avoidance behavior only occur after Panic attacks, conditioning on those two variables amounts to conditioning on whether an individual had panic attacks. This conditional relationship is then driven mostly by the positive relationship in the (much larger) group of individuals who have no panic attacks, indicated by the black dots in Figure H.1 (c).

Finally, we can explain the weak positive relationship between Arousal and Perceived Threat in the Ising model (Figure 11.5 (d) in the main text): The levels of these variables are defined by a median split of their mean values, depicted as dotted lines in Figure H.1 (c). Unlike in the GGM, there is a positive marginal relationship between these binarized variables, as the majority of individuals without panic attacks (denoted by the black points) end up in the low Perceived Threat and low Arousal groups (lower left quadrant Figure H.1 (c)) or high Perceived Threat and high Arousal groups (upper right quadrant). How then do we end up with a weakly positive conditional relationship between these two binary variables? Similarly to the GGM above, it turns out that conditioning on variables such as Escape behaviour and Avoidance almost entirely separates individuals into either the low Arousal and low Perceived Threat category (e.g., for low Escape values) or the high Arousal and high Perceived Threat category (for high Escape value). This means that, once we have conditioned on other variables which have direct and indirect causal connections to Arousal and Perceived Threat, there is very little additional information which Arousal can add to predicting Perceived Threat levels (and vice versa). This produces the weak positive conditional relationship between Arousal and Perceived Threat, as well as the stronger positive connections between Avoidance and Perceived Threat.

## H.3 Details Empirical vs Simulated Ising Model

In this appendix we describe in more detail how the theory-implied and empirical Ising Models presented in Section 11.3.3 are obtained.

### H.3.1 Simulated Data and Implied Ising Model

To obtain the theory-implied Ising Model we use the raw time series data generated from the Panic Model and described in Appendix H.1

To create the binary symptom variables in Section 11.3.3 we transformed the raw time-series data of each individual as follows. First, we define Anxiety at a given time point as the geometric mean of the Arousal and Perceived Threat components at that point in time. Second, we define a panic attack as short, sharp peak of Arousal and Perceived Threat. We code a panic attack to be present in

the time series data if Anxiety takes on a value greater than 0.5. The duration of a panic attack is the length of time Anxiety variable stays above this threshold, and so we define a single panic attack as a sequence of consecutive time points in which Anxiety stays over this threshold. This allows to define our first binary symptom variable, Recurrent Panic Attacks:

1. Recurrent Panic Attacks (PA) : PA is present if the individual experience more than three panic attacks over the observation window.

We define recurrent as more than three over the observation window for consistency with how this symptom is defined in the CPES dataset, detailed below.

Next, we can define the symptom Persistent Concern (PC), again using the time series of Anxiety. This symptom is typically described as experiencing a heightened level of anxiety following a panic attack (American Psychiatric Association, 2013). To define this, for each individual who experiences a panic attack, we calculate the mean level of Anxiety in a window of 1000 minutes (16.67 hours) following the end of each panic attack. If another panic attack occurs in that window, we instead take the mean level of Anxiety between the end of one panic attack and before the beginning of the next. This gives us a vector of mean Anxiety levels per person, one for each panic attack experienced. Next, we must define what we consider to be a “heightened” level of anxiety. We do this by obtaining the distribution of mean Anxiety levels for healthy individuals, that is, those members of our sample who never experience a panic attack. We consider mean Anxiety levels following an attack to be “heightened” if they are greater than the 90th percentile of mean Anxiety levels in the healthy population. This gives us our second binary symptom variable.

- 2 Persistent Concern (PC): PC is present if, following at least one panic attack, higher average levels of Anxiety are present than in the healthy population, as defined by the 90th percentile of average Anxiety in the healthy population.

Finally we take a similar approach to defining the symptom Avoidance (Av), typically described as engaging in a heightened level of avoidance behaviour following a panic attack. For this symptom, we use the time series of the Avoid component. For each individual who experiences a panic attack, we calculate the mean level of Avoid in a window of 1000 minutes (16.67 hours) following the end of each panic attack, or before the beginning of the next attack, whichever is shorter. Heightened avoidance behaviour is defined relative to the 90th percentile of Avoid levels in the healthy population. This gives us our third binary symptom variable.

- 3 Avoidance (Av): Av is present if, following at least one panic attack, higher average levels of Avoid are present than in the healthy population, as defined by the 90th percentile of average Avoid in the healthy population.

The Ising model of these three symptom variables is fit using the *EstimateIsing* function from the *IsingSampler* package (Epskamp, 2015), that is, using a non-regularized pseudolikelihood method.

### H.3.2 Empirical Symptom Data

To test the empirical predictions of the Panic Model, we made use of the publicly available Collaborative Psychiatric Epidemiology Surveys (CPES) 2001-2003 (Alegria et al., 2007). The CPES is a nationally representative survey of mental disorders and correlates in the United States. The CPES is attractive to use for testing the Panic Model, first because of the large sample size (20,013 participants) ensuring reliable estimates of empirical dependencies, and second, because approximately 146 items in the survey assess either panic attack or panic disorder experiences, characteristics, and diagnoses, typically in terms of lifetime prevalence.

To define our three panic disorder symptoms, we first use the diagnostic manual of the CPES to define whether individuals have ever experienced a panic attack based on responses to 18 items. There are three criteria which must be met for the individual to be classed as having experienced at least one lifetime panic attack. These are shown in Table H.1. In coding the presence or absence of a panic attack, individuals must positively report at least four out of the thirteen characteristics of a panic attack, according to the second criteria in Table H.1. Missing values were taken as a failure to report that characteristic.

Criterion	Description	Item number(s)
A	A discrete period of intense fear or discomfort	SC20 or SC20a
B (four or more)	Palpitations, pounding heart	PD1a
	Sweating	PD1e
	Trembling or shaking	PD1f
	Sensation of shortness of breath or smothering	PD1b
	Feeling of choking	PD1h
	Chest pain or discomfort	PD1i
	Nausea or abdominal distress	PD1c
	Feeling dizzy, unsteady, lightheaded or faint	PD1d or PD1m
	Derealization or depersonalization	PD1k or PD1l
	Fear of losing control or going crazy	PD1j
	Fear of dying	PD1n
	Paresthesias (numbing or tingling sensations)	PD1p
	Chills or hot flushes	PD1o
C	Symptoms developed abruptly and reached a peak within 10 minutes	PD3

**Table H.1:** Three criteria necessary to code an individual as having one lifetime panic attack based on items from the CPES survey, based on the CPES diagnostic manual

With this definition of a panic attack in place, we define the three binary symptoms of panic disorder, following the definitions laid out in the diagnostic manual for Panic Disorder.

1. Recurrent Panic Attacks (PA). PA is present if participant reports more than three lifetime occurrences of an unexpected, short, sharp attack of fear or

panic (item PD4 and criteria in Table H.1), more than one of which is out of the blue (PD17a)

2. Persistent Concern (PC). PC is present if reported that following an attack, a month or more of at least one of: a) persistent concern about having another attack (PD13a), or b) worry about the implications or consequences of having an attack (PD13b)
3. Avoidance (Av). Av is present if participant reports at least one of a) following an attack, changing everyday activities for a month or more (PD13c), b) following an attack, avoiding situations due to fear of having an attack for a month or more (PD13d), or c) in the past 12 months, avoiding situations that might cause physical sensation (PD42).

In coding this, if two out of three PA criteria were present, and the third was missing, we assigned a positive value to the PA item. The empirical Ising model was fit using the same procedure as the theory-implied Ising model.

---

# LIST OF PUBLICATIONS

---

\* indicates shared first authorship

## I.1 Under Review

Robinaugh, D. J., **Haslbeck, J. M. B.**, Ryan, O., Fried, E. I., & Waldorp, L. (under review). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction.

Dablander, F.\*, Ryan, O.\* & **Haslbeck, J. M. B.\*** (under review). Choosing between AR(1) and VAR(1) Models in Typical Psychological Applications.

**Haslbeck, J. M. B.\***, Ryan O.\*, Robinaugh D. J.\*, Waldorp L. J., Borsboom D. (under review). Modeling Psychopathology: From Data Models to Formal Theories.

**Haslbeck, J. M. B.\*** & Ryan O.\* (under revision). Recovering Bistable Systems from Psychological Time Series.

Robinaugh, D. J., **Haslbeck, J. M. B.**, Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A., McNally, R. J., van Nes, E. H., Scheffer, M., Kendler, K. S. & Borsboom, D. (under review). Advancing the Network Theory of Mental Disorders: A Computational Model of Panic Disorder.

## I.2 Published

**Haslbeck, J. M. B.** & Waldorp, L. J. (2020). mgm: Structure Estimation for Time-Varying Mixed Graphical Models in high-dimensional Data. *The Journal of Statistical Software*.

**Haslbeck, J. M. B.** & Wulff, D. U., (2020). Estimating the Number of Clusters via Normalized Cluster Instability. *Computational Statistics*.

**Haslbeck, J. M. B.**, Bringmann, L. F., & Waldorp, L. J. (2020). A Tutorial on Estimating Time-Varying Vector Autoregressive Models. *Multivariate Behavioral Research*.

**Haslbeck, J. M. B.**, Epskamp, S. & Marsman, M., Waldorp, L. J. (2020). Interpreting the Ising Model: The Input Matters. *Multivariate Behavioral Research*

**Haslbeck, J. M. B.**, Borsboom, D. & Waldorp, L. J. (2019). Moderated Network Models. *Multivariate Behavioral Research*

Kieslich, P. J., Henninger, F., Wulff, D. U., **Haslbeck, J. M. B.**, & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods*. New York, NY: Routledge.

Wulff, D. U., **Haslbeck, J. M. B.**, Kieslich, P. J., Henninger, F., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: Detecting types in movement trajectories. In M. Schulte-Mecklenbeck, A. Kuehberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods*. New York, NY: Routledge.

Fried E. I., von Stockert, S., **Haslbeck, J. M. B.**, Lamers F., Schoevers, R.A. & Pennix B. W. J. H. (2019). Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. *Psychological Medicine*.

Dablander, F., Epskamp, S., & **Haslbeck, J. M. B.** (2019). Studying Statistics Anxiety Requires Sound Statistics: A Comment on Siew, McCartney, and Vitevitch (2019). *Scholarship of Teaching and Learning in Psychology*.

**Haslbeck, J. M. B.**, Waldorp, L. J. (2018). How well do Network Models predict Future Observations? On the Importance of Predictability in Network Models. *Behavior Research Methods*.

**Haslbeck, J. M. B.**, Fried, E. I. (2017). How Predictable are Symptoms in Psychopathological Networks? A Reanalysis of 18 Published Datasets. *Psychological Medicine*.

Kossakowski, J. J., Groot, P. C., **Haslbeck, J. M. B.**, Borsboom, D., & Wichers, M. (2017). Data from 'Critical Slowing Down as a Personalized Early Warning Signal for Depression'. *Journal of Open Psychology Data*, 5:1.

---

# NEDERLANDSE SAMENVATTING

---

## J.1 Data Models

In het eerste deel van mijn dissertatie heb ik verschillende modellen ontwikkeld, die de multivariate afhankelijkheden tussen symptomen en andere variabelen in de psychopathologie omvatten.

Toen ik, in 2015, begon te werken aan mijn dissertatie, waren toegepaste onderzoekers in het veld van netwerkmethoden in de psychopathologie beperkt tot het gebruik van slechts enkele modellen: Het Ising model voor binaire variabelen, de multivariate Gaussiaanse verdeling voor continue variabelen, en het Vector Autoregressive (VAR) model voor continue tijdreeksmodellen. De schattingsmethoden die geïmplementeerd zijn in het R-pakket *mgm*, dat ik ontwikkeld heb, hebben de beschikbare modellen aanzienlijk uitgebreid. In Hoofdstuk 2 heb ik Mixed Graphical Models (MGMs) geïntroduceerd. Deze modellen kunnen de afhankelijkheidsstructuur tussen variabelen omvatten die gedefinieerd zijn op verschillende domeinen, zoals continu en categorisch. Zulke gemixte data komt vaak voor in psychopathologisch onderzoek. Symptomen en psychologische constructen worden vaak gedefinieerd op een continue of ordinale schaal, terwijl variabelen die gerelateerd zijn aan sociale context, werk omgeving, of behandeling vaak gedefinieerd worden op een (nominale) categorische schaal. Tevens is het nu mogelijk om non-lineaire interacties te vinden door het modelleren van continue en ordinale variabelen als zijnde categorische variabelen. Ten slotte heb ik de schattingsmethoden voor MGMs aangepast aan Vector autoregressieve (mVar) modellen, waar verschillende types variabelen elkaar voorspellen over tijd.

Statistische modellen worden doorgaans gerapporteerd met behulp van netwerk visualisaties. In deze figuren worden absolute waarden van parameters vertegenwoordigd door de wijdte van de verbindingen, die relatief aan de grootste parameter afgebeeld worden. Deze relatieve verbindingswijdtes geven een optimale visuele representatie van de relatieve groottes van de parameters, maar laten niet zien hoe goed de variabelen elkaar voorspellen op een absolute schaal. Om dit probleem te behandelen, heb ik in Hoofdstuk 3 voorgesteld om een knoospgewijze voorspelbaarheidsmaat te berekenen en om deze te verwerken in de netwerkvisualisaties. In Hoofdstuk 4 heb ik een heranalyse gedaan van de toen ontluikende empirische netwerk artikelen op het gebied van psychopathologie, en heb ik de nadruk gelegd op voorspelbaarheid en hoe voorspelbaarheid een theoretisch interessante kwantiteit kan zijn, die aan kan geven tot in hoeverre een systeem zelfvoorzienend kan zijn.

Een cruciale beperking van de statistische netwerkmodellen die gebruikt worden in de netwerk literatuur, is dat deze modellen enkel interacties tussen paren van variabelen bevatten. In andere woorden, deze modellen kunnen geen moderatie effecten modelleren. Dit is een grote beperking omdat moderatie effecten erg plausibel zijn in een contextgebonden veld zoals psychologie. Moderatie effecten zijn ook van centraal belang in de netwerkmethodologie voor psychopathologie, omdat deze effecten kunnen wijzen op mogelijke effecten van interventies op de knopen van het netwerk (Borsboom, 2017). Om zulke moderatie effecten te vinden, heb ik in Hoofdstuk 5 Moderated Network Models (MNMs) geïntroduceerd. Deze modellen maken het mogelijk om elke paarsgewijze interactie te modelleren als een functie van alle andere variabelen in het model. Omdat MNMs geïmplementeerd zijn binnen het MGM kader, is deze methode in staat om een grote verscheidenheid aan modellen te schatten, zoals paarsgewijze interacties tussen continue en categorische variabelen die tegelijkertijd gemodereerd worden door zowel continue als categorische variabelen. Dit impliceert dat een MNM met een enkele categorische moderatorvariabele een alternatief kan bieden om verschillen te detecteren tussen verschillende groepen, zonder gebruik te maken van meerdere stappen zoals ‘resampling’ en significantietoetsing.

Het centrale idee van de netwerkbenadering van psychopathologie is dat mentale stoornissen voortkomen uit causale interacties tussen symptomen. Dit suggereert dat interacties tussen symptomen in gezonde en ongezonde individuen verschillen, en dat deze interacties verschillen per individu wanneer het individu zich in een transitie tussen een gezonde en ongezonde staat bevindt. Om zulke transities te detecteren in individuele tijdreeksen, heb ik in Hoofdstuk 6 een methode geïntroduceerd om VAR modellen te schatten die variëren over tijd. Ook heb ik gestipuleerd tot in hoeverre de parameters van deze modellen gedetecteerd kunnen worden in realistische data door middel van een uitgebreide simulatiestudie. Hoewel dit hoofdstuk zich voor de eenvoudigheid richt op VAR modellen, is het R-pakket *mgm* ook in staat om de parameters te schatten van MGMs en mVAR modellen die variëren over tijd. Deze modellen zijn cruciaal voor het beantwoorden van diverse onderzoeksvragen: ze kunnen gebruikt worden voor het detecteren van veranderingen in de structuur van de interacties in observationele onderzoeken en voor het verklaren van deze veranderingen door middel van bijkomende variabelen. Een andere toepassing van deze methode is het monitoren van patiënten en om de tijdsvariërende modellen te gebruiken als multivariate “Early Warning Signals” (EWS; Scheffer et al., 2009). Op deze manier kunnen de periodes vastgesteld worden waarin een behandeling het effectiefst is (Olthof et al., 2019). Ten slotte kunnen tijdsvariërende modellen gebruikt worden om te bestuderen hoe de structuur van een interactie verandert als gevolg van een behandeling (Wichers et al., 2016).

Hoewel Hoofdstukken 2 - 6 gericht zijn op het beschikbaar maken van nieuwe modellen voor data voor toegepaste onderzoekers, zijn de laatste twee hoofdstukken over data modellen toegespitst op het oplossen van specifieke methodologische vraagstukken. In Hoofdstuk 7 bespreek ik de vertekening-variantie afweging en modelselectie in de context van de keuze tussen het VAR model en een speciaal geval van het VAR model, het AR model, wat enkel de autore-



gressieve coëfficiënten bevat. Naast het bespreken van deze theoretische uitdaging, bevat dit hoofdstuk een simulatiestudie die aantoont hoeveel waarnemingen er nodig zijn alvorens het VAR model beter presteert dan het AR model, door middel van simulaties uit verschillende types VAR modellen. In Hoofdstuk 8 bespreek ik hoe de interpretatie en dynamiek van het Ising model veranderen als het domein van de binaire variabelen verandert van  $\{0, 1\}$  naar  $\{-1, 1\}$ . Dit hoofdstuk toont aan dat de gebruikelijke aanname dat sterk verbonden netwerken leiden tot een verhoging van de symptomen, over het algemeen niet waar is, maar dat het resultaat afhangt van de eigenschappen van het gebruikte model.

## **J.2 Formele Theorieën**

In het tweede deel van mijn dissertatie bespreek ik hoe formele theorieën opgesteld kunnen worden met behulp van datamodellen, en introduceer ik een formele theorie over de paniekstoornis.

Tot nu toe berust het merendeel van het modellerwerk in psychopathologie op het passen van statistische tijdreeksmodellen. In Hoofdstuk 9 hebben we onderzocht tot in hoeverre er inferentie gedaan kan worden over onderliggende systemen in de psychopathologie door middel van tijdreeksmodellen. Dit hebben we als volgt gedaan: we hebben een bistabiel systeem voor emotionele dynamieken gedefinieerd als het ware onderliggende systeem. Dit systeem was gekozen omdat het plausibel is voor veel psychologische fenomenen en ingewikkelder is dan de meeste populaire tijdreeksmodellen, ondanks dat dit systeem relatief makkelijk te begrijpen is. Nadat het systeem gedefinieerd was, hebben we de eigenschappen van het systeem proberen terug te schatten door tijdreeksmodellen toe te passen op de data die gegenereerd was door het ware systeem. We hebben hierbij gefocust op twee uitdagingen. De eerste uitdaging is dat het ware systeem hoogstwaarschijnlijk niet een speciaal geval is van het tijdreeksmodel, wat betekent dat het tijdreeksmodel verkeerd gespecificeerd is. We hebben aangetoond dat het in zulke situaties erg moeilijk is om betrouwbare inferentie te doen met het tijdreeksmodel over de dynamieken van het ware systeem. De tweede uitdaging is dat we het systeem moeten meten met voldoende frequentie. We tonen dit probleem aan door het generen van een typische ESM tijdreeks vanuit het ware systeem, en laten vervolgens zien dat de steekproeffrequentie, voor het huidige systeem, te laag is om de dynamieken van het ware systeem te achterhalen. Dit probleem doet zich waarschijnlijk voor in veel toepassingen, zoals het bestuderen van de emotie dynamieken (op een tijdschaal van seconden of minuten) met ESM metingen (op een tijdschaal van uren). Hoewel het moeilijk is om de dynamieken op een korte tijdschaal te achterhalen, hebben we ook aangetoond dat het mogelijk is om op een betrouwbare wijze de globale eigenschappen van het systeem te beschrijven. We gebruiken deze resultaten om een methode te onderbouwen voor het construeren van formele theorieën over dynamieken binnen het individu die verder gaan dan het enkel toepassen van tijdreeksmodellen. Deze methode is uitgewerkt in Hoofdstuk 11.

In Hoofdstuk 10 introduceer ik een formele theorie over paniekstoornis, waar

ik heb bijgedragen aan het ontwikkelen en implementeren ervan. Om deze theorie te ontwikkelen, hebben we een benadering gebruikt die ontwikkeld is door van der Maas et al. (2006) om het mutualisme model voor intelligentie op te stellen, en door Dalege et al. (2016) om het Causal Attitude Network (CAN) verder te ontwikkelen. In plaats van beginnen met het toepassen van een data-model op een specifieke dataset, begint deze aanpak met het opsommen van door literatuur onderbouwde feiten. In het geval van paniekstoornis betekende dit het vinden van de onderliggende componenten van de stoornis en hoe deze componenten zich tot elkaar verhouden, maar ook relatief simpele empirische basisbeginselen zoals de typische duur van een paniekaanval (5-20 minuten). We hebben eerst de kerndynamieken van paniekaanvallen geïmplementeerd door een model uit de ecologie dusdanig aan te passen, dat het in staat was om realistische paniekaanvallen te produceren. Daarna hebben we een langzaam leerproces toegevoegd, wat nodig was om paniekstoornis te verklaren. Tot slot hebben we de formele theorie aangepast zodat het plausibele voorspellingen kon maken, en om het consistent te maken met de empirische basisbeginselen van een paniekstoornis. In Hoofdstuk 11 wordt het proces waarmee deze theorie opgesteld is, verder ontwikkeld tot een expliciet kader voor het vormen van theorieën.

In Hoofdstuk 11 hebben we drie verschillende routes onder de loep genomen om van datamodellen naar formele theorieën te gaan. De eerste route is om datamodellen te behandelen als zijnde formele theorieën; de tweede route is om inferentie te doen op basis van datamodellen om formele theorieën op te stellen; de derde route is om datamodellen te gebruiken om theorieën op te stellen middels een abductieve aanpak. De eerste route "heeft weinig kans van slagen" omdat datamodellen over het algemeen niet dezelfde complexiteit delen als de stoornissen die gemodelleerd worden, of omdat het onrealistisch is om de modellen te schatten op basis van de data. De tweede route is problematisch omdat het doorgaans onduidelijk is hoe we inferentie doen over formele theorieën op basis van datamodellen. Hierdoor stellen wij voor de derde route te gebruiken, waar de datamodellen op een abductieve manier gebruikt worden om theorieën te vormen. Bij deze route worden datamodellen afgeleid uit concurrerende formele theorieën, om deze vervolgens te gebruiken om te evalueren welke formele theorie het beste bij de data past. Tevens geven wij een expliciete beschrijving van de theorievorming methodologie die gebruikt wordt in Hoofdstuk 10 middels een procedure. Dit kader bestaat uit de volgende vier stappen: vaststellen van het fenomeen, formulering, ontwikkeling, en testen van de formele theorie. Hiermee geven we een algemene methodologie om theorieën te vormen over mentale stoornissen.

## ACKNOWLEDGEMENTS

---

Luckily, writing my dissertation was not a journey I had to make alone. I was very fortunate that a lot of people contributed in many ways to the four years of my PhD.

First of all, I would like to thank my two supervisors. Lourens, you were already the supervisor for my master's thesis, and most of what I know about statistics I learned from you. I was extremely fortunate to be able to rely on your vast knowledge about methodology, statistics, and mathematics. No matter which topic I brought up, it always seemed like you already happened to have worked through a couple of text books on it. Thank you for patiently walking me through all sorts of statistical territory that was way over my head. I also very much appreciated the clarity of your explanations and feedback. After our meetings I always knew where I stood and had new ideas and material I could use to move forward. Next to working on specific projects, I also very much enjoyed our many discussions about statistics and science and going on all sorts of random tangents.

Denny, you do not know this but we already "met" in 2011. That summer I ranted to a psychometrician at the University of Luxembourg about how nothing makes sense in psychological measurement, upon which she recommended an interesting book to me of some "crazy guy from Amsterdam". That book turned out to be your PhD dissertation. Two years later I saw a talk you gave in front of international students in which you gave us the advice that if we feel that something is wrong in science we should ask questions and investigate, and not let ourselves be convinced by authority that things are fine as they are. This spirit resonated a lot with me and from then on I knew where I wanted to do my PhD. I therefore felt incredibly fortunate to indeed be able to start my PhD research in your group in 2015. I am very impressed by your vast scientific interests, your enthusiasm and optimism, and your imagination for how things can be different. I thoroughly enjoyed all our meetings, during which we hardly ever managed to talk about what I had prepared but almost immediately got side-tracked into fundamental debates about psychometrics and philosophy of science.

I would like to thank both of you for creating such a fantastic environment for me to learn and do research. I never felt any pressure to get anything done quickly, which gave me the confidence to embark on bigger projects. You provided me with a lot of ideas, but at the same time gave me complete freedom in which projects to pursue. This way I experienced my PhD research as four years during which I could get up every morning and study something I find interesting, which I recognize as an extreme privilege. Finally, I really appreciated that

## *K. Acknowledgements*

---

you were not only interested in my academic development but also in that I am doing well and that I am enjoying my work.

Oisín, I am very happy that we started working together. Thank you for all the ludicrous rants and fights about wide-ranging topics. I greatly enjoyed resolving questions about system recovery and time series analysis by shouting at each other for hours, and I also appreciated the lengthy discussions about where to put the “also” in the sentence and whether to capitalize after the colon. Also, I am happy that we did not murder each other during our visit in Boston, but instead wrote at least one nice paper together.

Don, thank you for putting up with Oisín and me for a month in Boston and investing so much time to really start from scratch and think through how to build formal theories about psychopathology. Our month-long discussion was one of the most productive and fun academic experiences I ever had. I wish I could spend most of my work time on fundamental discussions like this. I am inspired by your determination to take the time to really get to the bottom of things. And you made me a better writer through the slightly infuriating experience of seeing you rewrite my paragraphs, which in the end still contained the same information but read a hundred times better.

Eiko, thank you for many discussions about psychopathology research and psychometrics and for a very nice collaboration on our predictability paper. Also thank you for allowing me to constantly harass you with questions about research on psychopathology and that very specific example data set I just happen to need. Laura, thank you for our discussions about time series analysis and for our nice collaboration on time-varying models. Dirk and Michael, it was a pleasure to work with you at various locations at a record-breaking pace on our mouse-tracking papers. I am looking forward to future meetings and I hope we will manage to publish our main paper before there are no more computer mice.

Jonas, thank you for patiently explaining to me every little detail on how to graduate and for sending me the LaTeX file of your dissertation, so I only had to change the text and the surname. Without you I would have graduated half a year later. Also thanks for all the fun discussions and for being part of our ridiculous hypercube project. For now it is put on hold, but I think it can still inspire many other papers we also really should not write.

Sacha, thanks a lot for many fun discussions, for helping me out with all sorts of (programming) issues especially during the beginning of my PhD, and for reacting so quickly to qgraph bugs (usually by declaring them as features). And thanks for making those pretty circles around the nodes!

Marie and Tessa, thank you for the cozy coffee and lunch sessions and for being part of our Fit Girls group. My heart was broken when Marie left to Berlin, but maybe we can reunite for some Super Flow once the virus blew over. Adela and Sacha, thank you for all the nice activities in and around the UvA and the cozy board game nights. Thank you for organizing the great summer and winter schools and for letting me steal the sweets for the participants.

Dear inhabitants of G040, Riet, Adela, Lisa, Gaby, Pia, Ria, Angelika and Claudia, thank you for creating such a nice office atmosphere including yoga, cocktails and Greek dancing. Thank you to all members of the Psychosystems

---

group and other PML/UvA people, Denny, Lourens, Han, Angelique, Mijke, Eiko, Max, Maarten, Sacha, Claudia, Riet, Jonas, Marie, Tessa, Jolanda, Adela, Julian, Ceren, Gaby, Lisa, Pia, Ria, Quentin, Akash, Alexandra, Angelika, Joost, Udo, Fabian, Johnny, Don, Don, Alexander, Dylan, Alexander and Nathan for many stimulating discussions and for creating such a fantastic work environment.

I would like to thank Han van der Maas, Eric-Jan Wagenmakers, Ellen Hamaker, Casper Albers, Don Robinaugh, Maarten Marsman and Sacha Epskamp for taking the time to plow through my dissertation and to be in the committee for my defense.

Kees, Erik-Jan, Laura, Machiel, Annette, Elian and Pascal thank you for many fun board meetings and for organizing a couple of very cool Statistics Cafés, Workshops and Company Visits with the Young Statisticians Netherlands.

Katha, Ivan and Peter thanks for the invitations and visits, the many fun activities together and for many interesting scientific discussions over distance. Dawid, Mark, Matteo and Joris, thank you for an interesting and fun excursion into the world of conflict studies. And thanks to Niccolo and Yuki for those conversation vacillations and the many fun times in Santa Fe and elsewhere.

Johnny, you are certainly one of the top discoveries during my PhD and I am very lucky that I already made it so early on. Thank you for being such an important part of my Amsterdam adventures and for never telling me anything about statistics. Also thanks for helping me to cover up my questionable Dutch language skills in this dissertation, and for being one of my paranympths.

Fabian, good to have you in here Amsterdam after we just missed each other in Austria. Thank you for the many amazing shared experiences, for being one of my favorite collaborators (hypercube!) and a constant intellectual inspiration. Also thank you for being one of my paranympths.

I would also like to thank friends outside the UvA PML circle. Giacomo, we did not end up finding that house but luckily we went on quite a journey nonetheless. Thank you for everything. And thank you for patiently drawing, re-drawing and re-arranging 3D objects and triangles for many hours to create the cover of this dissertation. Joris, thank you for so many amazing shared experiences and all this craziness over so many years, for always surprising me and for everything else. Damiano, thank you for all the good vibes and fun times together and for heating up the Netherlands with your Sicilian fire. It always works! Javier, I am very happy that we finally met in New Mexico after applying for the same PhD, working on the same campus and basically living next to each other for years. It's great to have you as a part of my Amsterdam crew. Daniel, thank you for all the dinners, concerts and walks, I very much enjoy your slow pace, the self-mockery and all the bullshit talk. Ani, thank you for a bumpy but amazing ride. Thank you all for being such amazing people and for making Amsterdam my new home.

Fuchs, Luci, Sam, Bumal, Gu, Andi, Marina, Sabrina, Simon and other non-Amsterdammers, thank you for your visits for the Sommerknallers and other questionable activities. Your visits alone reduced my dissertation by at least one chapter (totally worth it). Max, thank you for being a brother I can really connect to. Mama & Papa, thank you for all your love and support, for always believing in me, and for never telling me what to do so I could find my own path.