



Taka, Evdoxia (2023) *Interactive animated visualizations of probabilistic models*. PhD thesis.

<https://theses.gla.ac.uk/83903/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Interactive Animated Visualizations of Probabilistic Models

Evdoxia Taka

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow



University
of Glasgow

April 2023

Abstract

Bayesian probabilistic models' structure (determined by the mathematical relations of the model's variables) and outputs (i.e., the posterior distributions inferred through Bayesian inference) are complex and difficult to grasp and interpret without specialized knowledge. Various visualizations of probabilistic models exist but it is very little known about whether and how they support users' comprehension of the models. The aim of this thesis is to investigate whether adding interaction or animation to visual representations of probabilistic models help people better understand the structure of models and interpret the (causal and non-causal) relations of the variables.

This research presents a generic pipeline to transform a probabilistic model expressed in a Probabilistic Programming Language (PPL) and associated inference results into a standardized format which can then be automatically translated into an interactive probabilistic models explorer (IPME). IPME provides at-a-glance communication of a model's structure and uncertainty, and allows interactive exploration of the multi-dimensional prior or posterior MCMC sample space. A collapsible tree-like structure represents the structure of the model in IPME. Each variable is represented by a node that presents graphically the prior or posterior distribution of the variable. Slicing on indexing dimensions or forming conjunctive restrictions on variables by interacting with the distribution visualizations is supported. Each user interaction with the explorer triggers the reestimation and visualization of the model's uncertainty. This closed-loop exchange of responses between the user and the explorer allows the user to gain a more intuitive comprehension of the model. IPME was designed to enhance informativeness, transparency and explainability and ultimately, the potential of increasing trust in models.

This research investigates also whether adding interactive conditioning to classical scatter plot matrices that present samples from the prior distribution of probabilistic models helps users better understand the models, and if there are levels of structural detail and model designs for which it is beneficial. A user study was conducted. The analysis of the collected data showed that interactive conditioning is beneficial in cases of sophisticated model designs and the difference in response time between the interaction and static group becomes less important in higher levels of structural detail. Participants using interactive conditioning were more confident about their responses overall with the effect being stronger in tasks of lower level of detail.

This research proposes a pipeline to generate simulated probabilistic data from interven-

tions applied on causal structures that are expressed in PPLs using probabilistic modeling and Bayesian inference. An automatic visualization tool for visualizing the simulated probabilistic data generated by this pipeline was developed. A user study to evaluate the proposed tool was conducted. How effectively and efficiently people identify the causal model of the presented data and make decisions on interventional experiments when the uncertainty in the simulated data of interventions was presented using static, animated, or interactive visualizations was investigated. The findings suggested that participants were able to identify the causal model of the presented data either given a single intervention or by exploring various interventions. Their performance in identifying sufficient interventions was poor. Participants did not rely on the sufficient interventions to identify the causal model in the case of multi-interventional tasks. They might have relied more on combining information from multiple interventions to draw their conclusions. There were three different visual exploration strategies of the information in the scatter plot matrices which participants followed; roughly 1/3 of them relied on both the scatter and KDE plots, another 1/3 of them relied more on the scatter plots, and the last 1/3 of them relied more on the KDE plots. Those who followed the last strategy had a better performance in identifying the causal model given a specific intervention. Most participants judged the design of the visualization positively with many having mentioned that “it was informative”.

Contents

Abstract	i
Acknowledgements	xxi
Declaration	xxii
Acronyms	xxiii
1 Introduction	1
1.1 Summary	1
1.2 Probabilistic Models: Opportunities and Obstacles	1
1.2.1 Opportunities	1
1.2.1.1 Modeling Data Generating Mechanisms	1
1.2.1.2 Modeling Uncertainty	3
1.2.1.3 Incorporating Prior Knowledge	5
1.2.1.4 Estimating Uncertainty through Bayesian Inference	6
1.2.1.5 Accounting for Uncertainty when Modeling Causal Relations	9
1.2.2 Obstacles	11
1.3 Communicating Probabilistic Models through Visualization	12
1.3.1 How is Visualization Used to Communicate Probabilistic Models?	12
1.3.1.1 Visualizing the Structure of Probabilistic Models	13
1.3.1.2 Visualizing the Inference Results	13
1.3.2 Challenges in Visualizing Probabilistic Models	15
1.4 This Work	16
1.4.1 Thesis Statement, Aims and Scope	16
1.4.2 Contributions	16
1.4.3 Research Challenges	17
1.4.4 Thesis Outline	18
2 Theory	20
2.1 Summary	20

2.2	Bayesian Inference	21
2.2.1	Schools of Statistical Inference and Definition of Probability	21
2.2.2	Summary of Useful Probabilistic Concepts	22
2.2.3	Bayes' Rule	23
2.2.4	Difficulty with the Computation of Bayes' Rule	24
2.2.4.1	Explaining Complexity of Bayes' Rule Components	24
2.2.4.2	The Difficulty with the Denominator	28
2.2.4.3	Markov Chain Monte Carlo	28
2.2.5	Predictions	29
2.2.6	Example of Bayesian Inference	29
2.2.7	Why to Use Bayesian Statistics for the Analysis of Research Data?	31
2.3	Probabilistic Programming Languages	33
2.3.1	Purpose	33
2.3.2	Expressing a Probabilistic Model in a PPL	33
2.3.3	From the Definition of a Model to a Trace of Samples	35
2.4	Why Does Modeling Causal Relations Require Extra Modeling Methodologies?	35
2.4.1	Correlation is not Always Causation	35
2.4.2	How can I tell if a correlation is a true causal effect?	37
2.4.3	Causal Modeling Methodologies	38
2.5	Theories of Human Visual Perception	41
2.5.1	Why is it Important to Understand Human Visual Perception?	41
2.5.2	Visual Perception of Uncertainty	41
2.5.2.1	Heuristics and Biases When Judging Probabilities	41
2.5.2.2	Frequency Framings of Probability Improve Probability Judgments	42
2.5.3	Visual Perception of Animation	43
2.5.4	Visual Perception of Interaction	44
3	Literature Review	45
3.1	Summary	45
3.2	Visualization of Uncertainty	46
3.2.1	Existing Approaches	46
3.2.1.1	Static Visualisations of Uncertainty	47
3.2.1.2	Animated Visualisations of Uncertainty	49
3.2.1.3	Interactive Visualisations of Uncertainty	50
3.2.2	Evaluation of Uncertainty Visualizations	52
3.2.2.1	Evaluation of Decision-Making Under Uncertainty	52
3.2.2.2	Evaluation of Uncertainty Visualization Through Comparison of Different Designs	53

3.2.2.3	Evaluation of Uncertainty Visualization in Bayesian Reasoning	57
3.2.3	Challenges in Designing and Evaluating Uncertainty Visualizations . . .	58
3.3	Visualization of Models' Structure	59
3.3.1	Probabilistic Models	60
3.3.1.1	Bayesian Network	60
3.3.1.2	Existing Tools for Graphical Representation of Probabilistic Models	62
3.3.2	Causal Models	63
3.3.2.1	Causal Diagrams	63
3.3.2.2	Existing Visualization Tools for Causal Reasoning and Explo- ration	64
3.3.2.3	Evaluation of People's Causal Reasoning with Visualization .	65
3.4	Inspiring Ideas From the Literature on Interactive Visualization	65
4	Using Interaction for Visualizing Probabilistic Programming Models	68
4.1	Summary	68
4.2	Purpose	68
4.3	Cognitive and Practical Tasks that Users of Bayesian Probabilistic Models Un- dertake	70
4.4	Interactive Probabilistic Models Explorer	75
4.4.1	PPL Model Encoding	75
4.4.1.1	Model-related Information	75
4.4.1.2	Inference-related Information and Data	77
4.4.2	Design and Implementation of IPME	78
4.4.3	Objectives of IPME	83
4.5	What Makes IPME a Unique Tool	84
4.5.1	Presentation of the Probabilistic Programming Model's Graph	85
4.5.2	Presentation of the Inference Results	85
4.5.3	Comparison of IPME with Existing Visualization Libraries for Bayesian Analysis	88
4.6	Use Case Scenarios	90
4.6.1	Drivers' Reaction Time	90
4.6.1.1	Model Check	91
4.6.1.2	Interactivity	96
4.6.2	Stochastic Volatility	101
4.6.2.1	Model Check	101
4.6.2.2	Interactivity	103
4.6.3	Coal Mining Disasters	107
4.6.3.1	Interactivity	109

4.7	Discussion	111
4.7.1	Contributions	111
4.7.2	Limitations	112
4.7.3	Future Work	113
4.7.4	Conclusions	115
5	Using Interactive Conditioning for Supporting Users' Understanding of Probabilistic Models	116
5.1	Summary	116
5.2	Purpose	116
5.3	Relations of Probabilistic Models' Variables	119
5.3.1	A Hierarchy of Variables' Relations	119
5.3.2	Visualization of Variables' Relations	120
5.4	Interactive Pair Plot	122
5.5	Research Questions, Tasks & Conditions	124
5.6	User Study's Design	126
5.6.1	Task Models' Design	127
5.6.2	Implementation Details	128
5.7	Analysis and Results	128
5.7.1	Evaluation Measures	128
5.7.2	Data & Bayesian Modeling of Responses	129
5.7.3	Results	130
5.7.4	Comparative Analysis	132
5.7.5	Analysis of Interaction Logs	133
5.8	Discussion	134
5.8.1	When is Interactive Conditioning (not) Beneficial?	134
5.8.2	Practical Implications	134
5.8.3	Limitations of User Study	135
5.8.4	Future Work	137
5.8.5	Conclusions	137
6	Visualizations of Simulated Data of Interventions to Support Users' Causal Reasoning and Decisions on Interventions	139
6.1	Summary	139
6.2	Purpose	140
6.2.1	Motivation	140
6.2.2	Purpose of This Work	140
6.2.3	Summary of This Work	141
6.3	From Observed Data to Simulated Probabilistic Data of Interventions	142

6.3.1	Approaches to Causal Inference	142
6.3.2	Using Bayesian Probabilistic Models to Simulate Interventions	145
6.3.2.1	Probabilistic Modeling of Causal Structures	146
6.3.2.2	Simulating Interventions Probabilistically	149
6.4	The Visualizer of Causal Structures' Simulated Interventions	151
6.4.1	Objectives of Design	151
6.4.1.1	Inclusion of Uncertainty	151
6.4.1.2	Parallel Presentation of the Simulated Data Before and After the Intervention	152
6.4.1.3	Visualization of the Pairwise Distributions of the Simulated Data	152
6.4.1.4	Exploration of the Outcomes From Applying Different Inter- ventions	152
6.4.1.5	Joint Presentation of Causal Models & Simulated Inter- ventional Data	153
6.4.1.6	Automatic Transformation of Simulated Interventional Data Into Visualizations	153
6.4.2	Design & Implementation	153
6.4.2.1	Input	154
6.4.2.2	The Components of the Design	155
6.4.2.3	The Design of the Scatter Plot Matrix	155
6.4.2.4	The Presentation Problem of the Interventional Data	156
6.4.2.5	The Design of Causal Diagrams	159
6.4.2.6	Implementation and Limitations	159
6.4.3	Use Case	160
6.4.3.1	Atomic Intervention On Insomnia	162
6.4.3.2	Shift Intervention On Tiredness	164
6.4.3.3	Variance Intervention On Anxiety	166
6.5	Evaluation User Study	166
6.5.1	Research Questions, Conditions & Participants	166
6.5.2	Study's Structure	169
6.5.3	Design of Study's Tasks	173
6.5.4	Statistical Analysis & Results	175
6.5.4.1	Evaluation Measures & Data	176
6.5.4.2	Bayesian Analysis & Levels of Analysis	176
6.5.4.3	Analysis of Accuracy in TT1 Tasks	177
6.5.4.4	Analysis of Accuracy in TT2 Tasks	183
6.5.4.5	Analysis of Response Times and Confidence	187

6.5.5	Other Analyses	190
6.5.5.1	Did participants' statistical or causal inference literacy play a role in their performance?	190
6.5.5.2	Did the visual exploration strategy followed by participants play a role in their performance?	191
6.5.5.3	Did the visualization condition favor a specific visual exploration strategy?	192
6.5.5.4	Are participants' responses valid?	192
6.5.5.5	How did participants judge the design of the visualization?	194
6.5.6	Summary of Findings	194
6.5.7	Limitations of the User Study	197
6.6	Discussion	198
6.6.1	Contributions	198
6.6.2	Future Work	200
6.6.3	Conclusions	202
7	Conclusions	203
7.1	Summary	203
7.2	Purpose of Research	203
7.3	Summary of Findings	204
7.4	Conclusions & Future Directions	206
A	Examples of Models	209
A.1	Average Minimum Temperature in Scotland	209
A.2	The Eight Schools Model	210
A.3	Drivers' Reaction Time Models	212
A.4	Stochastic Volatility Model	215
A.5	The Coal Mining Disasters' Model	216
B	<i>arviz_json</i> Package	219
B.1	Graph JSON Structure	219
B.2	<i>ArviZ InferenceData</i> Export Into <i>numpy</i> Arrays and Metadata File	220
C	User Study 1	224
C.1	Participants' Training	224
C.2	Task Models	225
C.2.1	Model 1	225
C.2.2	Model 2	225
C.2.3	Model 3	226
C.3	Analysis	227

C.3.1	Accuracy's Model	227
C.3.1.1	Model for T1	227
C.3.1.2	Model for T2-T3	228
C.3.2	Response Times' Model	229
C.3.3	Confidence's Model	230
C.4	Tasks	231
D	User Study 2	243
D.1	Generation of Synthetic Observations for the Insomnia-Anxiety-Tiredness Problem	243
D.2	Participants' Training	244
D.3	Tasks	244

List of Tables

- 1.1 The reaction times of a driver after n consecutive days of driving under sleep-deprivation conditions. The dataset is from [Belenky et al., 2003] and can be retrieved from [Lambert, 2018b] for all 18 available drivers. 7
- 1.2 Posterior statistics in a tabular format for the drivers’ reaction time probabilistic model. The 97% HDI is included in the statistics: the left end of the interval shown in column HDI_1.5% is represented by the value below which 1.5% of the posterior falls and the right end of the interval shown in column HDI_98.5% is represented by the value above which 1.5% of the posterior falls. 14
- 4.1 Posterior statistics in a tabular format for the eight schools’ hierarchical model. This is a rather simple model and the table only consists of ten rows. This number could rise immensely if the model had more parameters or the parameters had more (multi-valued) indexing dimensions. 87
- 4.2 Comparative presentation of existing Bayesian analysis visualization libraries including IPME. IPME offers unique features that are not encountered in any other of the existing tools; the interactive exploration of the (MCMC) sample space and the graphical analysis of the model. 89
- 5.1 **Summary of probabilistic models and tasks used in the user study.** 125
- 6.1 Summary of TT1 tasks of user study. 171
- 6.2 Summary of TT2 tasks of user study. The question asked in all tasks was the following: “*You want to design and run **one** interventional experiment, which will help you identify the causal model of the data. Which of the provided interventions (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.*” The ticked boxes in the interventions column represent the correct response in each task. 172

6.3 Participants' responses to the UE4 (*How did you find the design of the visualization for the simulated data? Was it informative, annoying, distracting?*) shown in the first column, the condition each was assigned to shown in the second column, and the code with which each comment was tagged shown in the third column. 195

List of Figures

1.1	Causal model of insomnia, anxiety, and tiredness.	3
1.2	Parameter estimates and reaction times' predictions from the probabilistic and non-probabilistic model of the drivers' reaction time problem.	9
1.3	Causal model of insomnia, anxiety, and tiredness.	10
1.4	Various graphical representations of the probabilistic model of the drivers' reaction time problem (Box 1.3 and 1.4). (a) Bayesian network, (b) DoodleBUGs' graph, (c) PyMC's graph, and (d) Kruschke-style diagram.	13
1.5	Kruschke-style parameters' posterior distributions of the probabilistic model of the drivers' reaction time problem.	15
2.1	Homogeneous, heterogeneous, and hierarchical estimations of the regression line in the drivers' reaction time problem.	31
2.2	Definitions of the homogeneous probabilistic model of the drivers' reaction time problem in (a) probabilistic statements, (b) PPL code (PyMC), and (c) Kruschke-style diagram.	34
2.3	Correlation of the x and y variables and the causal diagrams of the three basic three-variable causal primitives of the x , y , and z variables with z being a (b) confounder, (c) mediator, or (d) collider.	36
2.4	Toy example from Huszár [2019] demonstrating the effect of simulated interventions in three similarly-correlated two-variable (x and y) systems. (a) Three simulators for the similarly-correlated two-variable systems and scatter plots of the x - y simulated observations, (b) the simulators altered to reflect the $x = 3$ intervention and the x - y scatter plots after the intervention, and (c) the underlying causal structure of each system before (upper row) and after the intervention (lower row).	39
3.1	Various graphical representations of the homogeneous probabilistic model of the drivers' reaction time problem (Box 2.1). (a) Bayesian network, (b) DoodleBUGs' graph, (c) PyMC's graph, and (d) Kruschke-style diagram.	60
3.2	Bayesian network of the hierarchical model of the drivers' reaction time problem defined in Box 2.3.	62

4.1	(a) Kruschke-style diagram of the two-parameter Bayesian model. (b) Posterior MCMC samples of the model's parameters. The samples that lie within the intersection of the value range restrictions are colored in orange. (c) The estimated posterior distributions and marginal distributions based on the entire posterior MCMC sample set (dark blue) and the restricted sample set (orange).	72
4.2	Flowchart showing the steps that a modeler should take to export the PPL-specified probabilistic model and the inference data into a standardized output, which will then be used as an input to the IPME module. The presented pipeline adds three new steps to the typical model specification and inference running routine; the export of the PPL model graph into a JSON structure; the export of the inference data into an <code>ArviZ InferenceData</code> object, where the structured description of the graph from the previous step is attached; and finally, the export of all these into a collection of <code>.npy</code> arrays and metadata in a zipped file. A concrete implementation of this pipeline is provided by the <code>arviz_json</code> module [Williamson, 2019], which was created for providing a standardized output for the PyMC3 models and inference data.	76
4.3	The IPME representation of the hierarchical model of the eight schools problem presented in Box 4.2. The KDE plots shown present the prior marginal distributions of the model's variables. PyMC3 was used for the model's specification and inference.	79
4.4	The collapsed IPME representation of the hierarchical model of the eight schools problem presented in Box 4.2. The toggle buttons were used to hide the KDE plots.	80
4.5	The IPME representation of the hierarchical model of the eight schools problem presented in Box 4.2. The KDE plots shown present the posterior marginal distributions of the model's variables.	81
4.6	The posterior predictive test statistics (<code>min</code> , <code>max</code> , <code>mean</code> , <code>std</code>) of the model of the eight schools problem presented in Box 4.2.	82
4.7	Graphical representations of the eight schools model.(a) DoodleBUGs graph, (b) PyMC graph using the Graphviz interface, (c) Kruschke-style diagram, (d) IPME representation.	86
4.8	The posterior densities of the eight schools' hierarchical model presented in the style introduced by Kruschke [2015]. Although, this is a rather simple model, we can see that it can produce 10 different uncertainty visualizations. This number could rise even more if the parameters of the model had more (multi-valued) indexing dimensions.	88

4.9	The prior IPME representation of the homogeneous drivers' reaction time model. The model predicts negative slopes, negative and very big (tens of seconds) values of reaction times a priori, which indicates that the priors might not fit well with the prior knowledge about the problem.	92
4.10	The prior IPME representation of the hierarchical drivers' reaction time model. The model predicts negative slopes, negative and very big (tens of seconds) values of reaction times a priori, which indicates that the priors might not fit well with the prior knowledge about the problem.	93
4.11	The posterior predictive test statistics of the (a) homogeneous and (b) hierarchical drivers' reaction time model. The hierarchical model improves the representation of the <code>min</code> , <code>max</code> , and <code>std</code> test statistics of the observations in the predictions.	95
4.12	The IPME representation of the drivers' reaction time models, where we set the driver to 308 and observe the posterior predictive distribution of the reaction times for the homogeneous and hierarchical model. The posterior predictive distribution of the reaction times for driver 308 and the (a) homogeneous and (b) hierarchical model. The homogeneous model does not reveal significant differences among drivers.	96
4.13	The IPME representation of the drivers' reaction time models, where we set the driver to 309 and observe the posterior predictive distribution of the reaction times for the homogeneous and hierarchical model. The posterior predictive distribution of the reaction times for driver 309 and the (a) homogeneous and (b) hierarchical model. The homogeneous model does not reveal significant differences among drivers.	97
4.14	The IPME representation of the drivers' reaction time hierarchical model. We set a condition <code>mu_p > 12.1</code> . The initial and re-estimated posterior predictive distribution of the reaction times for driver 310 on day 6 are shown.	99
4.15	The IPME representation of the drivers' reaction time hierarchical model. We set a condition <code>mu_p > 12.1</code> . The initial and re-estimated posterior predictive distribution of the reaction times for driver 335 on day 6 are shown.	100
4.16	The prior IPME representation of the stochastic volatility model.	102
4.17	The posterior IPME representation of the stochastic volatility model. The updated posterior graphical representation after restricting the <code>volatility</code> to values greater than -4.5 is shown.	104
4.18	The posterior IPME representation of the stochastic volatility model. The updated posterior graphical representation after restricting the <code>volatility</code> to values smaller than -5.1 is shown. The greater the value of the <code>volatility</code> , the more uncertain the model becomes about the predicted values of the <code>returns</code>	105

- 4.19 The IPME representation of the stochastic volatility model where we set the condition $v \in [21.0, 30.0]$. The (a) prior and (b) posterior graphical representation. The posterior distribution of `volatility` and the posterior predictive distribution of `returns` become tighter. More informative priors make the model to be more certain about its predictions. 106
- 4.20 The posterior IPME representation of the coal mining disasters model. 108
- 4.21 The IPME representation of the coal mining disasters model. The updated (a) prior and (b) posterior graphical representation after restricting the `switchpoint` to values within the interval 1893 – 1897. The posterior predictive PMF of the number of disasters in year 1890 shifts towards 4. This is reasonable as we assumed with our conditioning that the regulations changed at a later time. . . . 110
- 5.1 Visual representations of Model 1 of user study. **Definitions in Textual Languages:** (a) Probabilistic statements. (b) PPL code (PyMC) of model for Bayesian inference. A likelihood is defined for the observed variable `temperature` to account for the list `temp_list` of N observed temperatures for a set of `years`. **Graphs:** (c)-(f) Transcriptions of model in various graph representations. . . . 117
- 5.2 Visualizations of variables' relations in Model 1 of user study. **Joint & Marginal Distributions:** (a)-(c) The prior and posterior joint (3D surface plots) and marginal distributions (line plots on cube faces) of variables `temperature`, and `b`, `a`, or `c`, respectively. The yellow stars represent the observations in `temp_list`. **Scatter Plots:** (d)-(k) Samples and contours of variables' pairwise prior joint distributions. Conditioning facilitates the interpretation of scatter plots' shape. For example, conditioning on `b` in sequential increased ranges in (e)-(g), increases the mean value (white dot) of `temperature`'s distribution. **Interactive Conditioning with IPME:** (l)-(s) IPME-like representation. Interactive conditioning is applied on the prior marginal distributions of `b`, `a`, or `c` and the conditional marginal distributions are drawn (in orange). 118
- 5.3 IPP of Model's 1 variables. A selection box is dragged and drawn on the KDE plot of variable `b` restricting its range to $[12 - 40]$. The conditional marginal distributions of the variables are drawn (in orange) in the KDE plots on the diagonal and the samples in the restricted sample space are highlighted (in orange) in the scatter and rug plots. 123
- 5.4 (a) Task t_3 (Model 1 - T2) of user study. Participants in SG were shown a static pair plot. (b) The interactive pair plot participants in IG were shown instead. Both pair plots showed the minimum necessary subset of model's variables. . . . 124

5.5	Demographic statistics of participants in the user study. Both groups of participants (SG and IG) comprised of more older participants (D1). There was a slight gender imbalance between the groups with IG having more males and SG more females (D2). The educational background was generally well-balanced between the groups (D3), while participants in SG had a slightly higher former training in Statistics (D4, D5).	126
5.6	Probabilistic models used for the Analysis presented as Kruschke-style diagrams.	129
5.7	Results. Forest plot (94% HDI) of the posterior distributions of the probability of correct answer for IG (<i>thetaIG</i>) and SG (<i>thetaSG</i>), difference of θ s (<i>diff_of_thetas</i>), effect size of response times (<i>effect_size</i>) between IG and SG (normalised difference of duration), and difference of the estimated mean confidence of participants about their responses (<i>diff_of_means</i>). Tasks are presented vertically grouped per model.	131
5.8	Results. Pair plot of mean values of the posterior distributions of <i>diff_of_thetas</i> for the accuracy, <i>effect_size</i> for the response times and <i>diff_of_means</i> for the confidence. The fitted linear regression line is drawn with a 90% bootstrap confidence interval in each scatter plot.	133
6.1	Proposed pipeline for modeling a causal model probabilistically and simulating interventions.	147
6.2	Definition of the insomnia, anxiety, and tiredness model in (a) probabilistic statements, and (b) PyMC3 code. (c) Causal diagram of insomnia, anxiety, and tiredness.	148
6.3	Scatter plot matrix of the synthetic observed and simulated posterior predictive data of the insomnia-anxiety-tiredness causal model.	149
6.4	The (a) interactive, (b) animated, and (c) static visualization mode of the <code>vicausi</code> tool. The simulated data before the intervention is shown in blue in the interactive and animated visualization mode, and in green in the static. The simulated data after the intervention is shown in orange in the interactive and animated visualization mode, and in colors determined by the colorbar (colors are mapped to slices of data of increasing ranges) in the static.	156
6.5	The <code>vicausi</code> view presenting posterior predictive samples (before any intervention) from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem.	161
6.6	The <code>vicausi</code> view presenting posterior predictive samples before and after an atomic intervention on insomnia from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem. (a) The static visualization mode. (b)-(d) Three instances of the interactive visualization mode with each corresponding to a different interventional value.	163

6.7	The <code>vicausi</code> view presenting posterior predictive samples before and after a shift intervention on tiredness from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem. (a) The static visualization mode. (b)-(d) Three instances of the interactive visualization mode with each corresponding to a different interventional value indicated by the interactive slider at the top of the scatter plot matrix.	165
6.8	The <code>vicausi</code> view presenting posterior predictive samples before and after a variance intervention on anxiety from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem. (a) The static visualization mode. (b)-(d) Three instances of the interactive visualization mode with each corresponding to a different interventional value.	167
6.9	Demographic statistics of participants in the user study.	170
6.10	(a) Task t1 (TT1) and (b) task t11 (TT2).	174
6.11	Count plots of number of participants scoring a particular level of accuracy per task independently of the visualization condition (LA1) and based on the visualization condition (LA2) for the (a) <i>Single Correct Option (SCO)</i> , and (b) <i>Multiple Correct Options (MCO)</i> approach.	178
6.12	Kruschke-style diagrams of the analysis models for participants' accuracy in TT1 tasks. Two models for each one of the <i>Single Correct Option (SCO)</i> and <i>Multiple Correct Options (MCO)</i> approaches for measuring accuracy are created. (a) SCO LA1 Model for all participants independently of visualization condition. (b) SCO LA2 Model for participants in each visualization condition. (c) MCO LA1 Model for all participants independently of visualization condition. (d) MCO LA2 Model for participants in each visualization condition.	179
6.13	Forest plots of the posterior distributions inferred based on participants' accuracy in TT1 tasks. The results of the analysis models for both the <i>Single Correct Option (SCO)</i> and <i>Multiple Correct Options (MCO)</i> approaches for measuring accuracy are included. Forest plot of the (a) SCO LA1 Model for all participants independently of visualization condition, (b) SCO LA2 Model for participants in each visualization condition, (c) MCO LA1 Model for all participants independently of visualization condition, (d) MCO LA2 Model for participants in each visualization condition.	181

- 6.14 Forest plots of the posterior distributions inferred based on participants' accuracy in TT2 tasks. The first row presents the forest plots of results from the *Single Correct Option (SCO)* model for participants' selections of causal models inferred on the (a) LA1 (all participants' responses are taken together independently of visualization condition), and (b) LA2 (participants' responses are taken together in each visualization condition) level of analysis. The second row presents the forest plots of results from the *Multiple Correct Options (MCO)* model for participants' selections of sufficient interventions inferred on the (c) LA1, and (d) LA2 level of analysis. (e)-(f) The third row presents the forest plots of results from the same models as in the second row but considering only the responses of participants who got the model correct. 184
- 6.15 Kruschke-style diagrams of the analysis models for participants' response time and confidence in the user study's tasks and analysis on the (a) LA1 level for all participants independently of visualization condition, and (b) LA2 level for participants in each visualization condition. 188
- 6.16 Forest plots of the posterior distributions inferred based on the analysis model for the response time and confidence of participants in the user study's tasks. Forest plot of response time analysis on level (a) LA1 and (b) LA2. Forest plot of confidence analysis on level (a) LA1 and (b) LA2. 189
- 6.17 Strip and box plots of participants' performance measured as the number of TT1 tasks that each participant got correct plotted against their statistical (D3 question) and causal inference (D4 question) level of knowledge as these were recorded in the demographic questions. Blue dot markers are used in the strip plots to represent the observations (each dot represents a participant in the user study). 191
- 6.18 Participants responses in UE2 question (*In how many tasks in the study do you estimate that you looked at the scatter plots (plots presenting data with dot markers) in the scatter plot matrix?*) represented by circles and UE3 question (*In how many tasks in the study do you estimate that you looked at the KDE plots (plots presenting the variables' distribution) on the diagonal of the scatter plot matrix?*) represented by squares. The darker a shape is, the higher the performance of the participant is. 192

6.19	Participants responses in UE2 question (<i>In how many tasks in the study do you estimate that you looked at the scatter plots (plots presenting data with dot markers) in the scatter plot matrix?</i>) represented by circles and UE3 question (<i>In how many tasks in the study do you estimate that you looked at the KDE plots (plots presenting the variables' distribution) on the diagonal of the scatter plot matrix?</i>) represented by squares. The color of the shape represents the visualization condition to which each participant was assigned.	193
6.20	Participants responses in UE1 question (<i>In how many tasks in the study do you estimate that you looked at the directed graphs (DAGs) of the causal models?</i>) represented by circles. The color of the circles represents the visualization condition of the participant.	193
B.1	The <i>ArviZ InferenceData</i> data structure of the PyMC3 hierarchical model of the eight schools' problem. (a) The groups in which inference data is organized in the <i>InferenceData</i> data structure. (b) The <i>xarray.Dataset</i> that corresponds to the posterior group of the <i>InferenceData</i> object.	221
C.1	Model 1 - Task t1 (T1).	232
C.2	Model 1 - Task t2 (T2).	232
C.3	Model 1 - Task t3 (T2).	233
C.4	Model 1 - Task t4 (T2).	233
C.5	Model 1 - Task t5 (T3).	234
C.6	Model 2 - Task t6 (T1).	234
C.7	Model 2 - Task t7 (T2).	235
C.8	Model 2 - Task t8 (T2).	235
C.9	Model 2 - Task t9 (T2).	236
C.10	Model 2 - Task t10 (T3).	236
C.11	Model 2 - Task t11 (T3).	237
C.12	Model 3 - Task t12 (T1).	237
C.13	Model 3 - Task t13 (T1).	238
C.14	Model 3 - Task t14 (T2).	238
C.15	Model 3 - Task t15 (T2).	239
C.16	Model 3 - Task t16 (T2).	239
C.17	Model 3 - Task t17 (T2).	240
C.18	Model 3 - Task t18 (T3).	241
C.19	Model 3 - Task t19 (T3).	242
D.1	Common cause model of the insomnia-anxiety-tiredness problem considered in the user study.	243
D.2	Page 1 of training.	244

D.3	Page 2 of training.	245
D.4	Page 3 of training.	245
D.5	Page 4 of training.	246
D.6	Page 5 of training.	247
D.7	Page 6 of training.	248
D.8	Task t1 (TT1).	248
D.9	Task t2 (TT1).	249
D.10	Task t3 (TT1).	249
D.11	Task t4 (TT1).	250
D.12	Task t5 (TT1).	250
D.13	Task t6 (TT1).	251
D.14	Task t7 (TT1).	251
D.15	Task t8 (TT1).	252
D.16	Task t9 (TT1).	252
D.17	Task t10 (TT1).	253
D.18	Task t11 (TT2).	253
D.19	Task t12 (TT2).	254
D.20	Task t13 (TT2).	255
D.21	Task t14 (TT2).	256
D.22	Task t15 (TT2).	257
D.23	Task t16 (TT2).	258

Acknowledgements

First of all, I would like to thank my supervisors, John H. Williamson and Sebastian Stein. They offered me loads of support, encouragement, experienced advice, feedback, and inspirational discussions during our regular meetings. With their discreet presence and guidance, they were keeping me on track and at the same time, offering me a sense of freedom and choices. This is how the ideas presented in this thesis were born, blossomed, and turned into fruitful research.

Of course, all these would not have gone true without the financial support I received from the Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics, EPSRC Project: EP/R018634/1. The scholarship I was awarded funded me in the first 3.5 years of my PhD.

I could not omit my family from these acknowledgements. I would like to thank my parents, Anastasia and Demosthenes, for everything but foremost for their pedagogy to teach me and my sister *ethos*. An ethos of ideals inspired by the history of our ancestors: *aristeia* (excellence), justice, high spirits, faith.

Last but not least, I would like to thank my beloved husband, Demetrios, my greatest supporter, emotionally and materially. His encouragement, affection, and strong belief that I can make it were escorting me and giving me strength in all hard times during this journey. Not to mention his cooking skills that fed us both during the writing-up period.

To my family and
husband

Declaration

I declare that this thesis was composed by myself, and that the work contained herein is my own except where explicitly stated otherwise in the text.

The author's original work presented in this thesis has contributed to a number of publications that have been co-authored with Dr Sebastian Stein and Dr John H. Williamson:

- E. Taka, S. Stein, and J. H. Williamson. Increasing interpretability of Bayesian probabilistic programming models through interactive representations. *Front. Comput. Sci.*, 2: 52, 2020. ISSN 2624-9898. doi: 10.3389/fcomp.2020.567344. URL <https://www.frontiersin.org/article/10.3389/fcomp.2020.567344>
- E. Taka, S. Stein, and J. H. Williamson. Does interactive conditioning help users better understand the structure of probabilistic models? *IEEE Transactions on Visualization and Computer Graphics*, pages 1–12, 2023. doi: 10.1109/TVCG.2022.3231967

Evdoxia Taka

Acronyms

A

API Application Programming Interface. 14, 77, 85, 124, 220

C

CDF Cummulative Density Function. 53, 54, 71, 113

D

DAG Directed Acyclic Graph. 3, 60, 62, 63, 73, 75–78, 80, 85, 113, 154, 159, 164, 170, 219

H

HCI Human Computer Interaction. 58

HDI Highest Density Interval. x, xvi, 14, 31, 131, 180–182, 185–190

I

IPME interactive probabilistic models explorer. i, v, x, xiii–xv, 70, 75–78, 80, 82–85, 87–90, 92, 93, 96, 97, 99, 100, 102–108, 110–119, 121–124, 153–155, 157, 159, 201, 204, 206, 220

IPP interactive pair plot. xv, 116, 118, 119, 121–124, 135–137, 141, 152–157, 159, 204, 206

K

KDE Kernel Density Estimate. ii, xiii, xv, 14, 15, 47–49, 71, 79, 80, 82–84, 112, 113, 117, 120–124, 126, 128, 137, 155, 158, 162–164, 166, 170, 172, 191, 196, 197, 200–202, 206

M

MCMC Markov Chain Monte Carlo. i, x, 11, 17, 18, 20, 28, 30, 35, 68, 70–73, 75, 77, 78, 80, 83, 84, 89, 90, 111, 113, 115, 135, 148, 198, 200, 204, 220, 222

ML Machine Learning. 32

O

OLS Ordinary Least Squares. 7, 9

P

PDF probability density function. 22, 28, 47, 49

PMF probability mass function. xv, 22, 47, 107, 109, 110

PPL Probabilistic Programming Language. ii, xiii, xv, 11–14, 17–20, 33–35, 63, 68–71, 74–77, 83–85, 89, 90, 111, 112, 115–117, 120, 135, 139, 141, 144–146, 148–150, 160, 198, 200, 202, 204, 205, 220, 222

V

vicausi Visualizer of Causal Assumptions and Uncertainty-Aware Simulations of Interventions. xvi, xvii, 154–157, 159–163, 165, 167–170, 175, 199–201, 205, 206, 244

Chapter 1

Introduction

1.1 Summary

The main focus of the research presented in this thesis is how visualization could be used to represent probabilistic models and their outputs more effectively and efficiently for the users. Probabilistic models constitute the focal point of all visualization-related research efforts presented in this thesis. Section 1.2 explains what probabilistic models are. It discusses the advantages of probabilistic modeling and the obstacles in becoming widely adopted as a modeling and analysis approach. Section 1.3 presents how probabilistic models and their outputs are communicated (visually) based on the current practices and discusses the challenges in using visualization to achieve this. Finally, Section 1.4 presents the research scope, aims, contributions, and research challenges of this work in regards with visualizing probabilistic models. It also provides the outline of the thesis.

1.2 Probabilistic Models: Opportunities and Obstacles

1.2.1 Opportunities

1.2.1.1 Modeling Data Generating Mechanisms

Probabilistic models are models describing data generating processes. They are similar in purpose to any other type of model used in science, engineering, industry etc. A simple and straightforward description of what a model is, is given by Martin [2018, Chap.1]:

Models are simplified descriptions of a given system or process that, for some reason, we are interested in. Those descriptions are deliberately designed to capture only the most relevant aspects of the system and not to explain every minor detail. This is one reason a more complex model is not always a better one.

Models are designed to capture only the “relevant aspects of the system” that are of interest. A set of *variables* and *relations* (i.e., mathematical associations) are defined appropriately to describe these “relevant aspects of the system”. A model consists of two types of variables; the directly observed or measured variables called *observed*, and the unobserved hidden variables called *latent*. The latter are often referred as *parameters*. Parameters define the behaviour of the model through the mathematical associations relating them to each other or to the observed variables. In cases when more than one observed variables are considered in the model, these can also be associated to each other.

Box 1.1 presents a simple linear (non-probabilistic) regression model that consists of two parameters, a and b , and two observed variables, rt and n . The parameters are independent (one cannot be used to predict the other) of each other and are used to define the observed variable rt . The observed variable n is also used to define rt . This model can be modelled in a probabilistic way, as well, as it will be shown in Section 1.2.1.2.

Box 1.1 Linear regression model

An important task in a logistics company is the allocation of routes to drivers. The company wants to minimize the risk of accidents when allocating long routes to drivers that are susceptible to tiredness under sleep-deprivation conditions, and thus, wants to model drivers’ reaction time in regards with the number of consecutive days of driving under sleep-deprivation. Assuming that the reaction time of drivers increases linearly with the number of driving days under sleep deprivation, a linear regression model could be used for the prediction of drivers’ reaction time.

$$rt = b \cdot n + a, \quad (1.1)$$

where a is the intercept and b the slope of the regression line, n is the predictor (independent variable) representing the number of consecutive days of driving under sleep-deprivation with $n \in \{0, 1, 2, \dots, 9\}$, and rt is the dependent variable to be predicted representing the reaction time in milliseconds after n consecutive days of driving under sleep deprivation conditions.

All models do not necessarily define *causal* associations between variables, namely relations that determine which variable causes the other (which variables are the *cause* of others and which variables are the *effects* of others), based on the actual data generating mechanism that governs the modelled system. For example, in the linear regression model in Box 1.1 the independent variable (n) is used as a *predictor* for the dependent variable (rt). This association does not necessarily mean that n causes rt . It simply expresses the belief of the modeller that some kind of predictive relation exists between these two variables. Hence such variables’ relations in a model reflect simple predictor-predicted relations. These relations are used to compose a proxy for a data generating process with the aim to approximate the actual process as to the “relevant

aspects of the system”. The aim of such models is usually prediction or forecasting, in which cases a proxy for the data generating process suffices.

There are special types of models called *causal models* that can be used to model the cause-effect relations of observed variables. These can be graphical or mathematical models defining the causal relations among observed variables. The aim of these models is to explain the actual data generating mechanism. Box 1.2 presents a graphical causal model called *causal diagram*. This consists of a *Directed Acyclic Graph (DAG)* where variables are represented by nodes and the direct causal relations by directed edges starting from a variable-cause and pointing to a variable-effect.

Regression models can be used as mathematical models to describe the causal relations of variables under some assumptions [McElreath, 2020a, Chap. 5,6;McElreath, 2021; Westreich and Greenland, 2013; Bulbulia et al., 2021; Textor and Gilthorpe]. The independent observed variables would represent the causes of the dependent observed variable. In Box 1.2 a causal model of three variables is expressed mathematically by a set of two linear (non-probabilistic) regression models.

Box 1.2 A causal model

The causal diagram in Fig. 1.1 is a graphical causal model describing the causal relations among three observed variables; `insomnia`, `anxiety`, and `tiredness`.

The linear regression models 1.2 can be used as a mathematical model to describe the causal relations of these three observed variables (under some assumptions). Both models describe the following causal relations: `insomnia` causes `anxiety` and `tiredness`, and `anxiety` causes `tiredness`.

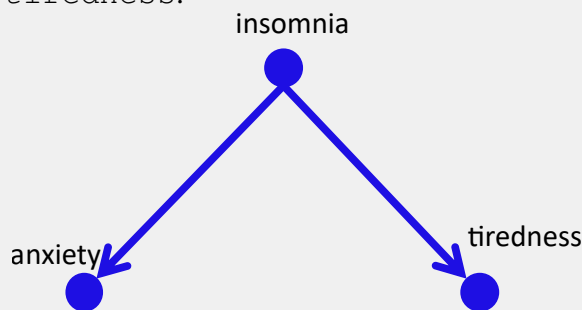


Figure 1.1: Causal model of insomnia, anxiety, and tiredness.

$$\begin{aligned}
 \text{anxiety} &= b_{\text{anx_ins}} \cdot \text{insomnia} + a_{\text{anx}} \\
 \text{tiredness} &= b_{\text{tir_ins}} \cdot \text{insomnia} + b_{\text{tir_anx}} \cdot \text{anxiety} + a_{\text{tir}}.
 \end{aligned}
 \tag{1.2}$$

1.2.1.2 Modeling Uncertainty

An advantage of probabilistic models is that they account for uncertainty. Uncertainty encompasses phenomena in a system or process from the natural processes of their occurrence or evo-

lution (e.g., is the weather going to get colder?) to the measuring and modeling of them (e.g., is there noise in the measurement of temperature? is the linear regression model a proper model for predicting the temperature?). In existing literature uncertainty is mainly categorised as *aleatoric* or *epistemic* [Alleman, 2013; Gelman, 2022; Spiegelhalter and Riesch, 2011].

Aleatory uncertainty is part of the naturally occurring stochastic processes [Alleman, 2013]. For example, the uncertainty about the outcome of a coin flip (heads or tails) is aleatoric. Epistemic uncertainty is the result of lack of knowledge deriving from various sources like the incomplete knowledge or understanding of the underlying processes of the phenomena, or the imprecise evaluation of their related characteristics (e.g., noisy measurements) [Alleman, 2013]. While aleatoric uncertainty is irreducible because it refers to the inherent uncertainty of random processes, epistemic uncertainty can be reduced if additional information about the underlying processes (e.g., more measurements) becomes available.

Both types of uncertainty can be modelled by a probabilistic model. The aleatoric uncertainty expressed as “naturally occurring variances” in the observations of a system [Alleman, 2013] is modelled in a probabilistic model “by placing a distribution over the output of the model” [Kendall and Gal, 2017]. The output of a probabilistic model is represented by a *random variable*. These types of variables do not take a fixed value like the variables in non-probabilistic models, but may take any value within a range with some *probability*.

The definition of the term “probability” has a long story, which will be narrated in Chapter 2 (Section 2.2.1). For the moment the *Bayesian* definition of it is given: probability is a measure of belief that a possible event (or outcome) will happen in a scale from 0 (no belief) to 1 (maximum belief) [Lambert, 2018a, Chap. 2; Martin, 2018, Chap. 1]. A random variable follows a *probability distribution* which assigns a probability to each possible value of the variable so that the sum of all probabilities or the area under the distribution is 1 depending on whether the random variable is discrete or continuous, respectively.

In Box 1.3 the linear regression model described in equation 1.1 (that was a non-probabilistic model) is rewritten in a probabilistic way to account for aleatoric uncertainty. The following paragraph describes how probabilistic models can account for epistemic uncertainty.

Box 1.3 Probabilistic linear regression model

In a probabilistic approach, the reaction time in the drivers' reaction time problem could be viewed as a random variable distributed according to a normal distribution. The mean (μ) of this normal distribution would be provided by the linear predictor $\mu = b \cdot n + a$, with some standard deviation (σ) such as $\sigma = s$. Equation 1.1 can be written in a probabilistic way as following:

$$\begin{aligned} \text{rt} &\sim \text{Normal}(\mu, \sigma) \Rightarrow \\ \text{rt} &\sim \text{Normal}(b \cdot n + a, s). \end{aligned} \tag{1.3}$$

Probabilistic statement 1.3 quantifies aleatoric uncertainty. For any 4-tuple of values (a, b, s, n) , a normal distribution of drivers' reaction time (rt) is defined. This normal distribution quantifies the aleatoric uncertainty.

1.2.1.3 Incorporating Prior Knowledge

Placing a prior distribution over a model's parameters, and then capturing how much these parameters vary given some data accounts for epistemic uncertainty [Kendall and Gal, 2017]. These *prior* distributions are the probability distributions of the model's parameters. They quantify any prior (incomplete) knowledge of the modeller or expert about the underlying data generating process of the modelled system *before observing any data*. Box 1.4 shows how priors could be set for the parameters of the drivers' reaction time problem.

Box 1.4 Definition of priors for the drivers' reaction time problem

An expected reaction time of a driver under normal conditions might be known to be around 100 ms with a standard deviation of 100 ms. This knowledge could arise from previous studies or experience. This forms a prior belief in the form of a distribution for the intercept a . The intercept equals the reaction time of a driver on the first driving day; $rt = b \cdot 0 + a = a$. Based on this information, the prior distribution of a takes the form:

$$a \sim \text{Normal}(100, 100). \quad (1.4)$$

The more days a driver drives under sleep deprivation conditions, the more tired he is expected to get. Hence, the slope of the regression model is expected to be positive. The bigger the slope of the regression line is, The more tired (bigger reaction times) the drivers get after consecutive days of driving, the bigger the slope of the regression line becomes. It is not known though how sharp this slope could be. For this reason, a wide prior distribution is set for the slope of the regression line: a normal distribution with mean value 10 and a standard deviation 10. This forms the prior distribution of parameter b :

$$b \sim \text{Normal}(10, 10). \quad (1.5)$$

The standard deviation of the drivers' reaction time cannot be judged a priori, but it is known to be a positive number. Thus, a wide prior distribution is set for parameter s such that it is defined over positive values:

$$s \sim \text{HalfNormal}(50). \quad (1.6)$$

The prior distributions express how much or little is known about the system's underlying data generating process before observing any data. A prior is *more informative* (i.e., the distribution is tighter) when there is more known and hence, the epistemic uncertainty is less. Otherwise, it is *less informative*.

1.2.1.4 Estimating Uncertainty through Bayesian Inference

The epistemic uncertainty can be reduced in the presence of observations especially in the case of less informative prior distributions. *Bayesian inference* is the mechanism that describes how this reduction of uncertainty can take place in the presence of observations in probabilistic modeling. In Bayesian inference, the *prior beliefs* about the value of the model's parameters get updated in the light of the observations and are turned into *posterior beliefs*. The update mechanism of prior beliefs in the light of observations is described mathematically by the Bayes' rule as it will be explained in detail in Section 2.2.

Box 1.5 explains why the observations can cause an update in the prior beliefs.

Box 1.5 Uncertainty reduction in the drivers' reaction time problem

Let us assume that a dataset of drivers' observed reaction times is available in the database of the logistics company. The data of one of the drivers is shown in Table 1.1.

Driver 330's reaction time on day 0 is 321 ms. Observing this data-point, the model might want to change the prior belief about the regression line's intercept (initially centered around 100 ms in statement 1.4) and shift it to higher values. This mathematically is interpreted as an assignment of more probability density to higher values of the parameter a in the posterior distribution.

Table 1.1: The reaction times of a driver after n consecutive days of driving under sleep-deprivation conditions. The dataset is from [Belenky et al., 2003] and can be retrieved from [Lambert, 2018b] for all 18 available drivers.

Driver ID	Days of Driving (n)	Reaction Time (τ)
330	0	321
330	1	300
330	2	283
330	3	285
330	4	285
330	5	297
330	6	280
330	7	318
330	8	305
330	9	354
330	6	280
330	7	318
330	8	305
330	9	354

The observed data of a system can be used to estimate the values of the parameters and consequently to generate predictions. The approach followed in the case of non-probabilistic modeling is to apply some algorithm to fit the parameters of a model to observations. For example, Ordinary Least Squares (OLS) is a commonly used algorithm for fitting linear regression models to observations. It fits the parameters of the regression model to the data by minimizing the sum of square differences between the observed and predicted values of the dependent variable. Such modeling approaches would generate fixed-value estimates for the parameters of the model. The predicted data would be fixed-valued given the value of the model's parameters.

The parameters of a probabilistic model are random variables defined by a probability distribution. The outcome of a probabilistic model is also represented by a random variable. A

probabilistic model produces predictions for the outcome with a confidence (or uncertainty) attached to them in contrast to non-probabilistic models that produce a prediction with certainty. The probability distributions of the parameters and observed variables in a probabilistic model can be inferred by Bayesian inference through the observations. Starting from a prior distribution that gets updated in the light of observation, a *posterior* distribution can be inferred through Bayesian inference, with the epistemic uncertainty being reduced as the observations accumulate.

Box 1.6 demonstrates how the outputs of a non-probabilistic and a probabilistic model differ within the context of the drivers' reaction time problem. The Python code for the modeling of the drivers' reaction time problem in this and the following chapter can be found in Taka [2023a]. Fig. 1.2 presents the outputs of both models. The parameters' estimates and the predicted reaction time take fixed values in the case of the non-probabilistic model, while they can take any value from a range of possible values in the case of the probabilistic model according to the estimated posterior distribution. The figure presents also the prior distributions of the parameters. The uncertainty included in the prior distributions is reduced in the posterior distributions in the light of data.

Modeling uncertainty is a valuable modeling attribute to achieve realistic predictions. For example, both the non-probabilistic and probabilistic models in Box 1.6 can predict the drivers' reaction time after n consecutive days of driving under sleep-deprivation conditions. Based on the non-probabilistic model all drivers would have the exact same reaction time if they all drive 3 consecutive days under sleep-deprivation conditions. These types of events are unlikely based on the human experience, while they are certain events based on this model. The probabilistic model estimates the distribution of the predicted reaction time. Based on this estimate the predicted reaction time will not be the exact same value for all drivers for any specific number of days driving under sleep-deprivation conditions.

Box 1.6 Estimation of parameters' value and generation of predictions in the drivers' reaction time problem

For the non-probabilistic linear regression model of equation 1.1, OLS is used to fit the parameters to the available observed data (found in [Lambert, 2018b]). The values of the parameters a and b are estimated; $a = 251.41$ and $b = 10.47$. The estimated values of the parameters can then be used in equation 1.1 to predict the reaction time of a driver after 3 consecutive days of driving, $n = 3$; $rt = 10.47 \cdot 3 + 251.41 = 282.82$ ms.

For the probabilistic linear regression model of statement 1.3, the priors given in statements 1.4-1.6, and the available observed data are used to conduct Bayesian inference and infer the posterior distributions of the parameters a , b , and s and the *posterior predictive* distribution of rt .

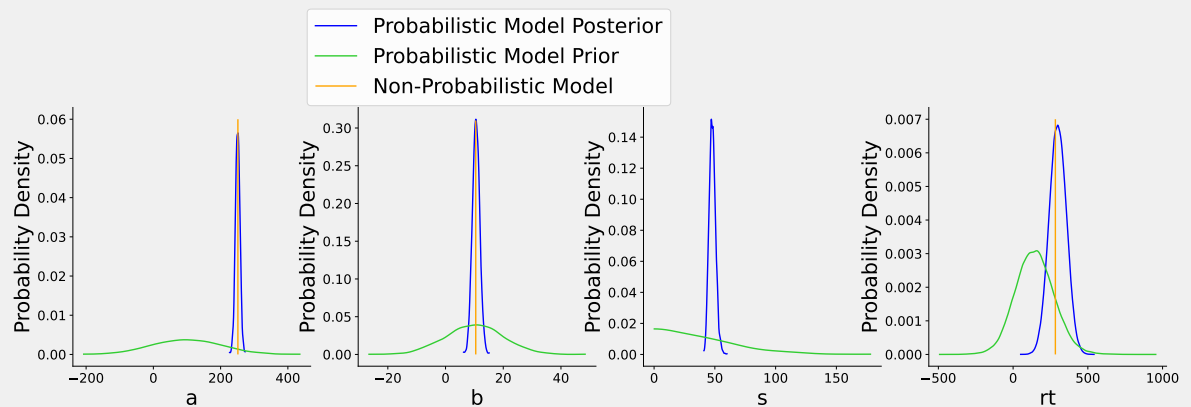


Figure 1.2: Parameter estimates and reaction times' predictions from the probabilistic and non-probabilistic model of the drivers' reaction time problem.

1.2.1.5 Accounting for Uncertainty when Modeling Causal Relations

As it was discussed in Section 1.2.1.1, the relation of two observed variables in a model is not necessarily truly causal, unless the model is a causal model or is designed under a causal model. For example, the observed variable days of driving n used as a covariate in the non-probabilistic and probabilistic regression models to predict the reaction time variable rt in Box 1.1 and 1.3, respectively, is not necessarily a cause of drivers' reaction time. On the other hand, the *insomnia* variable used as a covariate in the non-probabilistic regression models in Box 1.2 is a cause of the *anxiety* and *tiredness* variables. This is because the *insomnia-anxiety-tiredness* model in Box 1.2 was designed under a causal model represented by the causal diagram shown in Fig. 1.1.

Probabilistic models designed under a causal model can model the causal relations of observed variables in a probabilistic way. Box 1.7 demonstrates how the *insomnia-anxiety-tiredness* model shown in Box 1.2 is modelled probabilistically through probabilistic linear regressions. Appropriate priors could be set for the b , a , and σ parameters and Bayesian inference could be used to infer the posterior distributions of the parameters and the posterior predictive distribu-

tions of the observed variables.

Box 1.7 Modelling causal relations probabilistically

Fig. 1.3 replicates Fig. 1.1 to show again the causal diagram of the insomnia-anxiety-tiredness causal model here. The linear regression models 1.2 shown in Box 1.2 are rewritten below in a probabilistic way to account for uncertainty.

The `insomnia`, `anxiety`, and `tiredness` are assumed to be continuous random variables and linearly related. `anxiety` and `tiredness` are assumed to follow a normal distribution with mean $\mu_{\text{anx}} = b_{\text{anx_ins}} \cdot \text{insomnia} + a_{\text{anx}}$ and $\mu_{\text{tir}} = b_{\text{tir_ins}} \cdot \text{insomnia} + b_{\text{tir_anx}} \cdot \text{anxiety} + a_{\text{tir}}$ and standard deviation σ_{anx} and σ_{tir} , respectively.

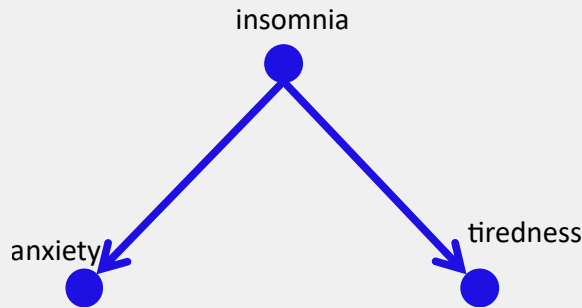


Figure 1.3: Causal model of insomnia, anxiety, and tiredness.

$$\begin{aligned}
 \text{anxiety} &\sim \text{Normal}(b_{\text{anx_ins}} \cdot \text{insomnia} + a_{\text{anx}}, \sigma_{\text{anx}}) \\
 \text{tiredness} &\sim \text{Normal}(b_{\text{tir_ins}} \cdot \text{insomnia} + b_{\text{tir_anx}} \cdot \text{anxiety} + a_{\text{tir}}, \sigma_{\text{tir}}).
 \end{aligned}
 \tag{1.7}$$

The regression coefficients b (the parameters that multiply the independent variables) in a linear regression model describe the correlation between the independent and dependent variables. When the independent variables are causes of the dependent ones, the regression coefficients b describe the *causal effect* the independent variables have on the dependent ones [Bulbulia et al., 2021; McElreath, 2021; Textor and Gilthorpe]; that is a measure of how much a variable-effect would change by a change in its variable-cause.

Using probabilistic models to model causal relations of variables can account for the uncertainty about the value of their causal effects and the predicted data. Estimations of the uncertainty about the size of causal effects or the values of predicted data could be valuable information in many settings of crucial decision-making. A simplistic example could be a doctor, who needs to decide which treatment to give to a patient. The effect of the possible treatments on the disease are not expected to be the same for all patients but to present variability. A more informative decision could be made if the uncertainty about the size of each treatment's effect on the disease is known.

1.2.2 Obstacles

The roots of Bayesian inference go back to the work presented by Thomas Bayes in 1763 [Bayes and Price, 1763] but Bayesian inference had remained in obscurity for a very long time since then and was mainly used within academia. A reason for this was the computational challenges in applying Bayes' rule for the computation of the posterior distribution when the complexity of the probabilistic models was increased. As the number of the parameters in the model increases, some components of Bayes' rule become intractable. This will be explained through concrete examples in Section 2.2.3. Alternative estimation techniques could be used for the calculation of the posterior in certain cases but these required specialized statistical knowledge.

The foundations of a wider adoption of Bayesian statistics were laid in early 1990's, when indirect estimation methods of the posterior were introduced; Markov Chain Monte Carlo (MCMC) algorithms [Spiegelhalter and Rice, 2009] and a decade later the Probabilistic Programming Languages (PPLs) [Poole and Wood, 2022]. The creators of PPLs started working on integrating semantically the inference in a programming language to hide the mathematical details and automate the inference process. They also incorporated powerful MCMC algorithms for the estimation of the posterior [Poole and Wood, 2022].

Bayesian probabilistic modeling has many advantages; it accounts for uncertainty systematically; it allows precise incorporation of prior expert knowledge; and the intrinsic structure of models is well-defined in terms of relations among random variables: the mathematical and statistical dependencies are explicitly stated. Bayesian probabilistic models can be implemented via PPLs, which provide automatic inference via efficient MCMC sampling algorithms. Nevertheless, probabilistic models and Bayesian inference are still not widely adopted.

From the perspective of an analyst, building a probabilistic model still requires some statistical knowledge. For example, the specification of the probability distributions for the priors can be a difficult task especially if the analyst has poor background in statistics [Phelan et al., 2019; Sarma and Kay, 2020]. Many analysts might avoid probabilistic modeling and Bayesian inference because they are not able to specify appropriate priors. Validating how well an analyst's prior knowledge aligns with the probability distribution of a prior or refining the model to more accurately capture the "relevant aspects of the system" are tasks that cannot be easily done by novices. Comprehension of the statistical context of the model is required for the validation or refinement of probabilistic models.

From the perspective of a user of a probabilistic model, parameter tuning and decision-making under uncertainty are tasks that require comprehension of model's structure and insight into the uncertainty of parameters' value and predicted data. Users need to know how variables in a model are related, namely which variable affects the other and how, to tune a parameter appropriately, or make a decision for an intervention on a variable that might affect other variables in the model. Insight into variables' uncertainty could help decision-makers make better assessment of the risk over all possible outcomes and thus, make more informed decisions.

Textual representations of probabilistic models like probabilistic statements or PPLs hide most of the mathematical details of probabilistic models but still require some statistical knowledge to infer variables' relations. The complexity of the model plays also an important role in the ability of users' to comprehend model's structure and uncertainty. A very simple probabilistic model with few parameters could allow a user to contemplate the entire model at once and comprehend how parameters interact with each other and the predictions of the model. This becomes challenging as the model becomes more complex, perhaps with multivariate distributions, complex inter-dependencies and increasingly abstract latent states.

Communication of uncertainty has challenges, too. The uncertainty in a probabilistic model can be complex and multidimensional (one dimension per parameter in the model) and is depicted in two types of distributions, the prior and the inferred posterior. Communicating uncertainty and especially a complex distribution may confuse people if the design of its representation does not account for the needs of a specific application [Fernandes et al., 2018; Greis et al., 2017; Kay et al., 2016a] or for the ways that people naturally reason about probability [Belia et al., 2005; Cosmides and Tooby, 1996; Gigerenzer and Hoffrage, 1995; Joslyn and LeClerc, 2013].

1.3 Communicating Probabilistic Models through Visualization

This section presents the current practices in using visualization to communicate probabilistic models given the available tools (Section 1.3.1) and discusses the challenges in visualizing probabilistic models (Section 1.3.2).

1.3.1 How is Visualization Used to Communicate Probabilistic Models?

A probabilistic model consists of the following components:

- a *structure*; this reflects the relations (i.e., the statistical and mathematical associations) of the variables in the model;
- *inference results*; these consist of the prior and posterior distributions of the parameters and the (prior and posterior) predictive distributions of the observed variables.

Both these aspects of a probabilistic model need to be communicated to users to help them draw a complete picture of the model. In this way users would be able to perform tasks like checking the validity of the model, refining it, using it to make decisions, or tuning its parameters. The following subsections discuss the existing practices in representing the structure and the inference results of probabilistic models visually.

1.3.1.1 Visualizing the Structure of Probabilistic Models

Graphs are a common way of representing a probabilistic model’s structure visually. The *nodes* correspond to model’s variables. The *edges* are directed arrows from one variable to another indicating the direction of their dependency based on the definition of the model; the arrow starts from the independent variable and points to the dependent variable. An informationally minimal graph is the Bayesian network [Koller and Friedman, 2009] (Fig. 1.4(a)). More informative versions of graphs are provided by the graphical tools of some PPLs. For example, in the DoodleBUGs’ graph [Spiegelhalter et al., 2003], nodes contain information about variables’ dimensions (Fig. 1.4(b)). In PyMC’s graphs [Ellson et al., 2004], nodes also contain the name of the prototype distribution of the variables (Fig. 1.4(c)). The Kruschke-style diagram [Kruschke, 2015, Chap. 8] (Fig. 1.4(d)) elaborates the graph with the iconic “prototypes” of the variables’ distribution on each node and annotations for the parameters of distributions being set by other parameters in the model.

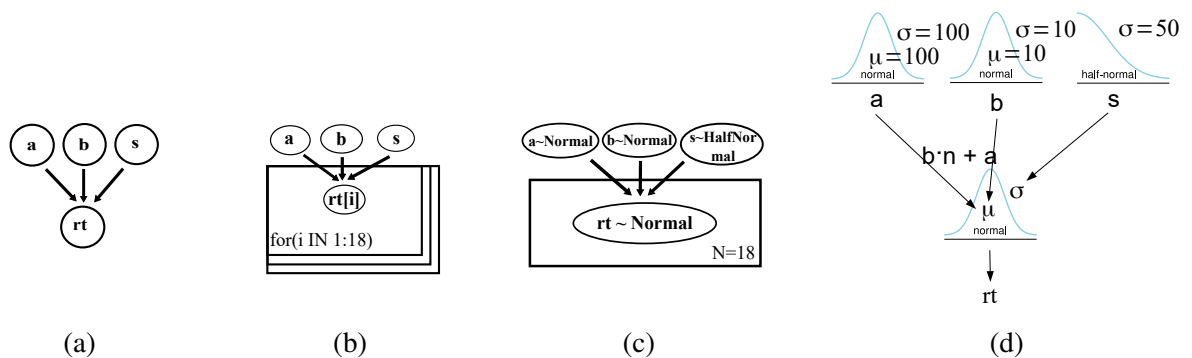


Figure 1.4: Various graphical representations of the probabilistic model of the drivers’ reaction time problem (Box 1.3 and 1.4). (a) Bayesian network, (b) DoodleBUGs’ graph, (c) PyMC’s graph, and (d) Kruschke-style diagram.

Diagrammatic representations of probabilistic models like these ones could provide a “comprehensive overview of the relations between parameters and their meanings with respect to each other and to the data” [Kruschke, 2018a]. Sketching out a diagram of the probabilistic model could facilitate tasks like model validation or specification in PPL code [Kruschke, 2018a].

1.3.1.2 Visualizing the Inference Results

Although the diagrammatic representations of probabilistic models are useful in communicating the structure of the models, Bayesian modeling needs more than this. For example, Gabry et al. [2019] highlight the importance of visualization in all the stages of a Bayesian workflow that comprise of an iterative process of model building, inference, model checking and evaluation, and model expansion.

The most common practice in reporting the inference results of Bayesian analysis is tables that present summary statistics of the posterior distributions. For example, Table 1.2 presents the

summary statistics of the posterior distributions of the parameters in the drivers' reaction time probabilistic model (Box 1.3 and 1.4). The mean value, standard deviation, and Highest Density Interval (HDI) of the posterior distributions are reported in this table. The HDI (sometimes encountered as Highest-Posterior Density (HPD) in the literature when referred to the posterior distribution) is a measure of spread of the distribution and designates the shortest interval containing a given portion of the probability density (e.g., 94%).

Table 1.2: Posterior statistics in a tabular format for the drivers' reaction time probabilistic model. The 97% HDI is included in the statistics: the left end of the interval shown in column HDI_1.5% is represented by the value below which 1.5% of the posterior falls and the right end of the interval shown in column HDI_98.5% is represented by the value above which 1.5% of the posterior falls.

	mean	std	HDI_1.5%	HDI_98.5%
a	251	6.6	237	265
b	11	1.2	8	13
s	48	2.6	43	54

The inference results of a probabilistic model could be represented visually through uncertainty visualizations. Depending on the type of uncertainty visualization used (e.g., error bar, Box plot, Kernel Density Estimate (KDE) plot), this way of communicating the inference results could be more or less informative in comparison to the tables of summary statistics. The use of visual means like color, transparency, gradient, etc., could make the uncertainty in the inference results of a probabilistic model more intuitively comprehensible.

A common uncertainty visualization used for the communication of inference results is the Kernel Density Estimate (KDE) plot. This can be used to communicate visually a smoothed summary of a parameter's prior or posterior distribution. The width of the KDE plot indicates the range of possible values that the variable can take and its height at each value point how probable that value is.

Fig. 1.5 presents the Kruschke-style [Kruschke, 2015] posterior distributions of the parameters in the drivers' reaction time probabilistic model (Box 1.3 and 1.4) in the form of KDE plots. The Kruschke-style representation of the posterior distributions is a quite informative visual representation as it provides information about the exact statistics of the distributions (the mean value and HDI) in the form of annotations.

There are various visualization tools that can represent the inference results visually. Many of them are suitable for Bayesian analysis as they can accept the outputs of various PPLs. The most common are ArviZ [Kumar et al., 2019] in Python, and bayesplot [Gabry and Mahr, 2020], tidybayes [Kay, 2020], and shinystan [Stan Development Team, 2017] in R. The ArviZ API was

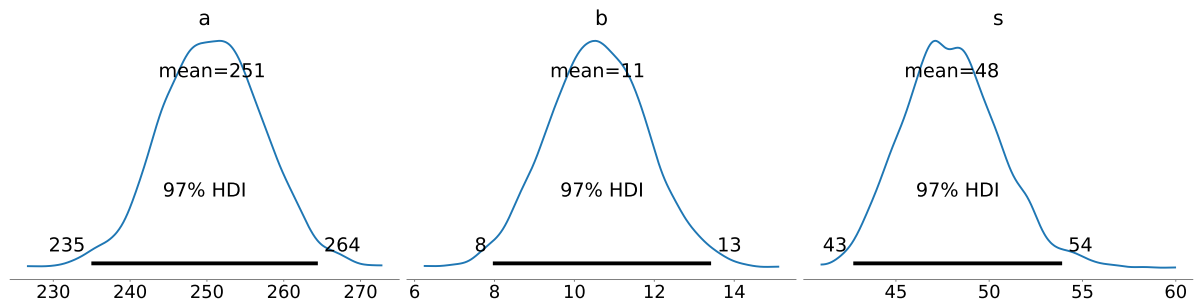


Figure 1.5: Kruschke-style parameters' posterior distributions of the probabilistic model of the drivers' reaction time problem.

used to generate Table 1.2 (`arviz.summary`) and the Kruschke-style KDE plots in Fig. 1.5 (`arviz.plot_posterior`).

1.3.2 Challenges in Visualizing Probabilistic Models

Although the means for the communication of probabilistic models seem to exist, very little is known about how effective they can be in improving users' comprehension of and trust in probabilistic models, and helping them in tasks. For example, diagrammatic representations of probabilistic models like graphs can help users view the relations among variables in a model at a glance (through the existence or absence of edges). In the case of the more informed graphs like Kruschke-style diagrams, users could even observe the exact statistical associations or mathematical equations. But inferring what exactly the effect of one variable on another is, is still very much dependent on the ability of the users to understand the mathematical details.

Similarly, interpreting representations of inference results in the form of tables of summary statistics or KDE plots depends on users' level of statistical knowledge. The numerical data presented are usually statistics like mean, standard deviation or confidence intervals, which could mislead or overwhelm unfamiliar users. A static representation of the inference data in summary tables or uncertainty visualizations could not communicate the sensitivity of the parameters, which would allow a decision-maker to assess the impact of parameter inter-dependencies and associated risks.

The efficiency of the existing communication means of probabilistic models can also be questioned especially in cases of complex models. The tables of summary statistics or the uncertainty visualizations of the parameters' posterior distributions can become unwieldy when the number of the parameters in the model increases. The limited capacity of human cognition could be a hurdle for users to grasp the uncertainty presented and make required assessments of the risk. Communicating the prior would imply communicating a second table of similar complexity or as many KDE plots. Thus, the priors or many of the posteriors might be omitted in reports of Bayesian analysis leaving the reader with an incomplete picture of the analysis and findings.

1.4 This Work

1.4.1 Thesis Statement, Aims and Scope

The hypothesis investigated in this research is that interactive animated visualizations of Bayesian probabilistic models and their outputs aid end-users in better understanding probabilistic models and causality in data. Specifically, it is investigated whether adding interaction and animation to visual representations of a probabilistic model helps people better understand the structure of the model and interpret the (causal and non-causal) relations of the variables within the context of the model. This is investigated through user studies.

To that end, novel and automated tools for visualizing Bayesian probabilistic models and their outputs are proposed and created in this research. The proposed visualizations incorporate structural information about the model and its sample-based (MCMC) inference and integrate interaction and animation. The communication and exploration of variables' prior and posterior distributions and relations is facilitated through the proposed visualizations. The ultimate aim is the proposed visualization tools to help make Bayesian analysis and its outputs more intuitive and interpretable for non-experts.

The created visualization tools are offered to the research community in the form of packages or libraries. Also anything related to the user studies conducted within the context of this research (collected data, analysis code, training material) is published along with the findings to encourage replication or meta-analyses.

The targets of this research are summarized by the points below:

1. The creation of novel and automated tools for visualizing Bayesian probabilistic models to:
 - (a) communicate variables' distributions;
 - (b) explore the prior and posterior MCMC sample space of the inference;
 - (c) visualize variables' (causal and non-causal) relations (dependencies) in a probabilistic model.
2. The investigation of whether animation and interaction improve the understanding of a model's structure and its variables' dependencies.
3. The investigation of whether animation and interaction improve the understanding of variables' causal relations in probabilistic models that model a causal structure.

1.4.2 Contributions

The research presented in this thesis consists of three parts which all focus on the visualization of probabilistic models and their outputs but from a different perspective.

The *first part* (Chapter 4) investigates two things; first, how an automated visualizer of probabilistic models that are expressed in any PPL and their outputs can be designed, and second how interaction could be incorporated to this design to make sample-based (MCMC) inference results more intuitive and easily explorable. A concrete implementation of such a tool is presented.

The *second part* (Chapter 5) investigates whether interactive conditioning when used to explore samples from the prior distributions of probabilistic models' variables helps users better understand the structure of the probabilistic model, namely the relations among the variables' of the model. The design of a concrete visualization presenting these distributions and integrating interactive conditioning is presented. The results from the analysis of data collected through a user study using the suggested interactive visualization as a visualization instance are presented and discussed.

The *third part* (Chapter 6) investigates two things; first, how probabilistic models and existing PPLs can be used to simulate causal models and *interventions* on them, namely to generate data with uncertainty from specific causal structures before or after an external change imposed on variables of the model. The second thing that is investigated is whether interaction or animation when used in visualizing simulated data from these models and interventions applied on them helps users better understand their structure or make decisions about the design of actual interventional experiments. The conduction of actual interventional experiments could help users (e.g., researchers, doctors, psychologists) infer the causal relations of observed variables of interest. The design of concrete visualizations presenting the simulated data of causal models and integrating interaction and animation are presented. The results from the analysis of data collected through a user study using the suggested visualizations as visualization instances are presented and discussed.

1.4.3 Research Challenges

The research on the visualization of probabilistic models presented in this thesis entails many challenges from a research, technical and dissemination perspective. The identified challenges are discussed in this subsection to give the reader an initial idea of the breadth and depth that this research could reach.

The challenges from the research perspective are mainly two-fold; how to design the visual representations and how to investigate their effectiveness. The communication of uncertainty in complicated multi-parameter models should employ visualization designs that account for people's ability to interpret uncertainty and in fact doing this in many dimensions, one per parameter. The incorporation of interaction and animation in these designs with the aim to make model's structure and output more intuitive rather than confusing for the naïve users constitutes the main research challenge in this thesis. The design of appropriate evaluation protocols for the user studies is another challenge. Decisions like whether participants should get trained and

how, how the questions and training could be simplified to enable participants of any statistical background to take part in the user studies, what modeling choices (e.g., hierarchical modeling, sophisticated parameterizations for setting the parameters of variables' distribution etc.) to be investigated to ensure a broad enough range is included should be made.

The challenges from the technical perspective are of lower importance but are there for creating limitations to how far the designs of the suggested visualizations could go. The use of animation and interaction imply the need for a *real-time* retrieval, presentation, and update of information in the visualizations which sets many technical challenges in terms of implementation especially as the models complexity increases and the visualizations need to scale up. The capabilities that existing development tools (e.g., existing visualization libraries for interaction) could offer sometimes put restrictions on the flexibility that the visualization designs could have. The creation of shareable code in the form of packages or libraries sets also certain technical specifications that should be met.

The challenges from the dissemination perspective are great. The aim of this research is to create tools that will be useful and helpful not only for the experts or already users of Bayesian analysis, but also for the less familiar audience. The communication of this research to a broader audience with possibly limited background in statistics is challenging both within the context of the evaluation user studies and generally for purposes of research dissemination. The definition of a set of appropriate models and applications for explaining and demonstrating the impact of this research is required. The context of the applications should be comprehensible and interesting to the broader audience and enable them to understand the benefits they can get by utilizing visualization means in Bayesian analysis. The technical and mathematical details should be simplified.

1.4.4 Thesis Outline

Chapter 2 explains the theoretical background to the work by expanding on the themes presented above. The technicalities of Bayesian inference and PPLs that are relevant to the context of the thesis are explained. Why modeling causal relations requires extra methodologies is explained. The relevant theories of visual perception are also presented in this chapter.

Chapter 3 presents the existing work in the field of uncertainty visualization; the existing approaches, the challenges in communicating uncertainty, the evaluation of uncertainty visualization, the uncertainty visualization in Bayesian reasoning, and the decision-making under uncertainty.

Chapter 4 presents the first part of the research presented in this thesis, which focuses on the design and creation of an automated interactive visualizer of probabilistic models that are expressed in any PPL and use sample-based (MCMC) inference.

Chapter 5 presents the second part of the research presented in this thesis, which focuses on the investigation of the role of interactive conditioning in the comprehension of variables'

relations in a probabilistic model.

Chapter 6 presents the third part of the research presented in this thesis, which focuses on the simulation of causal models and interventions by using probabilistic modeling and PPLs, the design of appropriate visualizations for representing the simulated data, and the evaluation of these designs.

Chapter 7 discusses the conclusions drawn from the thesis, and the dimensions that this work could take in the future.

Chapter 2

Theory

2.1 Summary

This chapter expands on the theoretical background of key themes encountered in this thesis. Bayesian inference is the main theoretical concept that readers need to know to follow the work presented in this thesis. All visualizations proposed and evaluated in this work present the outputs of sample-based (MCMC) Bayesian inference. Section 2.2 presents useful concepts from probability theory and explains Bayes' rule. The difficulties with the computation of Bayes' rule and the alternative methods to conduct Bayesian inference are discussed. The algorithm to generate predictions from the inferred distributions is presented. An example of outputs of Bayesian inference is provided. A discussion about why to use Bayesian statistics to analyse data is also provided.

PPLs are automated tools enabling the definition of probabilistic models and the conduction of Bayesian inference. A part of this research focuses on the automatic transformation of PPLs' outputs into interactive visualizations. For this reason, Section 2.3 discusses the purpose that PPLs want to serve, how a probabilistic model is expressed in PPL code and the advantages of doing this, and finally, how this code leads to arrays of samples within the context of a PPL.

Probabilistic models can be used to model variables' causal relations under a causal model (i.e., a model that explains which variable causes which other variables) as it was explained in Chapter 1. A systematic way to do this and to simulate interventions using PPLs is presented in Chapter 6. But before presenting this work in Chapter 6, it is important for the reader to understand why extra information, like that provided by causal models, is required to model causal relations of variables. Section 2.4 explains the reasons through illustrative examples and presents some well-known causal modeling methodologies.

Finally, there are various theories or empirical knowledge of the role that human visual perception plays in the effectiveness of visualization. This is a very relevant topic to the subject of this thesis as it sets the theoretical background of this research: why to use visualization to communicate probabilistic models? why to expect them to be effective? how should they be

designed? Section 2.5 introduces the reader to the relevant theories and empirical knowledge in human visual perception to provide the context in which this research is developed and for which it is purposed. It explains why it is important to understand how people perceive information visually. It discusses the most relevant theories of visual perception of uncertainty, animation, and interaction that constitute main research focuses in this thesis.

2.2 Bayesian Inference

2.2.1 Schools of Statistical Inference and Definition of Probability

A definition of *statistical inference* is given by Johnson et al. [2012]: “Statistical inference is the process through which inferences about a population are made based on certain statistics calculated from a sample of data drawn from that population.” The notion of “probability” plays a leading role in a statistical inference process. Specifically, the aim of statistical inference is the computation of the probability of a hypothesis about the population given the observed sample data, $\Pr(\text{hypothesis}|\text{data})$ with \Pr denoting the probability [Lambert, 2018a, Chap. 2.8].

The definition of probability varies depending on the school followed for the conduction of statistical inference. There are mainly two schools of statistical inference: the *frequentist* and the *Bayesian*. In frequentist statistics, the probability represents the frequency of an event occurring in an infinite number of repetitions of an experiment [Lambert, 2018a, Chap. 2.5]. In Bayesian statistics, probability is a measure of certainty about subjective beliefs, which can be updated in the light of data [Lambert, 2018a, Chap. 2.6]. Triggered by the different approaches in defining the notion of “probability” and for reasons of reference later in the thesis, a short discussion about the main differences between the two schools of statistical inference will be provided here, although the main focus of this work is on Bayesian inference.

The two schools of statistical inference differ in their approach of computing the probability $\Pr(\text{hypothesis}|\text{data})$. Given some hypothesis about the population the frequentist statistics assumes that this hypothesis is true, collects sample data from the population, and estimates the probability of this data occurring given the hypothesis is true, $\Pr(\text{data}|\text{hypothesis})$, and if this probability is very small (smaller than an arbitrary threshold), the hypothesis is rejected: $\Pr(\text{hypothesis}|\text{data}) = 0$. Frequentist statistics focuses on estimating the probability of obtaining the sample data given the hypothesis is true to infer the probability of the hypothesis is true given the sample data through a hypothesis testing. Estimating the probability of obtaining the sample data given the hypothesis is true, $\Pr(\text{data}|\text{hypothesis})$, requires that the data is considered random and the parameters of the model fixed to reflect the hypothesis [Lambert, 2018a, Chap. 2.5].

Bayesian statistics can directly estimate the probability of the hypothesis is true given the

sample data, $\Pr(\text{hypothesis}|\text{data})$. This is possible through Bayes' rule, which is the mathematical tool to invert the probability of obtaining the sample data given the hypothesis is true, $\Pr(\text{data}|\text{hypothesis})$, and compute the intended probability of the hypothesis is true given the sample data, $\Pr(\text{hypothesis}|\text{data})$. Estimating the $\Pr(\text{hypothesis}|\text{data})$ requires that the parameters are considered to vary and the data is considered to be fixed [Lambert, 2018a, Chap. 2.6]. Section 2.2.3 presents Bayes' rule and how it can be used for the estimation of this probability.

2.2.2 Summary of Useful Probabilistic Concepts

This section gives a very brief summary of the probabilistic concepts from the probability theory that will be encountered in this thesis. More details on any of these concepts could be found in any relevant textbook [Papoulis and Pillai, 2002; Rotondi et al., 2022; Speegle and Clair, 2021].

A random variable x is *discrete* when it takes one value from a countable set of values, or *continuous* when it takes values from a continuous range of values. A discrete random variable follows a valid *probability distribution* when its probability mass function (PMF) $p(x)$ assigns a probability value from 0 to 1 to each possible value of the variable so that the sum of the probabilities of all possible values equals 1:

$$\sum_{x \in X} p(x) = \sum_{x \in X} \Pr(x = X_i) = 1, \quad (2.1)$$

with $X = \{X_i\}$ and $i \in \{0, 1, \dots, n\}$. A continuous random variable follows a valid probability distribution when its probability density function (PDF) $p(x)$ assigns a probability density to each possible value of the variable so that the integral of the probability densities over the range of the variable equals 1,

$$\int_x p(x) dx = 1. \quad (2.2)$$

Well-known discrete probability distributions are the Poisson, binomial, Bernoulli distributions, and continuous are the normal, uniform, Cauchy distributions.

Given a set of random variables x_1, \dots, x_n , the joint probability distribution of these variables can be defined as $p(x_1, \dots, x_n)$. If the random variables are *independent* (none of the variables help to predict another), $x_1 \perp \dots \perp x_n$, the joint distribution of the variables is computed by the product of the distributions of the individual probabilities: $p(x_1, \dots, x_n) = \prod_i p(x_i)$. In case the random variables are not independent their joint probability distribution can be expressed in terms of the *conditional probabilities* of the variables based on the *chain rule*. For example, assuming $n = 2$ and $x_1 \not\perp x_2$: $p(x_1, x_2) = p(x_1|x_2)p(x_2) = p(x_2|x_1)p(x_1)$. For more than two variables, $p(x_1, \dots, x_n) = p(x_n|x_1, \dots, x_{n-1})p(x_1, \dots, x_{n-1})$.

The distribution of a random variable can be calculated by the joint distribution of this variable with other random variables by calculating the *marginal distribution* for this variable. For

example, the marginal distribution of x_n is $p(x_n) = \sum_{x_1} \dots \sum_{x_{n-1}} p(x_1, \dots, x_n)$ in the discrete case and $p(x_n) = \int_{x_1} \dots \int_{x_{n-1}} p(x_1, \dots, x_n) dx_{n-1} \dots dx_1$ in the continuous case.

2.2.3 Bayes' Rule

The aim of Bayesian inference is the computation of the probability of a hypothesis being true given some sample data, $\Pr(\text{hypothesis}|\text{data})$, as it was explained in Section 2.2.1. Let us denote as θ the vector with the parameters of the model that describe the hypothesis and as data the list of observations. In Bayesian inference the value of the parameters is considered to vary and thus, the parameters are represented by random variables. The probability distribution of the parameters given the list of the observations, $p(\theta|\text{data})$, can be computed with Bayes' rule according to the following formula:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}. \quad (2.3)$$

The $p(\theta)$ is a valid probability distribution called *prior distribution*. The prior distribution represents the analyst's prior belief about the parameters θ and expresses his pre-data uncertainty for the parameters' true value based on his knowledge, experience and expertise in the specific problem or context, or on potential previous analyses [Lambert, 2018a, Chap. 5.3]. The prior distribution $p(\theta)$ is a k -variate *joint distribution* with k denoting the number of parameters in the model.

The $p(\text{data}|\theta)$ is called *likelihood* and is a function of the parameters θ as the data is fixed [Lambert, 2018a, Chap. 4.4]. The likelihood models the distribution of the data for the various values of the parameters θ and it is not a valid probability distribution, as the value of the parameters varies. If the value of the parameters is fixed, a valid probability distribution is retrieved. This is often encountered in the literature as *sampling distribution*.

The $p(\theta|\text{data})$ is a valid probability distribution called *posterior distribution*. The posterior distribution represents the posterior belief about the parameters θ after updating the pre-data (prior) belief about them in the light of the data [Lambert, 2018a, Chap. 7.4]. The posterior distribution $p(\theta|\text{data})$ is a k -variate *joint distribution* with k denoting the number of parameters in the model like in the case of the prior distribution. The posterior distribution is the result of combining the likelihood and the prior and usually has less uncertainty than the prior because it includes the extra information from the observed data.

The *denominator* of Bayes' rule, $p(\text{data})$, is a normalising factor used to ensure the posterior is a valid probability distribution [Lambert, 2018a, Chap. 6.3]. The denominator of Bayes' rule is calculated by summing (for discrete variables) or integrating (for continuous variables) out all the parameter dependencies in the numerator, namely the product of the likelihood and priors. In the discrete case, that is:

$$p(\text{data}) = \sum_{\boldsymbol{\theta}} p(\text{data}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}} p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (2.4)$$

and in the continuous case, that's:

$$p(\text{data}) = \int_{\boldsymbol{\theta}} p(\text{data}, \boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad (2.5)$$

In the case of a single-parameter model where $\boldsymbol{\theta}$ consists of a single parameter, equations 2.4 and 2.5 consist of a single sum or integral, respectively, while in the case of multi-parameter models where $\boldsymbol{\theta}$ consists of a vector of parameters, the single sum or integral is replaced by a number of summations or integrals, respectively, one for each parameter in $\boldsymbol{\theta}$.

Having explained the components of Bayes' rule, it is clearer now what an analyst needs to do to compute the $\Pr(\text{hypothesis}|\text{data})$ using Bayesian inference. He needs to have a list of observed data, place a prior distribution over every parameter of the model, define a likelihood for the observed data, and finally, apply the Bayes' rule to estimate the posterior distribution $p(\boldsymbol{\theta}|\text{data})$. Having done so, he can then estimate the probability of the parameters to take the specific values of the hypothesis.

2.2.4 Difficulty with the Computation of Bayes' Rule

2.2.4.1 Explaining Complexity of Bayes' Rule Components

Bayes' rule is a valuable tool for conducting statistical inference. Nevertheless, its components can easily lead to high-dimensional distributions that are hard to calculate as the model increases in complexity, i.e., the number of parameters or the coordinates of the indexing dimensions of the parameters increases. Let us demonstrate in which ways the complexity of the components in Bayes' rule can be increased through a concrete example. Three different types of probabilistic models for the drivers' reaction time problem discussed in the previous chapter are considered; a *homogeneous*, *heterogeneous*, and *hierarchical* model.

A homogeneous model is the probabilistic model of the drivers' reaction time problem presented in Chapter 1 (Box 1.3 and 1.4). Box 2.1 presents again the definition of this model as it was presented in Boxes 1.3 and 1.4. This model uses a global set of parameters to model all drivers.

Box 2.1 Homogeneous probabilistic model of the drivers' reaction time problem

The homogeneous model assumes all drivers are modelled with a global set of parameters. The inferred posteriors of the parameters are common for all 18 drivers in the dataset.

$$a \sim \text{Normal}(100, 100) \quad (2.6)$$

$$b \sim \text{Normal}(10, 10) \quad (2.7)$$

$$s \sim \text{HalfNormal}(50) \quad (2.8)$$

$$\text{rt} \sim \text{Normal}(b \cdot n + a, s), \quad (2.9)$$

The whole set of the observations `data` is used for the inference of the parameters' posteriors.

In a homogeneous probabilistic model, data is pooled together for inference and a single posterior distribution is estimated for each parameter in the model. An inferred posterior in a homogeneous model is common for all subjects in the dataset. This model is good for modeling the uncertainty in the predictions. It also accounts for the common characteristics of the subjects in the dataset. For example, the homogeneous probabilistic model in Box 2.1 accounts for the common characteristics of the drivers in the dataset; all drivers are professional drivers that underwent similar training and are used to driving under sleep-deprivation conditions.

A homogeneous model does not account for the individual characteristics of the subjects. This could be achieved by conducting a separate inference for each subject and inferring a separate posterior distribution of the model's parameters based on the observed data of each subject. A *heterogeneous model* can be used in this case. This type of model indexes the parameters by numbers corresponding to each one of the subjects in the dataset. The indexing dimensions of the parameters in a heterogeneous model depends on the number of the subjects in the dataset. Box 2.2 presents the definition of the heterogeneous probabilistic model for the drivers' reaction time problem. The parameters in this model are indexed by numbers indicating a driver in the dataset. A heterogeneous model accounts for the individual characteristics of the subjects but does not account for their common characteristics as the homogeneous model does. For example, the heterogeneous probabilistic model in Box 2.2 accounts for the individual characteristics of the drivers; drivers might vary in skills, experience, and endurance.

Probabilistic modeling offers the option of *hierarchical modeling* [Gelman, 2006; Kruschke, 2012a]; a more realistic modeling approach that would account for both aspects of subjects' characteristics. A hierarchical probabilistic model captures the uncertainty of predictions for subjects overall taking into consideration the uncertainty of the individual subjects' behavior, at the same time.

Box 2.2 Heterogeneous probabilistic model of the drivers' reaction time problem

The heterogeneous model assumes that each driver is modelled with a separate set of parameters. The inferred posteriors of the parameters are unique for each one of the 18 drivers in the dataset.

$$a_i \sim \text{Normal}(100, 100) \quad (2.10)$$

$$b_i \sim \text{Normal}(10, 10) \quad (2.11)$$

$$s_i \sim \text{HalfNormal}(50) \quad (2.12)$$

$$rt_i \sim \text{Normal}(b_i \cdot n + a_i, s_i), \quad (2.13)$$

where $i \in \{1, 2, \dots, 18\}$ for each one of the 18 drivers included in the dataset of observations. Only the data_{a_i} subset of the observations corresponding to the i -th driver is used for the inference of the i -indexed parameters.

Hierarchical modeling enables inferences at both the individual and the whole population level at the same time. This is achieved by adding extra parameters in the model, the *hyperparameters*. Instead of fixing the parameters of the model's priors to constant numbers, they are estimated directly from the data by placing shared hyperpriors over them. Hyperpriors are inferred on the level of the population and thus, are common for all subjects. The priors are inferred on the level of the individual subjects like in heterogeneous modeling. The complexity of the model is further increased in hierarchical modeling by the addition of the extra (hyper)parameters. Box 2.3 presents the definition of the hierarchical probabilistic model for the drivers' reaction time problem.

The dimensionality of the distributions in Bayes' rule increases in line with the complexity of the probabilistic model. From $k = 3$ dimensions in the case of the homogeneous model of the drivers' reaction time problem, we move to $k = 3 \cdot 18 = 54$ dimensions in the case of the heterogeneous model, and $k = 3 \cdot 18 + 5 = 59$ dimensions in the case of the hierarchical model. This k determines the dimensionality of the vector $\boldsymbol{\theta}$ of the parameters used in Bayes' rule formula. The vector $\boldsymbol{\theta}$ of the parameters for each one of the homogeneous, heterogeneous, and hierarchical model takes the following form:

$$\boldsymbol{\theta}_{hom} = (a, b, s) \quad (2.14)$$

$$\boldsymbol{\theta}_{het} = (a_1, \dots, a_{18}, b_1, \dots, b_{18}, s_1, \dots, s_{18}) \quad (2.15)$$

$$\boldsymbol{\theta}_{hi} = (\mu_a, \sigma_a, \mu_b, \sigma_b, \sigma_s, a_1, \dots, a_{18}, b_1, \dots, b_{18}, s_1, \dots, s_{18}) \quad (2.16)$$

The dimensionality of the prior $p(\boldsymbol{\theta})$, posterior $p(\boldsymbol{\theta}|\text{data})$, and likelihood $p(\text{data}|\boldsymbol{\theta})$ is determined by the dimension of the $\boldsymbol{\theta}$ vector.

Box 2.3 Hierarchical probabilistic model of the drivers' reaction time problem

The hierarchical model assumes that all drivers are modelled with a global set of hyperparameters and each driver is modelled with a separate set of parameters. The inferred posteriors of the hyperparameters are common for all 18 drivers in the dataset and the inferred posteriors of the parameters are unique for each one of the 18 drivers in the dataset.

$$\mu_a \sim \text{Normal}(100, 100) \quad (2.17)$$

$$\sigma_a \sim \text{HalfNormal}(100) \quad (2.18)$$

$$\mu_b \sim \text{Normal}(10, 10) \quad (2.19)$$

$$\sigma_b \sim \text{HalfNormal}(10) \quad (2.20)$$

$$\sigma_s \sim \text{HalfNormal}(50) \quad (2.21)$$

$$a_i \sim \text{Normal}(\mu_a, \sigma_a) \quad (2.22)$$

$$b_i \sim \text{Normal}(\mu_b, \sigma_b) \quad (2.23)$$

$$s_i \sim \text{HalfNormal}(\sigma_s) \quad (2.24)$$

$$rt_i \sim \text{Normal}(b_i \cdot n + a_i, s_i), \quad (2.25)$$

where $i \in \{1, 2, \dots, 18\}$ for each one of the 18 drivers included in the dataset of observations. Only the $data_{a_i}$ subset of the observations corresponding to the i -th driver is used for the inference of the i -indexed parameters.

Assumptions:

- The parameters of each driver are independent, $a_1 \perp \dots \perp a_{18} \perp b_1 \perp \dots \perp b_{18} \perp s_1 \perp \dots \perp s_{18}$.
- The hyperparameters are independent, $\mu_a \perp \sigma_a \perp \mu_b \perp \sigma_b \perp \sigma_s$.
- Drivers' observations and the observations of each driver are independent; $data_{a_i} \perp$ and $data_{a_{i1}} \perp \dots \perp data_{a_{iM}}$.
- The data depends on the hyperparameters only through the parameters; the data is independent of the hyperparameters given the parameters: $data_{a_i} \perp \mu_a, \sigma_a, \mu_b, \sigma_b, \sigma_s | a_i, b_i, s_i$ [Hyvönen and Tolonen, 2019, Chap.6.1].

The complexity of the denominator of Bayes' rule $p(\text{data})$ is also determined by the dimensionality of the θ vector. In equations 2.4 and 2.5 the single sum or integral is replaced by a number of summations or integrals, respectively, one for each parameter in θ . For example, in the case of the homogeneous model of the drivers' reaction time problem this expression for the

calculation of the denominator consists of 3 integrals, in the case of the heterogeneous model it consists of 54 integrals, and in the case of the hierarchical model it consists of 59 integrals.

2.2.4.2 The Difficulty with the Denominator

The calculation of Bayes' rule's denominator for a single or double-parameter models with continuous parameters is possible, but in cases of multi-parameter models the integral in the expression 2.5 becomes multi-dimensional and its analytical solution can become intractable. Doing Bayesian statistics involves the solution of analytically intractable expressions like the denominator of Bayes' rule for the computation of the posterior distribution [Lambert, 2018a, Chap. 12].

Various methods were suggested through the years to give a solution to this problem and allow the use of Bayesian analysis for more complex models. The main tool to do Bayesian statistics for many years was *conjugate priors*, which allows the problem to be solved analytically. With conjugate priors the mathematical form of the prior and likelihood are jointly chosen to ensure that the posterior is in the same family of distributions as the prior [Lambert, 2018a; Spiegelhalter and Rice, 2009]. However, conjugate priors only allow the use of particular combinations of likelihoods and priors and therefore, they are often very limiting and too simple for most real life examples [Lambert, 2018a, Chapter 9.7].

In 70s and 80s, numerical integration methods based on analytic approximations or quadrature were developed [Spiegelhalter and Rice, 2009] providing more flexibility in the selection of prior distributions and likelihoods, but unfortunately, these methods suffer from the curse of dimensionality and do not scale well for models with many parameters [Lambert, 2018a, Chap. 12.4,12.5].

2.2.4.3 Markov Chain Monte Carlo

In early 90s, Markov Chain Monte Carlo (MCMC) became a promising indirect method for doing Bayesian statistics offering a solution to the aforementioned problems [Spiegelhalter and Rice, 2009]. MCMC algorithms are based on dependent sampling to produce an estimation of the posterior distribution's form by skipping the calculation of the denominator. As Lambert [2018a, Chap. 12.8] explains very eloquently, sampling the posterior based on the relative posterior density of pairs of points in the posterior space results in a histogram that proxies for the posterior distribution's PDF.

For example, assume two points in the posterior parameter space, ϑ_1 and ϑ_2 , and a ratio of the posterior at these two points $\frac{p(\vartheta_1|\text{data})}{p(\vartheta_2|\text{data})} = \frac{2}{1}$. Creating a sampler that samples twice as many times the first point is all that is needed. The use of the *relative posterior densities* removes the need for calculating the denominator based on the following equations:

$$\frac{p(\vartheta_1|\text{data})}{p(\vartheta_2|\text{data})} = \frac{\frac{p(\text{data}|\vartheta_1)p(\vartheta_1)}{p(\text{data})}}{\frac{p(\text{data}|\vartheta_2)p(\vartheta_2)}{p(\text{data})}} = \frac{p(\text{data}|\vartheta_1)p(\vartheta_1)}{p(\text{data}|\vartheta_2)p(\vartheta_2)}. \quad (2.26)$$

MCMC algorithms consisted of very simple steps of low computational complexity making them an implementable tool based on the technological means of that time. However, MCMC required specialized knowledge to implement and use and it remained a tool used mainly by academics for many years. The wide adoption of Bayesian statistics was inhibited for a long time, as well, because of the practical engineering challenges of the computational methods and the highly specialized knowledge that they required [Coyle, 2018].

2.2.5 Predictions

So far what has been discussed is how the parameters of a probabilistic model can be inferred by the data using Bayesian inference. The capabilities of Bayesian inference are not limited to that. Bayesian inference provides the tools for producing predictions for the observed variables of the probabilistic models which is the ultimate aim of any prediction or forecasting problem. Using random sampling from the posterior and sampling distribution of the model an approximation of the *posterior predictive distribution* can be obtained. This distribution represents the probability distribution of a new data sample, data' , given the current data sample, data [Lambert, 2018a, Chap. 7.8].

The following algorithm is used to retrieve a set of *posterior predictive samples*. These samples are used to graph the histogram of data' , which will consist the estimation of the *posterior predictive distribution*.

Algorithm 1 Sampling posterior predictive samples

Require: $k, n > 0$
1: **for** $i \leftarrow 1, k$ **do**
2: $\vartheta_i \sim p(\boldsymbol{\theta}|\text{data})$
3: **for** $j \leftarrow 1, n$ **do**
4: $\text{data}'_{ij} \sim p(\text{data}|\vartheta_i)$
5: **end for**
6: **end for**

In step 2 a value of the parameters is sampled from the posterior distributions. In step 4 a value of the observed variable is sampled from the sampling distribution conditional on the value of the parameters that was sampled in step 2.

2.2.6 Example of Bayesian Inference

This section demonstrates the differences in the estimates of the regression line among the homogeneous, heterogeneous, and hierarchical probabilistic models of the drivers' reaction time

model presented in Boxes 2.1, 2.2, and 2.3, respectively. Sample-based Bayesian inference was conducted based on MCMC sampling.

Box 2.4 presents a comparative graph with the estimates of the regression line of these probabilistic models and includes the corresponding estimate of the homogeneous non-probabilistic model presented in Chapter 1 (Box 1.1). The inferred posteriors for the intercept a and slope b parameters are same for all drivers in the case of the homogeneous models, while they are unique for each driver in the case of the heterogeneous and hierarchical models.

A phenomenon that is confirmed by the presented data in Fig. 2.1 is the shrinkage [Arnold; Kruschke, 2012a] of the hierarchical models' posteriors towards the overall mean of the corresponding posterior of the homogeneous model. The parameters whose posterior estimates have the highest uncertainty and lie furthest away from the overall mean, see their posterior to shrink the most in hierarchical models. For example, the posteriors of parameters a and b for driver 351 are shrunk towards the overall mean, and are between those of the homogeneous and heterogeneous model. The same happens with the posterior of parameter b for driver 308. Shrinkage is not observed for the posteriors of the hierarchical model in the case of driver 309 though, although they are further away from the overall mean. This happens because the posteriors of the parameters for this driver are quite certain (i.e., tight).

Box 2.4 Homogeneous, heterogeneous, and hierarchical estimations of regression line

Fig. 2.1 presents the estimations of the intercept a and slope b parameters of the regression line for four different drivers from the dataset [Lambert, 2018b]. The estimations of the *homogeneous non-probabilistic*, and the *homogeneous, heterogeneous, and hierarchical probabilistic* model are presented in a different color.

The last row of the figure shows the observed reaction times of each driver, the regression line estimated by the homogeneous non-probabilistic model, and the 97% HDI of the regression line estimated by the probabilistic models.

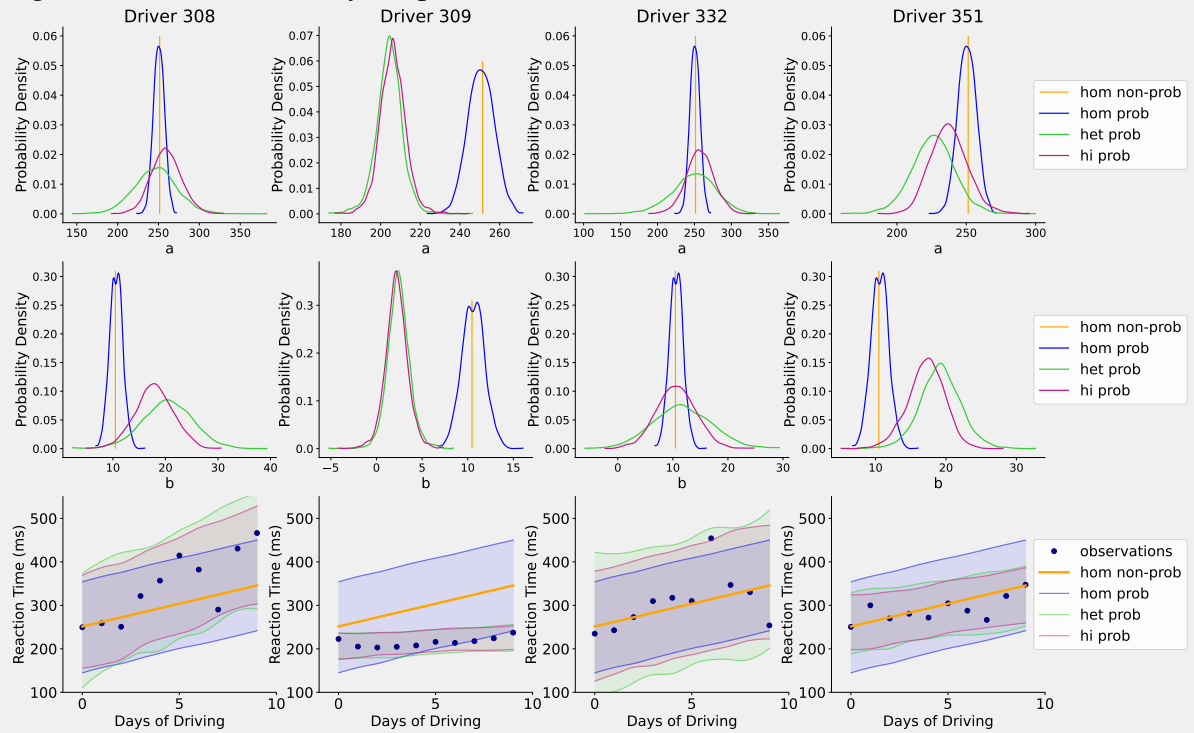


Figure 2.1: Homogeneous, heterogeneous, and hierarchical estimations of the regression line in the drivers' reaction time problem.

2.2.7 Why to Use Bayesian Statistics for the Analysis of Research Data?

In the field of analysis of experimental results, the frequentist and Bayesian statistics are the most common analysis approaches. Bayesian statistics came to the forefront as an alternative to the most common method used for the analysis of experimental data, frequentist analysis, after the breakout of the replication crisis in the 2010s. The replication crisis is an "ongoing methodological crisis in which it has been found that many scientific studies are difficult or impossible to replicate or reproduce" [Wikipedia:ReplicationCrisis].

Shrout and Rodgers [2018] reviewed the questionable methodological research practices in psychology and designated frequentist statistics as one of the sources of questionable practices. Colling and Szűcs [2018] argued that one of the main problems is the misinterpretation of the frequentist statistics' concepts by researchers. Greenland et al. [2016] presented a list of 25

common misconceptions of statistical concepts in statistical interpretation and reporting like *p-values*, confidence intervals, and power.

Cumming [2014] advocated a shift of emphasis from the *hypothesis testing* of the frequentist approach to the *estimation* of the Bayesian approach and proposed the term “new statistics” to infer all the substantial changes in the analysis of research data. Kruschke and Liddell [2018] explained how Bayesian methods achieve the goals of “new statistics” better than frequentist methods. Colling and Szűcs [2018] explored the benefits of the Bayesian statistics as an alternative to frequentist statistics and summarized them into the Bayesian statistics’ “immunity” to researchers’ intentions and its dependency only on the observed data. ShROUT and Rodgers [2018] reviewed the potential of Bayesian analysis for understanding the variation in the replication of an analysis. Researchers from other fields outside psychology advocate the Bayesian statistics, as well. Kay et al. [2016b] argued that Bayesian statistics was more appropriate for the data analyses of experimental data in HCI than the statistical significance testing. Buchinsky and Chadha [2017] encouraged the use of Bayesian statistics over the traditional frequentist statistics in otolaryngology research.

Bayesian inference is an iterative process of updating the prior beliefs by observing every incoming datapoint and estimating the posterior beliefs about the model’s parameters. Bayesian inference is possible for small datasets that may even contain a single datapoint (a single iteration of priors’ update will take place in this case). This attribute of Bayesian inference is valuable because most existing ML methods rely on the amount of available data for making predictions making them inappropriate for applications with only a limited amount of data or one-off problems such as the prediction of the outcome of an election. In such cases, Bayesian inference in comparison to other machine learning or statistical approaches, whose credibility is heavily dependent on the sample size [Jenkins and Quintana-Ascencio, 2020] could easily give an estimation of the outcome.

A problem that is often discussed as a criticism of Bayesian inference is the effect that the priors can have on the inferences [Martin, 2018, Chap. 1; Depaoli et al., 2020; Kruschke, 2018b] and the “subjectivity involved in choosing a prior” [Lambert, 2018a, Chap.5.2]. The more *informative* a prior is and the fewer the observed datapoints are, the stronger the effect of the prior on the posterior distribution is. A prior distribution is informative when the uncertainty it implies for a parameter is small. For example, tight distributions implying small ranges for a parameter are informative priors expressing a strong belief of the modeller about the parameter.

This topic has been extensively discussed in the existing literature with many researchers providing suggestions about the choice of priors [Gelman, 2020]. The use of uninformative (known also as flat or vague) priors, although it would have the least possible impact on the posterior, are not usually recommended [Gelman, 2020]. The use of weakly-informative priors is often suggested when restrictions of the parameters or some information on the scale of the parameters for the order of magnitude of the outcomes are known for a specific likelihood

[Martin, 2018, Chap. 1;Gelman et al., 2017;Kruschke, 2018b]. The use of informative priors is suggested only in the cases that strong prior knowledge or evidence is available through years of research and previous literature [Martin, 2018, Chap. 1]. In cases when the conclusions of a research might differ by a slight change in the priors, this should be reported by the researchers [Lambert, 2018a, Chap. 2.10] who are encouraged to conduct a sensitivity analysis of their priors [Depaoli et al., 2020; Kruschke, 2018b].

2.3 Probabilistic Programming Languages

2.3.1 Purpose

PPLs are a new “breed” of programming languages that are designed to enable inference through probabilistic models [Quddus, 2019]. PPLs can be an either entirely new languages like JAGs [Plummer, 2017], BUGs [Spiegelhalter et al., 2003], Church [Goodman et al., 2008], Stan [Stan Development Team] with some of them offering interfaces to other popular programming languages or hosted in an existing programming language like Edward [Tran et al., 2016] or PyMC [Salvatier et al., 2016] in Python or Webppl [Goodman and Stuhlmüller, 2014] in Javascript.

PPLs are designed to enable inference with general purpose representations. These include semantics for defining a random variable as observed and specifying a likelihood for it. PPLs’ predecessors, simulation languages like Simula [Dahl and Nygaard, 1966], lacked the possibility of defining observed variables and indicating the conditioning over observed data syntactically [Poole and Wood, 2022].

The notion of *interpretation* for the probabilistic models specified in a simulation language differs between the predecessors of PPLs and PPLs themselves. When a probabilistic program specified in such simulation languages was *interpreted*, a set of random samples was generated from the specified random variables (unconditionally) based on the built-in random number generators. When a probabilistic program specified in a PPL is interpreted, the posterior distribution conditioned on the observed data is computed. The main purpose of PPLs is to offer an environment that automates the interpretation of probabilistic programs into posterior distributions.

2.3.2 Expressing a Probabilistic Model in a PPL

In Section 2.2.4.1 probabilistic statements were used for the definition of the homogeneous, heterogeneous, and hierarchical models of the drivers’ reaction time problem. This is a form of textual language used often to hide the many mathematical details involved in the definition of a probabilistic model and offer a level of abstraction.

PPLs are designed to include semantics equivalent to these of probabilistic statements in their syntax as well as indicating conditioning on observed data. The aim is to make the definition of the model more intuitive, hiding the mathematical details and offering flexibility in

```

dataFile = 'evaluation_sleepstudy.csv'
reactions = pd.read_csv(dataFile,
                        usecols = ['Reaction', 'Days', 'Subject'])
with pm.Model() as homogenous_model:
    ## priors
    a = pm.Normal("a", mu = 100, sd = 100)
    b = pm.Normal("b", mu = 10, sd = 10)
    s = pm.HalfNormal("s", sd = 50)
    ## Likelihood
    rt = pm.Normal("rt", mu = a + b*reactions.Days,
                  sd = s,
                  observed = reactions.Reaction)

```

$a \sim \text{Normal}(100, 100)$
 $b \sim \text{Normal}(10, 10)$
 $s \sim \text{HalfNormal}(50)$
 $rt \sim \text{Normal}(b \cdot n + a, s)$

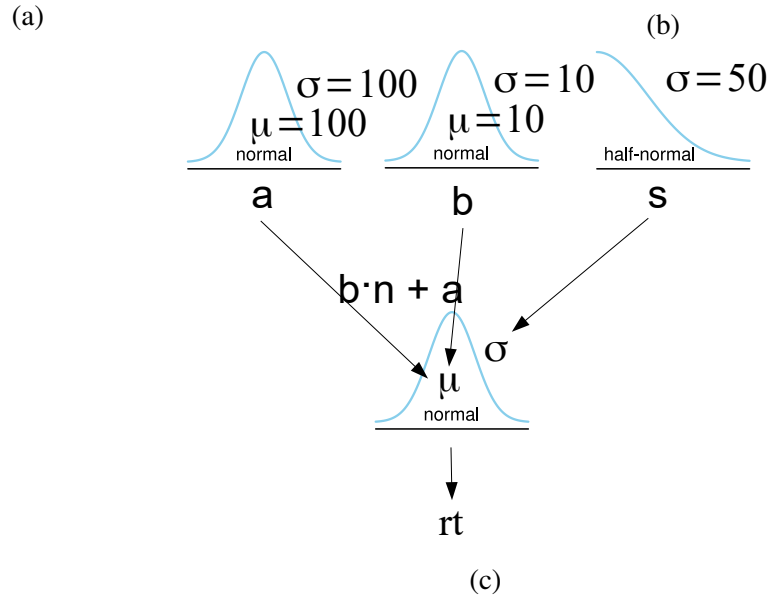


Figure 2.2: Definitions of the homogeneous probabilistic model of the drivers' reaction time problem in (a) probabilistic statements, (b) PPL code (PyMC), and (c) Kruschke-style diagram.

the specification of models. As Tejas D. Kulkarni, one of the pioneers in introducing probabilistic programming in computer vision, says “When you think about probabilistic programs, you think very intuitively when you’re modeling. You don’t think mathematically. It’s a very different style of modeling.” [Kurzweil, 2015].

Fig. 2.2 presents the definition of the homogeneous probabilistic model of the drivers' reaction time problem in probabilistic statements, in PPL code (PyMC) and through a Kruschke-style diagram side by side. There is a one-to-one correspondence among the lines of the probabilistic statements, the lines of PPL code and the nodes of the diagrammatic representation. This is an important attribute of PPLs because the definition of probabilistic models becomes simple, intuitive and more easily comprehensible.

2.3.3 From the Definition of a Model to a Trace of Samples

PPLs automate the specification of probabilistic models and inference benefiting users who only need to use minimal problem-specific engineering [Kulkarni et al., 2015]. As Kulkarni says: “The whole hope is to write very flexible models [...] as short probabilistic code, and then not do anything else. General-purpose inference schemes solve the problems.”

An important advantage of PPLs is that they offer interfaces to efficient and well-tested implementations of MCMC algorithms. Writing a MCMC algorithm and optimizing it requires specialized knowledge. PPLs automate the inference of probabilistic models with their interfaces to MCMC sampling algorithms; the inference can literally be conducted by the push of a button [Martin, 2018, Chap. 2].

Many contemporary PPLs like PyMC and Stan offer ready-made diagnostics tools for the users to check the convergence of the MCMC samplers. Such diagnostics include the Gelman-Rubin statistic [Gelman and Rubin, 1992] using multiple chains and computing the Rhat statistic to check for lack of convergence, autocorrelation, or prior and posterior predictive checks.

The output of a sample-based (MCMC) PPL inference is the *trace*, the set of samples resulting from the inference with an MCMC sampler; the output from “running” the model. The trace can be a complex object containing samples from the prior, posterior, prior predictive and posterior predictive distributions of the model.

2.4 Why Does Modeling Causal Relations Require Extra Modeling Methodologies?

The definition of a probabilistic model determines the relations of the variables and ultimately, the structure of the model; some variables are independent and others are used to determine the distribution parameters of other variables generating a set of hierarchical parent-child relations. Section 1.2.1.1 discussed that the relation of two variables in a model is not necessarily truly causal, unless the model is a causal model or is designed under a causal model. Extra information like this provided by a causal model is required to judge whether the relation of two variables is causal and to model this relation as causal. The fact that *correlation is not causation* could explain why this extra information is required to model causal relations.

This section explains why correlation is not always causation (Section 2.4.1) and how a correlation could be identified as a true causal relation (Section 2.4.2). Finally, it presents well-known existing methodologies for causal modeling (Section 2.4.3).

2.4.1 Correlation is not Always Causation

A research question often encountered in various fields (e.g., sociology, anthropology, medicine) is whether an observed correlation between two variables is a true causal effect: does one of the

variables causes the other?

Let us assume x and y are two variables of interest. Some observed data of these variables are available based on which the two variables appear to be positively correlated (Fig. 2.3(a); figure retrieved from Huszár [2019]); when one increases the other increases, too. This information is not enough to tell us which of the two is true: an increase in x is *because of* an increase in y , or an increase in y is *because of* an increase in x ? Which variable causes the other? In reality, there is a third possible scenario; none of the variables causes the other, they simply appear correlated for other reasons. Why?

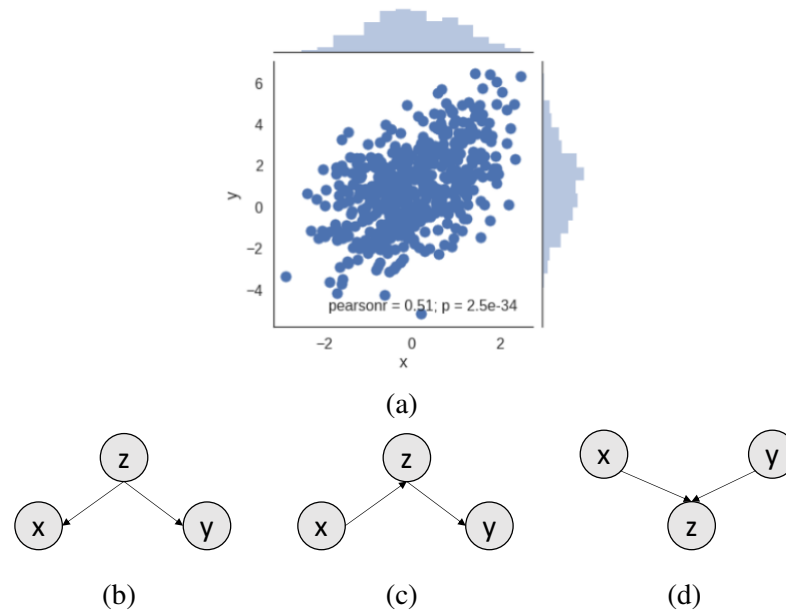


Figure 2.3: Correlation of the x and y variables and the causal diagrams of the three basic three-variable causal primitives of the x , y , and z variables with z being a (b) confounder, (c) mediator, or (d) collider.

Fig. 2.3(b), (c), and (d) present the causal diagrams of three different causal models of the same three variables, x , y , and z . The arrows on a causal diagram represent direct causal links with the direction of the arrow indicating the directionality of causality starting from a cause and pointing to an effect. Two variables in a causal diagram are *adjacent* if they are directly connected by an edge; e.g., x is adjacent with z but not with y in Fig. 2.3(b). According to Greenland et al. [1999] a *path* through the graph is “any unbroken route traced out along or against arrows or lines connecting adjacent nodes”. Correlation flows across paths on a causal diagram, sometimes even against arrows.

A correlation between two variables could be *spurious* due to the existence of *confounding*, namely a third variable (z) that intercepts the path from one to the other being a common cause of the two variables [McElreath, 2020a,b, Chap. 6]. Every time that the value of this third variable changes, the values of the two variables change at the same time and thus, they appear correlated, although they may not even have a causal link (arrow) between them (Fig. 2.3(b)). Confounding creates a bias in the observed correlation of the two variables.

In the absence of confounding, the correlation would express either a direct causal effect between the variables if a direct link exists between them or an indirect one through a *mediator* (Fig. 2.3(c)) [McElreath, 2020a,b, Chap. 6]. This is a third variable (z) intercepting the path from one variable to the other without changing the direction of the path (all arrows in the path follow the same direction). In the case of a confounder or mediator, the path is open for correlation to flow.

Collider is a third case of causal primitive that could associate two variables. The collider is a variable that intercepts the path of the two variables being the common effect of the two variables (Fig. 2.3(d)). The existence of a collider in a causal model blocks the correlation between its two causes; its two causes will not appear correlated unless they have a direct causal link, a common cause, or a mediator. A path between two variables is *blocked* if it has one or more colliders and does not allow correlation to flow from one variable to the other; otherwise it is unblocked. In the case of collider another type of bias, the *selection (Berkson's) bias* [Berkson, 1946; Greenland and Pearl, 2017], could cause observed correlation between the two independent ancestors of a common descendent when a selection from the population of the collider is done based on some criteria.

2.4.2 How can I tell if a correlation is a true causal effect?

To ensure that an observed correlation is a true causal effect, confounder paths should be closed and collider paths open to avoid biases such as confounding or selection bias. There are various methodologies for dealing with these biases. Conducting *interventional* experiments is one such way; an intervention is applied on a variable to externally force it to take a specific value. An intervention on a variable prevents any causes of the variable from affecting it because its value is completely determined externally. This approach is effective for confounding bias because the confounding path is blocked when the intervention is applied on a descendant of a confounder. Interventional experiments need attention with colliders; intervening on a collider opens the path and introduces selection bias.

In cases when interventions might not be possible or ethical (e.g., forcing people to smoke to investigate the effect of smoking on health) and only *observational* studies are possible, different approaches should be followed to make causal inference based on the observed data. One way is statistically *controlling* for confounders by applying stratification or conditioning on them or including them as a covariate to a multivariate (regression) model [Pourhoseingholi et al., 2012]. Attention is again required with the colliders; colliders should not be controlled for to avoid selection bias. The problem with the observational studies is that unless there is well-established prior knowledge about which variables can be colliders, which is usually not the case, it is not possible to know with certainty which variables can be safely controlled for.

A more modern approach involves simulating interventional experiments in cases when observed data is available [Han et al., 2018; Herd and Miles, 2019; Huszár, 2019; Ibeling, 2018;

Sofrygin et al., 2017; Witty et al., 2019] along with some assumptions about the possible underlying causal structures. Huszár [2019] presents a toy example of simulating an intervention on variable x probabilistically and using visualization to observe the effect of this intervention on another variable y in three different cases of causal models. Fig. 2.4 from [Huszár, 2019] demonstrates this example.

Three different Gaussian samplers are created for the two random variables, x and y , which have a similar joint distribution (first row of Fig. 2.4(a)). Each one reflects a different causal structure for the two variables presented. These causal structures are represented by causal diagrams shown in the first row of Fig. 2.4(c). The three simulators are run and observations are generated for the two variables. These observations are plotted in the scatter plots in the second row of Fig. 2.4(a).

The samplers are altered to include an intervention on variable x : $x = 3$ (first row of Fig. 2.4(b)). The causal diagrams are altered to depict the intervention in the second row of Fig. 2.4(c); any incoming arrow to x from its causes is removed because the causes of x cease affecting it after the intervention (these pruning operations applied on a causal diagram after an intervention are often encountered in the literature as *mutilation* of the causal diagram). The simulators are run again and the samples of y after the intervention are plotted in the scatter plots in the second row of Fig. 2.4(b).

The distribution of y changes after the intervention in the first simulator (the distribution's value range from $[-4, 6]$ became $[-1, 8]$), while it remains the same in the second and third simulator (the distribution's value range remains $[-4, 6]$). This happens because in the first causal model y is a descendant of x and is affected every time x 's value changes, while in the second and third causal model y becomes independent of x after the intervention (the intervention blocked the path from one to the other) and does not get affected by x when its value changes.

2.4.3 Causal Modeling Methodologies

The difficulty of interpreting the correlations in observed data as causal or non-causal relations leads to the need for methodologies of *causal inference* from observed data; how could observed data be used to infer whether the relations of variables are causal? The approaches to causal inference may be broadly divided into two schools of thought; the *potential outcomes* and the *causal graphical models*. The potential outcomes framework roots back to the work of Hume [1748], was first proposed by Neyman [1923], and extended into a general framework for causal inference by Rubin [Sekhon, 2007]. The causal graphical models were first used by Wright in his path analyses [Wright, 1921] and Pearl was instrumental in extending them into a general unified framework for causal inference [Larsen, 2021; Pearl, 2010].

The two schools differ in how they define the notion of causal effect. According to Rubin [1972] the causal effect is defined as following:

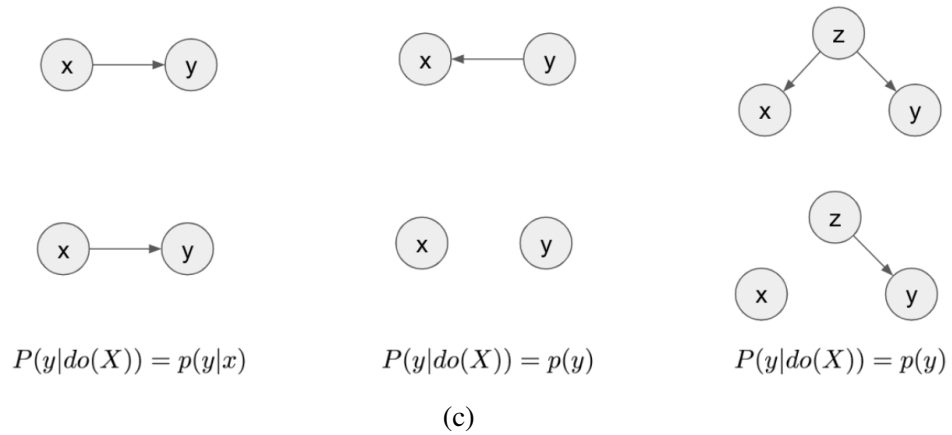
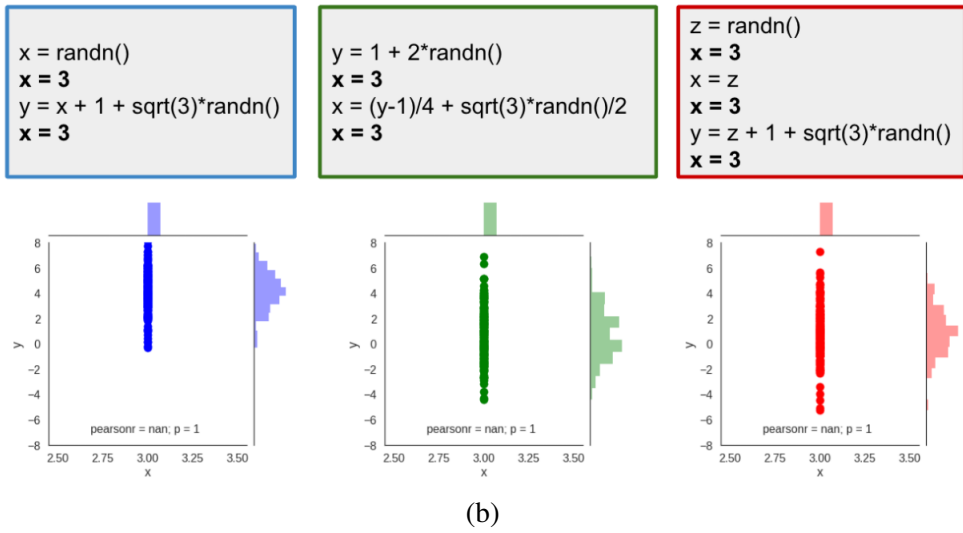
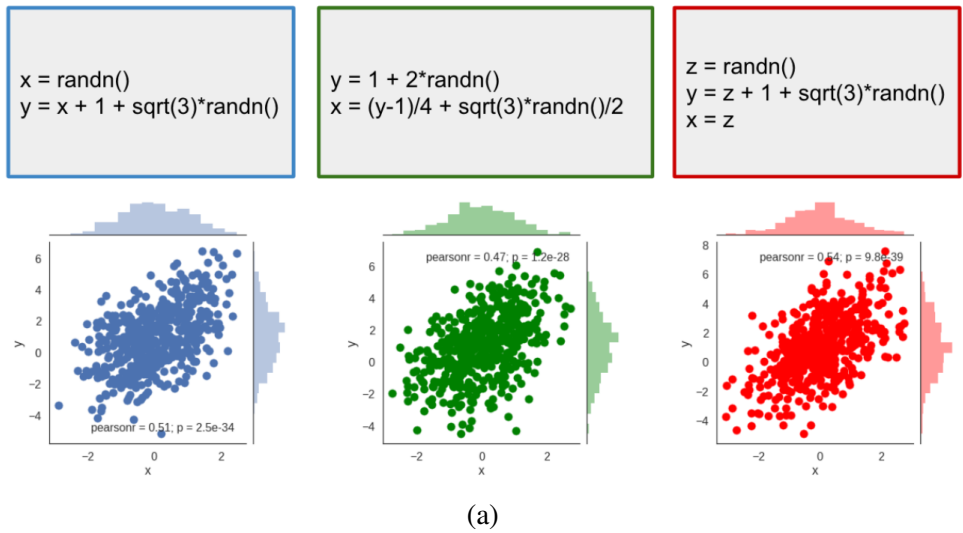


Figure 2.4: Toy example from Huszár [2019] demonstrating the effect of simulated interventions in three similarly-correlated two-variable (x and y) systems. (a) Three simulators for the similarly-correlated two-variable systems and scatter plots of the x - y simulated observations, (b) the simulators altered to reflect the $x = 3$ intervention and the x - y scatter plots after the intervention, and (c) the underlying causal structure of each system before (upper row) and after the intervention (lower row).

Intuitively, the causal effect of one treatment, E , over another, C , for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1 : 'If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone,' or 'because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone.' Our definition of the causal effect of the E versus C treatment will reflect this intuitive meaning.

Based on this school of thought a causal effect of a “treatment” on a unit exists when the difference of the two potential outcomes of the “treatment” on the unit is not zero. Then, it could be said the “treatment” and the unit have a causal relation.

The “fundamental problem of causal inference” [Holland, 1986] is that it is impossible to observe both potential outcomes on a single unit and estimate the causal effect of the treatment; you either take the aspirin now or you don't. Thus, the problem of causal inference according to the potential outcomes framework is reduced to a problem of missing data (counterfactuals). Then, given some assumptions various statistical tools could be used for the estimation of the missing information. Rubin contributed to the expansion of various statistical techniques for the conduction of causal inference within this context, like instrumental variables [Angrist et al., 1996].

Judea Pearl advocated that causal inference “requires new mathematics and that causal questions cannot be solved within existing paradigms for probabilistic inference” [Lattimore and Rohde, 2019b]. He introduced the *do-calculus* [Pearl, 1995], a graphical inference tool comprising a set of rules to express the *conditional probabilities of the variables given the interventions*, $P(y|\text{do}(X))$ ¹, in terms of the *observed conditional probabilities of the variables*, $p(y|x)$, through querying causal diagrams. According to Pearl, the causal effect of x on y is determined by whether the post-intervention distribution $P(y|\text{do}(X))$ differs from the pre-intervention distribution $p(y|x)$; if they are the same, it means that x affects y causally.

In Huszár [2019]'s example the mutilated causal diagrams after the intervention in the second row of Fig. 2.4(c) were generated after applying the rules of do-calculus. The third row Fig. 2.4(c) presents how the conditional distribution of y given the intervention on x , $P(y|\text{do}(X))$, is expressed based on the observed, conditional and non-conditional, probabilities of y in each one of the three causal models. The post-intervention conditional distribution $P(y|\text{do}(X))$ is the same with the pre-interventional conditional distribution $p(y|x)$ only in the case of the first causal model. This means that only based on the first model x affects y causally.

Causal diagrams and do-calculus gave rise to specialized algorithms for *automatic causal discovery*, which try to learn the causal structure of data by applying do-calculus given some assumptions [Glymour et al., 2019]. These estimates, albeit incomplete, can be used by users to

¹The symbols P and p are used to denote the post- and pre-intervention probability distributions, respectively.

decide upon the variables they can control for or intervene on.

2.5 Theories of Human Visual Perception

2.5.1 Why is it Important to Understand Human Visual Perception?

Visualization is an “external artifact” functioning as a cognitive tool for humans [Ware, 2012] who use it in tasks like making decisions or forming beliefs and judgements. The ways that humans perceive visual information and turn it into cognitive actions play an important role in visualization. Visualization cannot alone generate human actions without the intervention of human’s perception that will receive information, process it, draw conclusions from it, and then decide what it will do with it.

Understanding the ways humans perceive (process) the visual information is vital when designing visualization tools if the aim is to maximize the “cognitive throughput”, namely the amount of valuable cognitive work done per unit of time [Ware, 2012, Chap. 11]. As Shneiderman [1996] states very eloquently

the bandwidth of information presentation is potentially higher in the visual domain than for media reaching any of the other senses. Humans have remarkable perceptual abilities, that are greatly under-utilized in current designs. Users can scan, recognize, and recall images rapidly, and can detect changes in size, color, shape, movement, or texture. They can point to a single pixel, even in a megapixel display, and can drag one object to another to perform an action.

Designing visualizations that exploit humans’ perceptual and cognitive capabilities could lead to really powerful cognitive tools.

The field of visual perception has extensively been studied especially by psychologists. Many theories exist and much specialized knowledge is involved. There is also extensive empirical work which investigated how people perceive information visually and established new or confirmed existing theories. This section focuses on theories of visual perception that are relevant to the context of this thesis; Section 2.5.2 discusses theories of visual perception of uncertainty, Section 2.5.3 discusses the visual perception of animation, and Section 2.5.4 discusses the visual perception of interaction. Most of these topics will be revisited in Chapter 3.

2.5.2 Visual Perception of Uncertainty

2.5.2.1 Heuristics and Biases When Judging Probabilities

The ways that people reason about uncertainty have extensively been studied in literature. It has been found that people have difficulty reasoning about uncertainty based on standard probability

formats (e.g., probability, percentile, odds). They tend to rely on heuristics or rules of thumb to make judgments about probabilities rather than apply a correct mathematical solution to make rational choice.

Tversky and Kahneman [1974] have systematically studied the heuristics people use to make judgments about uncertainty. For example, they found that people tend to rely on *representativeness* (i.e., the degree to which an event (individual) is representative of a process (population)) to judge the probability of an event to originate from a process. The representativeness heuristic make people to be insensitive to prior probability or sample size.

Kahneman and Frederick [2002] talked about *substitution*, an unconscious strategy people often employ when they have to deal with difficult information; people tend to substitute a difficult mental computation with an easier one. The substitution heuristic can lead to quick decisions which might also be correct if the judgment is detrimental (e.g., the “deterministic construal error” occurring when people try to substitute visual uncertainty information for deterministic information [Joslyn and LeClerc, 2012]). However, this heuristic was not found to be always effective [Belia et al., 2005].

Another heuristic that Tversky and Kahneman [1974] found is the *availability* according to which people assess a probability of an event by the ease with which it can be brought to mind [Tversky and Kahneman, 1974]. A characteristic example that Tversky and Kahneman [1974] give is that “a class whose instances are easily retrieved [e.g., because they are more easily memorable] will appear more numerous than a class of equal frequency whose instances are less retrievable”. All these heuristics, although they might benefit judgments of probability up to a point, they lead to “predictable biases” [Tversky and Kahneman, 1974] making some researchers argue that “human logic is *systematically* flawed” [Padilla et al., 2021b].

2.5.2.2 Frequency Framings of Probability Improve Probability Judgments

In response to the theories of systematic flaw in people’s judgments of probabilities, Gigerenzer [1996] suggested that this systematic flaw in people’s judgments of probabilities is influenced by the format in which probabilities are communicated; confusing formats to express probability (e.g., as a percentile) that people cannot easily understand could be replaced by formats that naturally express probability in ways “how people experience probability throughout their lives (e.g., changing 10% to 1 out of 10)” [Padilla et al., 2021a]. Hence, Gigerenzer [1996] suggests expressing probabilities as frequencies to make probabilistic judgments more intuitive and less reliant on heuristics [Gigerenzer, 1996; Gigerenzer and Hoffrage, 1995; Hoffrage and Gigerenzer, 1998].

This theory gave rise to many visualization designs that present distributional information framed as frequency, e.g., quantile dotplots [Kay et al., 2016a] or Hypothetical Outcome Plots (HOPs) [Hullman et al., 2015] (these will be explained in detail in Chapter 3). Extensive empirical research has been conducted to evaluate such visualization designs [Fernandes et al., 2018;

Hullman et al., 2015; Kale et al., 2019; Kale et al., 2021; Kay et al., 2016a] and has found that frequency-framed uncertainty shown in visualizations can improve viewers' accuracy and recall compared to visualizations that show only probability distributions or summary statistics (e.g., intervals, mean, median) [Padilla et al., 2021a]. The frequency-framed uncertainty visualizations require that the viewer exploits their cognitive mechanisms to acquire an intuitive sense of the underlying distribution and according to Kale et al. [2021] they do not "allow the viewer to fixate on summary information such as a mean" [Padilla et al., 2021a]. Hence, presenting probability framed as frequency in visualizations is recommended as a best practice by researchers in literature.

More about the theories of uncertainty visualization can be found in [Padilla et al., 2021a].

2.5.3 Visual Perception of Animation

Moving patterns can be used to represent motion in dynamic systems like vector fields or motion can be used to display dynamic data [Ware, 2012]. For example, animation could be used to convey causality by exploiting the "temporal cue" an intervention produces when applied on a system: the intervention is the most likely cause of any subsequent changes [Lagnado and Sloman, 2004].

Animation can offer a rich "vocabulary of expressive motion" [Ware, 2012] that could be exploited for the communication of other aspects of data than its timely evolution. Ware [2012] argues that "the perception of dynamic patterns is not understood as well as the perception of static patterns, but we are very sensitive to patterns in motion, and if we can learn to use motion effectively it may be a good way to display certain aspects of data". The visual system's ability to "form gist representations from experience" [Kale et al., 2019] could be exploited to communicate difficult to grasp concepts like probability. For example, animation has been used to communicate samples from distributions [Hullman et al., 2015; Kale et al., 2019].

Tversky et al. [2002] was skeptical about the superiority of animation over static visualizations. The animated and static versions of visualizations often compared in studies are not informationally equivalent [Kim et al., 2019]. Tversky et al. [2002] suggested two high-level principles for effective animation; the *Congruence Principle* states "the content and format of the graphic should correspond to the content and format of the concepts to be conveyed" and the *Apprehension Principle* states that "graphics should be accurately perceived and appropriately conceived". According to the first principle studies comparing animated with static visualizations should ensure "equivalence between animated and static graphics in content or procedures". According to the second principle animations should not be "too complex or too fast to be accurately perceived".

2.5.4 Visual Perception of Interaction

An interactive visualization usually exploits various interactive graphical elements like sliders, drop-down menus, radio buttons to enable the exploration and query of data according to the users' demands. Ideally information should be displayed as needed and disappear when not needed for an interactive visualization to support users' thinking processes [Ware, 2012]. The need for drilling down and finding data according to users' demands requires visualizations that would be appropriately designed. Shneiderman [1996] developed the *Visual Information Seeking Mantra* to be used as a guideline for designing interfaces that would support "information-seeking behaviors" [Ware, 2012] more effectively: "Overview first, zoom and filter, then details on demand".

According to Ware [2012] interactive visualization consists of three levels of "interlocking feedback loops". In the *data manipulation* loop users select or move objects using basic eye-hand coordination skills. In the *exploration and navigation* loop users explore the data space, form a cognitive model for it, and try to navigate the data space to "find their way" to the information they need. In the *problem-solving* loop users form hypotheses about the data and refine the view of the data to confirm or reject their hypotheses.

These feedback loops depending on the application create requirements that should be accounted for by the designs of the visualization (see Ware [2012] for a detailed discussion about this topic). Human cognitive capacity is limited. The human "working memory" cannot store information for long or cannot support high cognitive load to conduct complicated "calculations". An implication of this is that delays in the feedback loops or increase of the cognitive load can lead to low rates of information uptake [Ware, 2012]. Another implication is that the "cognitive context" users form by perceiving the presented information and processing can be easily lost when users need to switch attention between thinking and interacting with the interface in a feedback loop [Ware, 2012]. All these reveal the difficulties and the precision that is required for designing interactive visualizations.

Chapter 3

Literature Review

3.1 Summary

This chapter presents the literature review of existing work in topics relevant to the research presented in this thesis. Uncertainty is a cornerstone of probabilistic models; it is encountered in the prior and the posterior distributions. When probabilistic models need to be communicated uncertainty cannot be omitted. Section 3.2 reviews the existing work as to uncertainty visualization approaches and empirical evidence about their effectiveness. Static, animated, and interactive visualization approaches are considered. The emphasis is primarily on univariate uncertainty representations as they are required for representing the (univariate) probability distributions of variables in a probabilistic model (i.e., the marginal distributions of the model's joint distribution).

Probabilistic models have a structure that is determined by the associations of their variables. This structure presents the order of the operations in the modelled data generating process. The structure of a probabilistic model should be communicated along with the uncertainty to provide a complete picture of the model. Section 3.3 explains how Bayesian network (i.e., a graphical representation) can be used to represent the structure of probabilistic models. It reviews the existing tools for generating such a graphical representation.

The structure of a probabilistic model does not necessarily reflect the cause-effect relations of the variables, unless it is designed under a causal model. Section 3.3 explains also how the structure of causal models is presented graphically. It reviews existing tools for graphically presenting, exploring, and validating learned causal structures of data. It reviews existing evaluation work of visualization in causal reasoning (i.e., how visualization can be used to help people reason about which variable causes which other variable). This last part in Section 3.3 is related to the last part of the research presented in this thesis (Chapter 6). A main focus of this research was how visualization could support causal reasoning.

Interaction is not often used for the communication of uncertainty in literature. Thus, Section 3.4 presents interesting ideas on interactive visualization from the literature that served as

inspiration for the work presented in this research.

3.2 Visualization of Uncertainty

The perception of uncertainty by people contains many misconceptions or biases leading to incorrect or imprecise judgements of uncertainty [Belia et al., 2005; Gigerenzer, 1996; Kahneman and Tversky, 1974; Tversky and Kahneman, 1971, 1973, 1974]. This fact makes many analysts sceptical about whether uncertainty should be communicated [Hullman, 2020]. However, hiding uncertainty from people could lead to uninformed decisions. This could prove disastrous in crucial decision-making occasions like in healthcare, or stock-market trading where the risk which a decision entails should be considered.

For this reason, there was a huge research effort conducted in the last decades about whether and how visualization could be used to communicate uncertainty effectively. This section reviews the most relevant work presented in this field. Section 3.2.1 presents the existing approaches in the visualization of uncertainty categorized as static, animated, and interactive. Section 3.2.2 presents the relevant experimental work in the evaluation of uncertainty visualizations. Section 3.2.3 provides a discussion about the challenges in designing and evaluating uncertainty visualizations overall.

3.2.1 Existing Approaches

The existing uncertainty visualization approaches that will be presented in this subsection are categorized in three categories; *static*, *animated*, and *interactive*. Static uncertainty visualizations rely on static graphics and visual encodings to communicate uncertainty. Animated uncertainty visualizations rely on animation to communicate uncertainty. Interactive uncertainty visualizations rely on interaction techniques and interactivity to communicate uncertainty.

The basis of this categorization is the level of *user engagement* required to exploit the provided uncertainty visualization for making a judgement or decision. Static visualizations lacking any motion or interaction require users to passively preceive the presented information. Animated visualizations use motion to add movement to static graphics and require users to more actively engage their mental mechanisms. Interactive visualizations require users to come into a closed-loop with the visualization by interacting with it, querying the information they need, wait until it is displayed, perceive it, process it, and query a new piece of information to refine the perceived “picture” of the information [Ware, 2012, Chap. 1,10].

This categorization serves the scope of this research: the use of interaction and animation in the communication of uncertainty in probabilistic models to increase user engagement and make difficult and unintuitive mathematical formulations more easily understood. The uncertainty visualization approaches suggested or investigated in empirical studies in the literature

are presented in this subsection. The related empirical work will be presented in the following subsection.

3.2.1.1 Static Visualisations of Uncertainty

Various visual attributes of graphics that can be interpreted by human visual perception encode uncertainty-related information; the position, shape, size, color, number of presented shapes, density of glyphs, etc. The existing uncertainty visualizations presented here are categorized based on the visual attributes they rely on to encode uncertainty.

Size and Position. Typical uncertainty visualizations that encode uncertainty in their size (i.e., width, height) and position are *error bars*, *bar charts with error bars*, and *Box plots*. These plots are called often *interval plots* because they encode some statistical interval (confidence or credible interval, interquartile range, standard error, standard deviation) in their size.

An error bar is a line passing from a point representing a measure of central tendency (e.g., mean or median) of a univariate variable and is drawn parallel to the axis corresponding to this variable. The two ends of the line can correspond to various statistics describing the variation in the observed data: e.g., a particular confidence interval of the variable's distributions, the mean value plus/minus the standard deviation, or the mean value plus/minus the standard error.

A bar chart with an error bar consists of a bar plot whose height represents some aggregation measure like the mean value, and an error bar drawn vertically to the top edge of the bar plot and passing from the center of it.

The Box plot [Frigge et al., 1989; Haemer, 1948; Spear, 1952; Tukey, 1977] is a more informed representation of uncertainty as it presents a five-number summary of the variable's data distribution. It is represented by a box with whiskers. The first (Q1) and third quartile (Q3) of the observed data correspond to the left-right or bottom-top edges of the box depending on the orientation of the box; horizontal or vertical, respectively. The whiskers are usually drawn within the 1.5 interquartile range ($IQR = Q3 - Q1$) but can stand for several other things; the minimum and maximum observed values, one standard deviation above and below the mean of the data, the 95% central quantile. A line enclosed in the box and drawn parallel to the Q1 and Q3 edges represents the median (Q2).

Shape. Various uncertainty visualizations use the shape of the presented graphic to encode the density of the distribution. These visualizations are often called *density plots* because they encode the density of the distribution.

In the case of continuous variables, *Kernel Density Estimate (KDE)* plots depict the PDF of continuous variables in the shape of the presented curve. In the case of discrete variables, *histograms* can be used instead to depict the PMF of discrete variables. KDE plots and histograms can be used to depict uncertainty in 2D, as well. Enhanced Box plots like *hist plots*

[Benjamini, 1988], *vase plots* [Benjamini, 1988], *Box-percentile plots* [Esty and Banfield, 2003], or *violin plots* [Hintze and Nelson, 1998] encode probability density in their shape. Another extended Box plot that encodes uncertainty information in its shape is proposed by Potter et al. [2010]; *Potter's summary plot* incorporates higher order descriptive statistics. Potter et al. [2010] presents also an extension of this to two-dimensional distributions.

Color Intensity. Various uncertainty visualizations use the color intensity to encode probability density. These visualizations are often called *gradient plots*.

A *density strip* [Jackson, 2008] is a strip of a usually single-hue colormap; the darkness level is proportional to the probability density at each point.

Correll and Gleicher [2014] present a version of *gradient plot*. A distribution is represented by a box whose left-right or bottom-top edges (depending on the orientation of the box) correspond to the 100% confidence interval. A solid enclosed line parallel to the confidence interval edges represent the median. The box contains densely stacked lines parallel to the median line. Within the 95% confidence interval the lines are fully opaque. Outside the 95% confidence interval the opacity decays with respect to the cumulative probability for the absolute value of the variable based on an underlying t-distribution, until the lines at the edges of the box become completely transparent.

Density and Spatial Arrangement. The uncertainty visualizations in this category present discrete outcomes (samples from distributions) and are often called *discrete plots*.

The *rug* and *strip plots* [Feigelson and Babu, 1992; Yi] represent the observations of a univariate variable as tick or dot markers, respectively, across their axis. *Scatter plots* can be used to represent two-dimensional observations on a 2D plane. The density of the markers in these plots encodes the probability density of the variable.

Dot plots [Wilkinson, 1999] are used to represent individual observations on a continuous scale using dots that are locally displaced in a direction orthogonal to their axis to prevent overlapping. Dot plots could be considered as a discrete analog of the KDE plot.

Number of Presented Shapes. The uncertainty visualizations in this category present uncertainty as discrete outcomes (i.e., draws from a probability distribution) and allow probability estimation through counting. They encode uncertainty in the number of presented shapes using a frequency framing for it (i.e., they represent a percentile 10% as a ratio, e.g., 1 out of 10). These visualizations are often called *frequency-framed (discrete) plots*.

Icon arrays use a shape or small picture that is repeated as many times as the denominator of the ratio with a number of them corresponding to the numerator altered usually by colour.

Quantile dotplots [Kay et al., 2016a] represent distributions where dots are sampled proportional to the quantiles of the distribution. Icon arrays are appropriate for a small number

of discrete possible outcomes, while quantile dotplots can be used as an “frequency-framed alternative for displaying uncertainty for a continuous variable” [Padilla et al., 2021a]. Quantile dotplots could be considered as a discrete analog of the KDE plot.

3.2.1.2 Animated Visualisations of Uncertainty

Communicating Uncertainty Through Animation of Simulated Data. In geospatial applications, Ehlschlaeger et al. [1997] suggested the use of animation to present the uncertainty in spatial data, which resulted from coarse coverage of the area in sampling. Spatial uncertainty is modelled by specialized stochastic models, which generate “many potential realizations” of the spatial data (e.g., surfaces). Ehlschlaeger et al. [1997] suggested the use of random and serial animation of these “potential realizations” visualizations. Animation could help researchers get a better insight into the uncertainty in spatial analysis and recognize spatial autocorrelation.

Evans [1997] suggested the use of “flickering” to present the reliability of data in land cover maps; a frame showing a map that presents all the data alternates with a frame showing a bivariate map that presents only highly certain data (color-on cells show highly certain data, color-off cells show highly uncertain data).

Kwock et al. [2010] proposed the use of animation to present the statistically modeled uncertainty about a large dataset in density plots instead of the actual PDF of the datapoints. The aim was to provide visual summaries of large datasets instead of conducting expensive computations of high quality density plots. They proposed the *probabilistic plots* which present a set of animated frames showing a scatter plot or a parallel coordinate plot. In each frame the plots present a different set of random samples from the underlying distribution. In these animated plots the regions of high probability density remain stable, while the outliers “intermittently flicker in and out of existence” [Kwock et al., 2010] drawing viewers’ attention to them.

Animated Frequency-framed Plots. Hullman et al. [2015] inspired by the ideas of using simulated data to communicate uncertainty introduced the Hypothetical Outcome Plots (HOPs), a form of *animated frequency-framed plot*. HOPs present a set of frames each of which shows a random draw from a distribution and is displayed for a short time (i.e., <500ms). Viewers of HOPs need to use their cognitive mechanisms to acquire an intuitive sense of the underlying distribution (i.e., estimate the mean value or identify the less possible outcomes). HOPs force users to “account for uncertainty in their understanding of the data” [Padilla et al., 2021a].

Animated Transitions for Statistical Data Graphics. This category of visualizations exploits *animated transitions* to transform a graphic into another by possibly applying some operation to the underlying data. *Transitions* can be thought of as “state changes” [Heer and Robertson, 2007]. Heer and Robertson [2007] describe the process of transition and explain what the challenge in designing animated transitions is:

Analytic operators make changes to the semantic model of the data graphic, editing the data schema, data values, or visual mappings. This in turn results in changes to the graphical syntax. In static transitions, the original syntactic form is simply replaced with the new one. The challenge of designing animations is to visually interpolate the syntactic features such that semantic changes are most effectively communicated.

Heer and Robertson [2007] created the *DynaVis*, a visualization framework supporting animated transitions for statistical data graphics like bar charts, pie charts, and scatter plots.

Kim et al. [2019] suggested the use of animated transitions to communicate the meaning of aggregation operations (i.e., min, max, mean, median, count, standard deviation, interquartile range) over univariate distributions. The original design concerned the transitioning from unaggregated to aggregated dot plots. They provide an example that illustrates the animated transitions for the arithmetic mean: individual dots shift along a horizontal line representing the average, and transform to residual lines, which then collapse synchronously in a way that the upper and lower parts cancel out to form the average. They extend their design to transitions to depict the construction of box plots, histograms, and means and confidence intervals calculated via bootstrapping.

3.2.1.3 Interactive Visualisations of Uncertainty

Although interaction is broadly used in the communication of complex data or ideas [Faith, 2007; Nguyen et al., 2020; Sankaran and Holmes, 2018], it is not often encountered in the designs of uncertainty visualizations. The effect of interaction was investigated in tasks like Bayesian reasoning [Khan et al., 2018; Mosca et al., 2021; Tsai et al., 2011], or graphical prediction of uncertainty or trends in data [Collective, 2019; Hullman et al., 2018; Kim et al., 2017, 2018]. The following paragraphs present the ways to use interaction in Bayesian reasoning and graphical prediction that were explored in the literature.

Interaction in Bayesian Reasoning. Tsai et al. [2011] suggested the use of an *interactive frequency box diagram* to present the different components of a conditional probability in Bayesian reasoning problems. The frequency box diagram is suggested to accompany the frequency-framed textual description of the problem. The frequency box diagram consists of a large box that is subdivided into small squares that symbolize the entire population (each small square represents one person in the population). A set of checkboxes is provided for toggling on or off the components of a problem (subpopulations) in any order or combination using a different color to highlight each component. Improvement in performance of participants who used the interactive frequency box diagram was found in comparison to participants who only relied on the textual descriptions of the problem.

Khan et al. [2018] explored the effect of adding interaction to the *double-tree diagram* [Khan et al., 2015] on people’s performance in a Bayesian reasoning task. The double-tree diagram can be used to visualize the “double branching structure of a Bayesian problem”; “the false-positive/true-positive and false-negative/true-negative symmetry of the problem” is directly represented [Khan et al., 2018]. Khan et al. [2018] asked participants in a user study to construct the double-tree diagram using drag-and-drop based on the textual description of the problem before responding to the questions of the task. The results showed that interactively constructing the double-tree diagram was not beneficial for participants’ performance.

Mosca et al. [2021] explored the effect of adding interaction to the stimulus used in a Bayesian reasoning problem concerning a disease in population. The stimulus was an icon array showing the subpopulations: “Have Disease”, “Do Not Have Disease”, “Test Positive”, and “Test Negative”. They experimented with adding *checkboxes* or *drag and drop* to hide or show subpopulations in the icon arrays. They also tested an implementation of *hover* in which users hovering their mouse over an area of text describing a subpopulation in the description of problem, the text and corresponding subpopulation in the icon array were highlighted. Finally, they tested *tooltips* showing up in the icon array when users hovered over a subpopulation and stating the part of the textual description of the problem. None of these interactions seemed to be beneficial in participants’ performance in a user study conducted.

Interactive Graphical Prediction of Uncertainty or Trends in Data. Hullman et al. [2018] suggested a novel interactive, graphical uncertainty prediction technique for communicating uncertainty in experiment results. With this technique users can predict the uncertainty in experiment replications graphically before they see the true sampling distribution. They can sketch their prediction of the uncertainty either as a probability density plot or a quantile dotplot. Hullman et al. [2018] argue that graphically predicting replication uncertainty is an effective way to communicate uncertainty in experimental science.

Kim et al. [2017] suggested the interactive graphical elicitation of users’ prior expectations about trends in data to improve users’ recall of the data. Users who draw their predictions, view their expectations against the actual data, and try to self-explain data are benefited. This interaction with their “internal representations” of the expectations about the data support learning and deepen users’ understanding of data [Kim et al., 2017]. Kim et al. [2018] showed that presenting users’ expectations against others’ expectations is also beneficial to the recall of data.

The results of these studies led to the creation of *TheyDrawIt!*, a publicly available authoring tool for eliciting people’s prior beliefs about the data in an interactive way [Collective, 2019]. *TheyDrawIt!* enables users to create line charts, sketch their beliefs about the trend in data, and view other users’ beliefs.

3.2.2 Evaluation of Uncertainty Visualizations

A multitude of experimental work on evaluation of uncertainty visualization exists in the literature. All these studies vary in the methodologies and evaluation strategies. Evaluating uncertainty visualizations involves many design-related decisions for the experiment like the evaluation goal, behavioral targets, expected effects, measures etc. Hullman et al. [2019] reviewed and analysed the evaluation decisions made in 86 studies of uncertainty visualization to define the more or less common decisions in the literature.

Most studies aimed at comparing multiple uncertainty visualizations or investigating the impact of uncertainty by evaluating at least one visualization that does not contain uncertainty. Aspects like how or why a visualization works or how it could interact with user characteristics were less often investigated. The greatest part of the user studies focused on the evaluation of the participants' performance and mostly by investigating effects on users' accuracy (i.e., the difference from a ground truth). Evaluation of user experience and effects on confidence, confidence/accuracy alignment, response time, decision quality, memorability etc. were not so commonly encountered.

In this subsection a relevant part of the existing literature on the evaluation of uncertainty visualization is presented: the evaluation of decision-making under uncertainty (does presenting uncertainty lead to better decisions than omitting it?); the evaluation of uncertainty visualization through comparisons of different designs, most of which presented in the previous subsection (are there particular designs of uncertainty visualizations that are more beneficial in specific tasks?); the evaluation of uncertainty visualization in Bayesian reasoning (is uncertainty visualization beneficial for people's performance in Bayesian reasoning tasks?).

3.2.2.1 Evaluation of Decision-Making Under Uncertainty

Uncertainty - No Uncertainty. The effect of communicating uncertainty in various decision-making contexts was evaluated in the existing literature.

One of these contexts was the communication of weather uncertainty information. Two studies evaluated people's decisions on taking precautionary measures for adverse weather conditions (salting the roads overnight). Nadav-Greenberg and Joslyn [2009] and Joslyn and LeClerc [2012] found that people made more optimal decisions when the uncertainty of the weather forecasts was communicated to them. Joslyn and LeClerc [2012] also found that displaying uncertainty increased people's trust in the forecast.

In transportation scenarios, Jung et al. [2015] argued that displaying the uncertainty of the remaining range in an electrical vehicle improved drivers' driving experience and behavior in regards to road and remaining range conditions in comparison to a single point estimate. Fernandes et al. [2018] found that some displays presenting uncertainty information about bus arrival times led to more accurate and consistent real-time transit decisions compared to displays with no uncertainty.

Conclusion. Communicating uncertainty helps people make more optimal decisions in comparison to point estimates according to the findings of empirical studies.

3.2.2.2 Evaluation of Uncertainty Visualization Through Comparison of Different Designs

Interval - Density - Gradient Plots. Uncertainty visualizations that display intervals have been found to be confusing or misinterpreted [Belia et al., 2005; Correll and Gleicher, 2014; Stock and Behrens, 1991]. Belia et al. [2005] found that interval plots like error bars are hard to interpret correctly for both naïve and expert users. [Stock and Behrens, 1991] found that viewers of Box plots underestimate the length of the whiskers for larger boxes and overestimate for smaller boxes. This was attributed to a visual perceptual illusion. Correll and Gleicher [2014] found that bar plots with error bars suffer from perceptual issues. The bar provides a false metaphor of containment; values within the bar are seen as likelier than those outside the bar. Viewers tend to adopt a binary interpretation in regards with the error bar; values are within the margins of error, or they are not. This leads to overestimated effect sizes.

Ibrekk and Morgan [1987] evaluated nine different representations of uncertainty including interval, density, or gradient plots (an error bar showing a certain confidence interval, discretised representations of the probability density function as a histogram and pie chart, a conventional probability density plot, violin plot, horizontal bar encoding density in the shading using either dots or vertical lines, a predecessor design of the modern Box plot (Tukey box), CDF plot) through a survey. Participants were asked to determine the location of the “best estimate” based on the presented visualization. No statistical differences were found but Cumulative Density Function (CDF) plots were found to be confusing when used to estimate the mean value.

Correll and Gleicher [2014] compared interval plots (i.e., bar plots with error bars and box plots) with density and gradient plots (i.e., violin and gradient plots) through a user study. The findings suggested that gradient and violin plots helped participants better understand uncertainty.

Conclusion. Overall density and gradient plots are preferred over interval plots according to the findings of the empirical studies. According to [Padilla et al., 2021a]: “More expressive visualizations provide a fuller picture of the data by depicting more properties, such as the nature of the distribution and outliers, which can be lost with intervals.”

Static Frequency-framed - Conventional Uncertainty Plots. The frequency-framed uncertainty visualizations were suggested in the literature on the basis that people were found to better understand and reason more accurately about uncertainty when this is framed as a frequency (e.g., 1 out of 10) instead of probability (e.g., 0.1) or percentile (e.g., 10%). This hypothesis was formulated and investigated by Gigerenzer [Gigerenzer, 1996; Gigerenzer and Hoffrage, 1995;

Hoffrage and Gigerenzer, 1998]. Gigerenzer argued that people reason more naturally about uncertainty when it is framed as frequency because this is how they experience probability in the world around them. Visualizations that implement frequency framing of uncertainty information have been suggested in the literature (e.g., icon arrays, quantile dotplots, HOPs). The evaluation work on the comparison of static frequency-framed uncertainty visualizations and other more conventional static designs is presented here.

Kay et al. [2016a] compared quantile dotplots with density plots in the context of realtime transit prediction scenarios. They found that the quantile dotplots reduced the variance of people's probabilistic estimates compared to density plots and facilitated more confident estimations.

Fernandes et al. [2018] compared displays with textual uncertainty, uncertainty visualizations (quantile dotplots, interval plots, density plots, and complementary CDF plots), or no-uncertainty in the ability of users' make optimal decisions in realtime transit scenarios. They found that low-density quantile dotplots and CDF plots led to better decisions and more accurate estimations of probability intervals.

Hullman et al. [2018] compared quantile dotplots with probability density plots when used in a graphical prediction interface for communicating uncertainty in experiment results. They found that quantile dotplots improved recall of a sampling distribution from a single experiment in comparison to the density plot.

Conclusion. Overall static frequency-framed plots are preferred over interval or density plots according to the findings of the empirical studies.

Staged Animated - Interpolated Animated - Static Transitions for Statistical Data Graphics. Heer and Robertson [2007] evaluated the effectiveness of *static*, *interpolated animated*, and *staged animated* transitions for common statistical data graphics (i.e., bar charts, pie charts, and scatter plots) in object tracking and value change estimation tasks. In *static* transitions the final chart replaces the initial one in a single step. In *interpolated animated* transitions the final chart replaces the initial one in a number of steps that show (usually linear) interpolations of the marks in the initial chart to its final state. This is a kind of direct animation [Heer and Robertson, 2007] and is different from *staged animated* transitions. In *staged animated* transitions the final chart results after a row of visual transitions in the position or value of the presented graphics (position and value might change synchronously or asynchronously).

Some of the transitions tested were "bar chart to donut chart (visualization change), stacked to grouped bars (drilldown), and sorting a bar chart (ordering)". The analysis of the collected data showed that participants' graphical perception was significantly improved when animated transitions were used in both types of tasks. Participants also reported strong preference for animation as they found it more helpful and engaging.

Kim et al. [2019] evaluated people's performance in identifying the aggregation performed over a distribution when animated transitions were used. They compared views with *staged animated* (the aggregated chart results after a row of visual transitions in the position or value of presented graphics), *static* (the aggregated chart replaces the unaggregated one) or *interpolated animated transitions* (interpolated transitions linearly interpolate marks in the unaggregated charts to their final values). Participants were asked to perform "binary identification tasks" to indicate whether a presented transition matches a provided aggregation. They found that the staged animated transitions improved participants' performance in identifying the aggregation performed and disambiguating the average and median, or stdev and iqr operations.

Conclusion. Overall animated transitions are more beneficial in comparison to static transitions according to the findings of the empirical studies.

Animated - Static Representation of Data Reliability in Cartography. Evans [1997] evaluated the use of animation in presenting the reliability of data shown in land cover maps. They compared a static composite bivariate map that presented only highly certain data with a "flickering" land cover map (animated alterations of the all-data and the high-certainty-data map). Both representations were found to be helpful according to participants' opinions.

Conclusion. Animated alterations of all-data and high-certainty-data maps were not found to be superior to the static high-certainty-data maps in cartography.

Animated Frequency-framed - Conventional Uncertainty Plots. The existing work in the evaluation of animated frequency-framed uncertainty visualizations in comparison to other more conventional static designs is discussed here.

Hullman et al. [2015] found that HOPs supported more accurate probability estimates than static alternatives (violin plots and error bars). Higher precision in participants' inferences came at a cost of time. Hullman et al. [2015] argued that HOPs required little statistical background to interpret as only simple mental processes were required to infer properties of the distributions.

Kale et al. [2019] compared the effectiveness of HOPs, error bars, and line ensembles (this is a form of discrete plot) as to users' ability in defining a trend in ambiguous data. They found that participants were able to correctly infer the underlying trend in presented data of lower level of evidence when using HOPs rather than static aggregate uncertainty visualizations.

Conclusion. Overall HOPs are preferred over interval, density, or discrete plots. According to Kale et al. [2019] the greater effectiveness of HOPs could be attributed to the fact that the sampling-oriented visualizations of uncertainty are in alignment with visual system's ability to "form gist representations from experience".

Interactive Graphical - No Prediction of Uncertainty or Trends in Data. Hullman et al. [2018] evaluated the effect of sketching one's predictions about uncertainty in the replications of an experiment prior to viewing the sampling distribution. They compared three conditions; in the first, participants were asked to graphically predict what would happen when an experiment was replicated; in the second, participants viewed the true sampling distribution for an experiment in a discrete representation; in the third, participants were asked to complete a rule training task about sampling distributions. Participants recall of the sampling distribution and accuracy in estimating the replication uncertainty in a transfer task were evaluated. The findings suggested that users were able to make better predictions about replications of new experiments when they graphically predicted the replication uncertainty in an experiment.

Kim et al. [2017] evaluated the effect of interactive graphical elicitation of users' prior beliefs about trends in data. Users' recall of the data was evaluated. Five conditions were investigated; no prior prediction and simple observation of data (control group); self-explanation of presented data typed in a text box; prior prediction of omitted data, comparative observation of prediction and actual data, and self-explanation of the gap between the prediction and the actual data; prior prediction of omitted data and comparative observation of prediction and actual data; prior prediction of omitted data, comparative observation of prediction and actual data, and observation of textual and visual annotation of the gap between the prediction and the data. For each one of these conditions there was one of two types of data representation; text and visualization. Participants who predicted the trend in data graphically and were prompted to self-explain data outperformed the control group in recall and comprehension. The effects were observed when visualization was used instead of text and persisted for participants with moderate or little prior knowledge on the datasets.

A relevant user study was conducted by Kim et al. [2018]. Although this study did not compare interactive and static conditions, it is interesting because it adds value to the findings of Kim et al. [2017]. Kim et al. [2018] used a similar interactive graphical elicitation interface and evaluated the effect of showing participants predictions against other users' predictions. Participants recall of the data was evaluated. Participants recalled the data more accurately when other users' predictions had high degree of consensus in comparison to participants who only viewed the data. Participants showed less trust in the validity of data and were more likely to maintain their initial expectations when other users' expectations aligned with their own initial expectations but not with the data.

Conclusion. Interactive graphical prediction of the replication uncertainty in an experiment help users make better predictions about the replications of new experiments. Also interactive graphical prediction of trends in data improves viewers' recall of the data when their predictions are presented against the actual data and the predictions' of other people who present high consensus.

3.2.2.3 Evaluation of Uncertainty Visualization in Bayesian Reasoning

Existing work in human Bayesian reasoning found that people perform poorly in Bayesian reasoning tasks, where they need to calculate the conditional probability of a hypothesis by updating their prior beliefs in the light of data (apply Bayes' rule). It was found that people's difficulty to perform well in Bayesian reasoning tasks is attributed to various cognitive biases [Cole, 1989; Gigerenzer and Hoffrage, 1995; Tversky and Kahneman, 1974] like the base rate neglect [Tversky and Kahneman, 1974]. Many research efforts in the literature focused on suggesting problem descriptions including visualization for debiasing people's Bayesian judgments and improving their performance. The evaluation work conducted in this direction is reviewed in this subsection.

Graphical Displays - Textual Descriptions. People's performance in Bayesian reasoning seemed to have been benefited when graphical displays (contingency tables, signal detection bar, detection bar, probability map or double-tree diagram) [Cole, 1989; Khan et al., 2018] or iconic pictorial representations [Brase, 2009] or interactive frequency grids with check boxes [Tsai et al., 2011] were combined with a textual description of the Bayesian reasoning problem. Expanding the sample through crowd-sourcing [Micallef et al., 2012; Ottley et al., 2012] led to inconsistent findings with previous work possibly because the wording of textual descriptions could significantly impact users' accuracy [Ottley et al., 2016]. Ottley et al. [2016] showed that (text-only or) visualization-only designs were more effective than those which blend text and visualization.

Conclusion. Text-only or visualization-only designs seem to be more beneficial for people's Bayesian reasoning.

Interactive - Static Representations. There are few studies having investigated the effect of interaction on users' performance in Bayesian reasoning, and the findings were unexpected.

Mosca et al. [2021] found no improvement in people's Bayesian reasoning by adding interaction (checkboxes, drag and drop, hover, tooltips) to static icon arrays to show or hide subpopulations or link the textual description to the visualization. Khan et al. [2018] found that interactively constructing the double tree diagrams [Khan et al., 2015] through dragging-and-dropping led to worse performances in the Bayesian reasoning tasks. Khan et al. [2018] suggested that people's worse performance when using interaction might result from the cognitive overload caused to them by interacting. Khan et al. [2018] argues that these findings arise the need for future research in the effects of interaction on cognitive load.

Conclusion. The existing work about the added value of interaction on static visualizations is little and thus, conclusions about this contradiction cannot be easily drawn.

3.2.3 Challenges in Designing and Evaluating Uncertainty Visualizations

Challenges in Designing Uncertainty Visualizations. Although the presented literature suggests that communication of uncertainty helps people make better and more informed decisions, it may also confuse them if the design of the uncertainty visualization does not account for 1) the needs and 2) the ways that people naturally reason about probability. These two important parameters should be considered in the process of designing uncertainty visualizations.

Regarding the first parameter, Greis et al. [2017] emphasized the importance of developing general design guidelines as to how to design for and with uncertainty in HCI. The lack of general design guidelines is a serious problem often leading designers to omit uncertainty. In many instances though the application may set design specifications for the representation of uncertainty. For example, in the case of designing visualizations for a transit prediction mobile application by Fernandes et al. [2018]; Kay et al. [2016b], users had to make quick, in the moment-decisions. The representations of uncertainty had to be glanceable and compact for the mobile phone display. That was a helpful cue for the designer to create appropriate displays of uncertainty.

Regarding the second parameter, the ways that people reason about uncertainty depend on various factors. The statistical literacy of viewers of uncertainty visualization is one such factor. Many typical uncertainty representations like error bars are misunderstood or misinterpreted especially by non-experts or researchers unfamiliar with statistical concepts like confidence intervals and standard error [Belia et al., 2005]. Another factor is cognitive biases in people's judgments under uncertainty because of the reliance on judgmental heuristics [Tversky and Kahneman, 1974]. According to [Hullman, 2016], people often employ

a form of intuition that provides a mental shortcut for hard decisions. Most heuristics work by substituting a simple but less accurate representation to turn a difficult decision about a situation with multiple parameters and uncertainties into an easier one.

The ways that people more naturally reason about uncertainty is another factor. For example, people more accurately understand frequency formats than probabilities [Gigerenzer, 1996; Gigerenzer and Hoffrage, 1995; Hoffrage and Gigerenzer, 1998].

Challenges in Evaluating Uncertainty Visualizations. The process of designing the evaluation of uncertainty visualizations needs attention. Hullman [2016] argued that this process is usually error-prone and categorizes the factors she believes influence the effect of evaluating uncertainty visualizations into three categories:

- the nature or definition of probability and subjective probability distributions (e.g., should participants' subjective uncertainty be compared to statistical uncertainty?);

- the sensitivity of participants' responses to the elicitation method (e.g., response modes as numeric responses, decisions between alternatives, Likert scale ratings, etc.);
- the use of heuristics in the judgements of probability by users who try to make their responses resemble normative ones.

She also suggested some good practices to be followed in the design of the evaluation routine.

Hullman et al. [2019] presented a taxonomy of six levels of decisions that should be made in the designing of an uncertainty visualization evaluation experiment: the behavioral targets of the study, expected effects from an uncertainty visualization, evaluation goals, measures, elicitation techniques, and analysis approaches. They reviewed existing work in the evaluation of uncertainty visualization and analysed the design decisions made for the evaluation of such visualizations. Based on this analysis they suggested a set of recommendations designed “to encourage more transparent evaluations aligned with the state of the art in knowledge on uncertainty comprehension”.

3.3 Visualization of Models' Structure

Uncertainty is inherent in probabilistic models (found in the prior distributions of the models' parameters) and their outputs (found in the posterior distributions of the models' parameters). Nevertheless, communicating only the uncertainty of parameters or predictions is often not enough. In tasks like model refinement, parameter tuning, or decisions of interventions on a variable a good comprehension of the dependencies of the model's variables in the model is required. For this reason the *structure* of the model should be communicated to users along with the uncertainty.

A common way to represent the structure of a model visually is through *directed graphs*. This section presents the existing work in presenting the relations of variables in a model graphically. Section 3.3.1 presents the Bayesian network, a “skeleton” for representing a probabilistic model's joint distribution in a factorized way [Koller and Friedman, 2009], and the available tools for generating such representations. Section 3.3.2 presents the causal diagram, a graphical representation of the causal relations among the variables in a model. It presents the existing tools suggested in literature for visualizing graphically the causal relations in datasets recovered by causal discovery algorithms. Finally, it reviews the evaluation work in the effects of visualization in causal reasoning.

3.3.1 Probabilistic Models

3.3.1.1 Bayesian Network

A common visual representation of the structure of a probabilistic model is the Bayesian network [Koller and Friedman, 2009] (see Fig. 3.1(a) which replicates Fig. 1.4 shown in Chapter 1). A Bayesian network comprises a Directed Acyclic Graph (DAG) with nodes corresponding to random variables and edges indicating the direction of the “influence” of one node on another. The Bayesian network apart from presenting an at-a-glance overview of the model’s variables and their dependencies, provides a framework for compactly representing the model’s joint distribution by factoring out independencies.

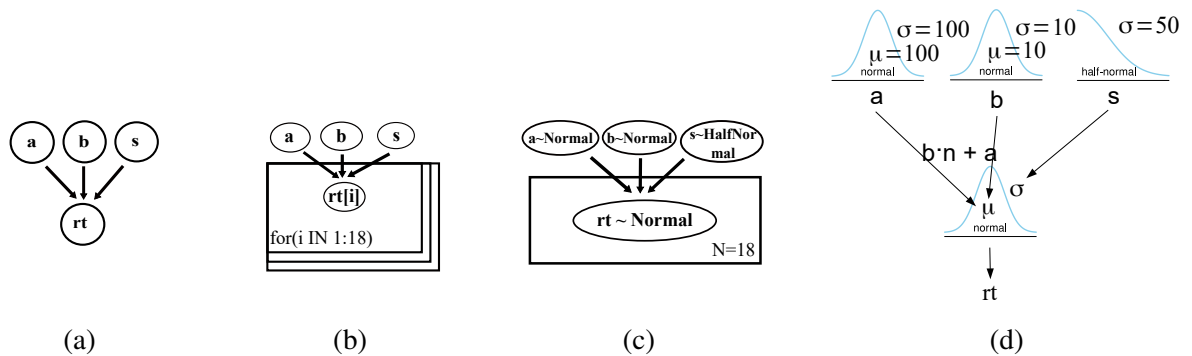


Figure 3.1: Various graphical representations of the homogeneous probabilistic model of the drivers’ reaction time problem (Box 2.1). (a) Bayesian network, (b) DoodleBUGs’ graph, (c) PyMC’s graph, and (d) Kruschke-style diagram.

The dimensionality of the joint distribution of a probabilistic model depends on the number of its parameters (i.e., it increases with the complexity of the model) (explained in Section 2.2.4.1). Expressing a multi-dimensional joint distribution in terms of the simpler marginal distributions of the model’s parameters is straight-forward when the model’s parameters are independent (the joint distribution equals the product of the marginal distributions of the parameters) but becomes challenging when the parameters present dependencies like in hierarchical models.

A Bayesian network encodes a set of conditional dependence and independence assumptions among model’s random variables. The conditional dependencies are represented as edges, while the conditional independencies as missing edges in the DAG. Bayesian networks satisfy the local Markov property, which states that a node is conditionally independent of its non-descendants given its parents. As Koller and Friedman [2009] very precisely state in their book, there are two equivalent ways that a Bayesian network can be viewed; “a skeleton for representing a joint distribution compactly in a factorized way”, and “a compact representation for a set of conditional independence assumptions about a distribution”. Box 3.1 demonstrates how the Bayesian network of a hierarchical model can be interpreted into a set of conditional independence assumptions and used to factorize the joint distribution of the model.

The Bayesian network is a specific instance of a probabilistic graphical model representing the model's joint distribution. There may be multiple valid probabilistic graphical models for a joint distribution because there may be more than one ways to factorize it. For example, any joint distribution of two variables (x , y) that are not independent may be represented by both $x \rightarrow y$ or $x \leftarrow y$ [Lattimore and Rohde, 2019a]. The direction of the arrows in probabilistic graphical models do not necessarily indicate the order of the operations in the modelled data generating process (e.g., the order of operations as determined by the definition of the model e.g., in probabilistic statements) but in Bayesian networks they do. Thus, Bayesian networks provide a natural representation of the model's joint distribution.

Box 3.1 Bayesian network of the hierarchical model of the drivers' reaction time problem

Fig. 3.2 presents the Bayesian network of the hierarchical model of the drivers' reaction problem defined in Box 2.3 in Chapter 2.

Conditional dependencies among variables in Bayesian networks are represented by the *directed* edges. For example, the (μ_b, b_i) edge expresses the conditional probability distribution $p(b_i|\mu_b)$. Bayesian networks satisfy the *local Markov property*: a node is conditionally independent of its non-descendants given its parents. For example, the conditional probability of node rt_i given its parent nodes a_i, b_i, s_i is conditionally independent of the hyperparameters' nodes, $\mu_a, \sigma_a, \mu_b, \sigma_b, \sigma_s$:

$$p(rt_i|a_i, b_i, s_i, \mu_a, \sigma_a, \mu_b, \sigma_b, \sigma_s) = p(rt_i|a_i, b_i, s_i). \quad (3.1)$$

The joint distribution of the model can be expressed as a product of the variables' conditional probability distributions based on the chain rule of probability, the local Markov property, and the independence assumptions of the model's variables (see Box 2.3 for the assumptions):

$$\begin{aligned} p(rt_i, a_i, b_i, s_i, \mu_a, \sigma_a, \mu_b, \sigma_b, \sigma_s) = \\ p(rt_i|a_i) \cdot p(rt_i|b_i) \cdot p(rt_i|s_i) \cdot p(a_i|\mu_a) \cdot p(a_i|\sigma_a) \cdot p(b_i|\mu_b) \cdot p(b_i|\sigma_b) \cdot p(s_i|\sigma_s) \cdot \\ p(\mu_a) \cdot p(\sigma_a) \cdot p(\mu_b) \cdot p(\sigma_b) \cdot p(\sigma_s) \end{aligned} \quad (3.2)$$

Equation 3.2 shows how a Bayesian network can provide a factorized representation of the model's joint probability distribution such that each factor in the factorized representation is represented by an edge in the DAG.

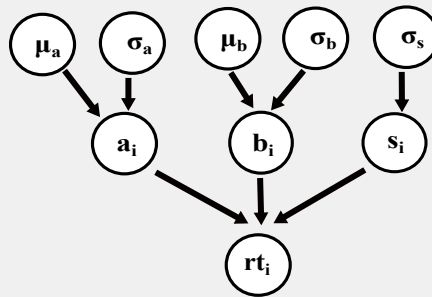


Figure 3.2: Bayesian network of the hierarchical model of the drivers' reaction time problem defined in Box 2.3.

3.3.1.2 Existing Tools for Graphical Representation of Probabilistic Models

The existing tools that can generate a graphical representation of a probabilistic model in the spirit of Bayesian networks are reviewed in this section. There are few tools that could automatically transform a specification of a probabilistic model into a graphical representation. There

are two PPLs that provide some form of graphical model interface; BUGS (via DoodleBUGS) and PyMC3.

DoodleBUGS is a software component of WinBUGs [Spiegelhalter et al., 2003] that provides a Doodle editor for creating probabilistic graphical models as DAGs and automatically transcribing DAGs into BUGs language [Lunn et al., 2009]. However it cannot do the opposite, namely transform a model written in BUGS into a DAG. Fig. 3.1(b) presents the graphical probabilistic model of the drivers’ reaction time homogeneous model that was created with DoodleBUGS. PyMC [Salvatier et al., 2016] provides a method that converts a probabilistic model expressed in PyMC code into a graphviz Digraph (`pymc.model_graph.model_to_graphviz`) using the Graphviz graph visualization software [Ellson et al., 2004]. Fig. 3.1(c) presents the graphical model of the drivers’ reaction time homogeneous model that was generated through PyMC’s interface.

Kruschke [2015] introduced a more informative DAG that shows iconic “prototypes” of the distributions on each node of the diagram. Kruschke [2012b] argues that this type of diagram (in comparison to the ones created with DoodleBUGS) has “a much more direct correspondence to lines of code in JAGS/BUGS: (Usually) each arrow in this diagram corresponds to a line of code in the JAGS/BUGS model specification.” Kruschke [2012b] explains that this type of diagram indicates which parameters participate in the same distribution, which is not visible in DoodleBUGS graphs. There is no automatic tool for the creation of this Kruschke-style diagram, but Kruschke [2013] presents two drawing tools that were created specifically for the creation of this type of diagram; a set of distribution and connector templates in LibreOffice Draw and R; and LaTeX/TikZ scripts. Fig. 3.1(d) presents the Kruschke-style diagram of the drivers’ reaction time homogeneous model created with the LibreOffice Draw template.

3.3.2 Causal Models

3.3.2.1 Causal Diagrams

A probabilistic graphical model in which a link $x \rightarrow y$ is assumed to mean x causes y [Lattimore and Rohde, 2019a] represents a *causal model*. Causal models are “mathematical models representing causal relationships within an individual system or population” [Stanford Encyclopedia of Philosophy, 2018]. The probabilistic graphical models that represent causal models are called *causal diagrams*.

Causal diagrams were first introduced by Wright [1921]. Causal diagrams are used for communicating causal models and facilitating causal reasoning (i.e., reasoning about what variable affects what other variable). They are also used for causal inference, namely for quantifying the causal effects among variables [Greenland et al., 1999; Pearl, 1995; Pearl and Mackenzie, 2018].

In this thesis the focus is more on how causal diagrams can be used as visual aids for causal

reasoning. The relevant existing work is reviewed in this section. First, existing visualization tools in literature that present graphically learned causal structures by automatic causal discovery algorithms are presented. Then, a review of relevant work in the evaluation of visualization in facilitating people's causal reasoning follows.

3.3.2.2 Existing Visualization Tools for Causal Reasoning and Exploration

Various algorithms have emerged recently that can learn the causal structure from observational data, as well as experimental data [Glymour et al., 2019; Malinsky and Danks, 2018]. The learned causal structure informs about whether the variables in a dataset have a cause-effect relation and sometimes even about the direction of this relation (which variable is the cause, and which one is the effect). Various visualization tools have been suggested in the existing literature for the visualization, exploration, and validation of these learned causal structures.

Dang et al. [2015] presented the “ReactionFlow”, an interactive visualization tool for causality analysis amongst proteins, complexes, and biochemical reactions in biological pathways. That tool enabled users to interactively filter, cluster, and select pathway components across linked views, or use animation to view the flow of activity through a pathway. Xie et al. [2020] presented the “Causality Explorer”, a visual analytics tool for validating causal relations and exploring causal effects through simulations of interventions and stratification of variables. Ge et al. [2020] presented the first web-based causal discovery tool that presents the graphical models of the learned causal models and provides interactive features for the exploration and annotation of the causal relations among the variables.

There are cases when automatic causal discovery algorithms are not able to determine the direction of the causal relation between variables (discriminate between a cause and effect) solely based on the observed data. In these cases, causal assumptions inferred through the analyst's or expert's experience and expertise could be valuable. Existing literature presents some notable causality elicitation and visualization tools that can bring the analyst in a closed-loop with the causal analysis tool.

Wang and Mueller [2015] presented the “Visual Causality Analyst”, an interactive visual tool that enables domain experts to verify and edit the causal links on the estimated causal model and explore causal effects by controlling for variables through stratification and conditioning or regression models. Wang and Mueller [2017] extended this work by developing more features which were integrated into a new tool called the “Causal Structure Investigator”. The new visualization tool had many advantages in comparison to the previous tool; among others, it presented the causal relations in path diagrams laid out using spanning trees to better expose the flow of causal dependencies, used a scoring function and corresponding visual hints to compare alternative causal models, and provided interactive facilities for the exploration of data subdivisions, which might imply different associations among variables depending on how the data is subdivided (this is described as the Simpson's Paradox [Simpson, 1951]).

3.3.2.3 Evaluation of People's Causal Reasoning with Visualization

Existing visual analytics tools aim at facilitating the exploration of causality in data through the use of visual means (e.g., graphs, interaction, annotations, bar/pie charts etc.) but there is very little known about how well people can infer the causal structure of data through visualization. The existing work is still quite poor in this field.

Kale et al. [2022] evaluate the quality of people's causal reasoning based on visualizations. They assessed the ability of users to detect a treatment effect and a confounding relationship in visualized count data. They tested their hypotheses for non-interactive visualizations (text contingency tables, faceted icon arrays, and faceted bar charts) and two interactive designs; aggregating bars with similar design to the faceted bar charts whose faceting could be interactively toggled, and cross-filtering bars, which were linked bar charts that could be interactively cross-filtered by clicking on the bars. The analysis of the collected data did not show any reliable improvement in participants' performance when they used the common visualizations in comparison to textual contingency tables.

Yen et al. [2019] evaluated people's performance, strategies, and pitfalls in identifying mediators in a dataset using bar charts. They compared an interface presenting data in bar charts with a similar design that included a component allowing an interactive graphical representation of the variables' causal relations. They found that participants' performance in identifying the mediator significantly decreased when a confounding variable influenced the variable being analyzed. Participants' individual visualization exploration strategies and the design of the interface seemed to have influenced participants' reasoning performance.

Lagnado and Sloman [2004] evaluated the ability of people to identify a causal chain model (a three-variable causal model that included a mediator) when simulated data was presented to them. The participants were either presented with observational data or were able to interactively intervene on one of the variables and set its value and then view the values of the other two variables. They found that interveners' performance was better than observers' having ruled out as possible explanation for this any informational differences between the two conditions. The experimental evidence confirmed that the advantage of interveners was driven by a temporal signal: "interveners exploited the cue that their interventions were the most likely causes of any subsequent changes".

3.4 Inspiring Ideas From the Literature on Interactive Visualization

Exploiting interaction for communicating complex ideas, methods, and results in science, research, or education is well-established in the research literature. Various interactive visualization techniques have been developed to either reveal dimensions of the information that cannot

be conveyed through static visualizations or to enhance comprehension of the presented information.

For example, there is an extensive literature on interactive projection of high dimensional data [Faith, 2007; Sankaran and Holmes, 2018]. These occur in Bayesian probabilistic models with many parameters in the joint distribution. Some studies focused on the importance of interaction for comprehension and engagement [Hullman et al., 2018; Kim et al., 2017, 2018; Tsai et al., 2011]. A novel interactive reporting method for scientific and research results brings interaction in the center of attention for statistical reporting. Dragicevic et al. [2019] presented a new approach to statistical reporting where readers of research papers can explore alternative analysis options by interacting with the paper. The “explorable multiverse analysis reports” allow authors to report the outcomes of many different statistical analyses and readers to dynamically change some elements of the analysis and get new “versions” of the paper based on the new produced results.

The “explorable multiverse analysis reports” rely on two key concepts. The first is the “multiverse analysis” in statistical reporting [Steege et al., 2016], where all processed data sets that are generated from raw data based on different choices of processing are analysed to produce a *multiverse* of statistical results. This multiverse reveals the fragility of the results across various processing options. The second is the idea of “explorable explanations” [Victor, 2011a], which aims at encouraging active reading through active engagement of the readers with a new form of interactive narratives (e.g. reactive documents, explorable examples, contextual information) that could allow readers dynamically change some elements and get a new “version” of the narrative.

Victor has also expounded how interaction, simulation and visualization could be used for simplifying abstract ideas to provide an intuitive understanding. For example, he suggests the creation of a high-level mathematical tool that could become “as ubiquitous as the pocket calculator”, which would transcribe mathematical problems into software simulations of simple physical models instead of abstract equations and symbols [Victor, 2009]. He argues that this kind of software could introduce a new form of practical mathematics that could “provide a broader context, allowing a deeper understanding of the problem; easily handle problems which are difficult or impossible to solve analytically; and be used to actively create, not just passively understand”. An illustrative example was the scrubbing calculator [Victor, 2011c] that interactively explores parameter spaces of algebraic problems by scrubbing over numbers until the desired result is reached.

Finally, Victor [2011b] highlights the importance of a “ladder of abstraction”. By moving between levels of abstraction starting from the lower one that indicates a concrete working system and stepping down to the higher one that indicates abstract equations or aggregate statistics, the system designers’ intuition and their design develop “side-by-side”. Victor [2011b] argues that interaction in this iterative process of exploring a system design is an essential element to

move around the “ladder of abstraction”.

Chapter 4

Using Interaction for Visualizing Probabilistic Programming Models

4.1 Summary

This chapter presents the first of the three parts of this research. The focus of this part is on the design and implementation of a novel interactive graphical representation of a probabilistic model's structure at varying levels of granularity, with seamless integration of uncertainty visualization. The probabilistic model is assumed to be expressed in any PPL and the intention is to achieve an automatic transformation of the probabilistic model expressed in a PPL into this interactive graphical representation. This interactive graphical representation supports the exploration of the prior and posterior MCMC sample space.

Section 4.2 discusses the purpose of this work in more detail, Section 4.3 presents the typical tasks that users of Bayesian probabilistic models need to undertake and could be supported by the proposed tool, Section 4.4 presents the main aspects of the design and implementation of the proposed tool, Section 4.5 provides a comparative presentation of the proposed tool and other existing relevant tools, Section 4.6 demonstrates illustrative examples of use, and finally, Section 4.7 discusses the contributions and limitations of this work and the potential future endeavours in the field.

4.2 Purpose

Bayesian probabilistic modeling has many advantages; it accounts for uncertainty, incorporates prior expert knowledge, has a well-defined intrinsic structure in terms of *relations* among random variables: the mathematical and statistical dependencies are explicitly stated. Extremely flexible Bayesian probabilistic models can be implemented via PPLs, which provide automatic inference via efficient MCMC sampling. Nevertheless, Bayesian probabilistic models' structure and inference results can be challenging to communicate as the model becomes more complex,

perhaps with hierarchical structure, multivariate distributions, complex inter-dependencies and increasingly abstract latent states. Communicating uncertainty has challenges, too. Bayesian reasoning is closely tied to reasoning about conditional probabilities. People with a weaker background in statistics can have difficulty reasoning about (conditional) probabilities [Díaz and Inmaculada, 2007; Gigerenzer and Hoffrage, 1995; Kahneman and Tversky, 1974; Koller and Friedman, 2009; Tversky and Kahneman, 1973, 1974]. For example, people find it difficult to distinguish conditional and joint probabilities, and recognize that conditional probability involves a restriction in the sample space [Díaz and Inmaculada, 2007]. But even in cases that people are fully aware of these issues in theory, it is practically difficult to reason about the conditional probabilities of a complex model.

There are visualization tools that seek to communicate the structure of a probabilistic model or its inference results to users in a compact and relevant way. For example, ArviZ [Kumar et al., 2019] is a unified library in Python that provides tools for diagnostics and visualizations of Bayesian inference for various PPLs. A very simple probabilistic model with few parameters could allow a human user to contemplate the entire model at once and comprehend how parameters interact with each other and the predictions of the model. However, a complex Bayesian model could result in a high-dimensional posterior distribution that would require unwieldy tables to present summary statistics or a multitude of uncertainty visualizations, one per marginal distribution, that are difficult to grapple with (as explained in Section 1.3.1). Also the more complex a Bayesian probabilistic model becomes, the more error-prone the specification and validation process of a Bayesian probabilistic model becomes given the existing ways to represent probabilistic models' structure and inference results.

Another difficulty users usually encounter is that most existing visualization tools for Bayesian analysis are PPL-dependent as they require the model- or inference-related information to be provided in formats (structures) compatible with the specific PPL they were built for. This might not allow the exploitation of the available representation possibilities of all existing visualization tools by the users of a specific PPL or might require users to switch from one PPL to another to be able to benefit from a certain visualization offered for a specific PPL backend.

There is need for tools that would automatically synthesize user interfaces to PPL inference results independently of the PPL used for the inference, creating a compact graphical representation that would convey model- and inference-related information in conjunction. These tools could exploit interactivity to vary the granularity of the presented information and facilitate the exploration of complex probabilistic models' structures and inference results. Such tools could replace large tables of statistics with interactive graphical representations which would integrate the structural relation of parameters along with their inferred distributions to convey uncertainty accurately. Such tools could also support interactive sensitivity analysis of the model's parameters.

In this chapter, one realization – design and implementation – of such a tool is presented; the

interactive probabilistic models explorer (IPME). The objectives of its design are:

1. automatic synthesis of a graphical representation of a model independently of the PPL used;
2. seamless integration of uncertainty visualization into the graphical representation of the model;
3. inclusion of both prior and posterior distributions of the model's parameters, and the corresponding predictive ones for the observed variables;
4. interactive exploration of inference MCMC sample space;
5. interactive sensitivity analysis of model's parameters;
6. granularity in the presented visual information according to user's choices and needs;
7. inclusion of predictive checks.

These objectives are set to support a series of probabilistic models-related tasks that might not be well supported by existing tools. Real users like model builders, decision-makers, or researchers need to rely on probabilistic models to conduct Bayesian analysis or interpret the results of Bayesian inference. These users need to be supported in their tasks. This chapter identifies and describes a number of cognitive and practical probabilistic models-related tasks that people dealing with probabilistic models and Bayesian inference usually need to undertake. Some of these tasks require new IPME-like tools to be supported. A comparison of existing visualization tools with the new suggested tool is presented in this chapter to emphasize the identified gap in the availability of such tools. Concrete use case scenarios demonstrating the use of IPME in a variety of scenarios (and probabilistic models) to support such tasks are also presented in this chapter.

4.3 Cognitive and Practical Tasks that Users of Bayesian Probabilistic Models Undertake

The purpose of this section is to present the main cognitive and practical tasks that most users (i.e., modelers, decision-makers, or researchers) are usually required to undertake when they have to deal with probabilistic models and Bayesian inference. Some of these tasks might be supported by existing visualization tools, but possibly not efficiently or effectively, while others are not supported at all. These tasks informed the design of IPME, which will be presented in the next section. The identified tasks are categorized in four categories, each indicating a different user intention.

Distributional Information Comprehension. The comprehension of the distributional information contained in the prior and posterior joint distribution of a Bayesian probabilistic model is critical for any relevant task, from building and validating the model to interpreting the inference results and making decisions based on them. The prior and posterior distributions of a Bayesian probabilistic model are multi-dimensional joint distributions (as explained in Section 2.2.4.1). The marginal distributions of the prior and posterior distributions of the model are slices which reveal the uncertainty of a subset of (usually one of) the model’s parameters. PPLs’ typical output consists of a *trace*, a set of samples from the posterior distribution resulting from inference with a MCMC sampler; that’s, the output from “running” the model (as explained in Section 2.3.3). PPLs can also produce a set of independent samples drawn from the prior joint distribution of the model. The marginal (prior or posterior) distribution of each parameter in a model can be easily estimated from these (MCMC) samples.

Fig. 4.1(b) in Box 4.1 presents the posterior MCMC samples of a two-parameter model and Fig. 4.1(c) shows the estimated 3D posterior distribution along with the 2D marginal distributions of the parameters (in dark blue). The distributional information of the prior or posterior joint distribution of a model is usually communicated in the form of summary statistics or uncertainty visualizations (e.g. KDE plot, error bar, Box plot, CDF plot, quantile dotplot) of the marginal distributions of the model’s variables. The complexity of the model (i.e., number of parameters and number of variables’ indexing dimensions) determines the number of the uncertainty visualizations or rows in the tables of summary statistics. Thus, as the complexity of the model increases, the distributional information of the model’s variables becomes harder to communicate, explore, and ultimately, comprehend. Some representation that would allow granularity of the presented information and some mechanism of requesting the information of interest from the outputs of probabilistic models expressed in a PPL would be really useful.

The density of the joint samples of the model’s prior or posterior usually is not uniformly distributed in the corresponding sample space. Understanding how the distribution of this density might vary given that a parameter or a subset of parameters take a specific value or values in a specific subrange is useful when users need to estimate conditional probability densities of the form $p(Y|X = x)$, where X and Y are two random variables and it is observed that variable X takes the value x . For example, this could be particularly useful for decision-makers who would like to have estimates of a parameter’s uncertainty under some specific conditions, which could represent a worst case scenario.

Estimating conditional probabilities from the output of a PPL leads to some form of *querying* the results of the MCMC process; for example, a conjunctive restriction like $(1.6 < \mu < 2.0)$ AND $(1.0 < \sigma < 1.4)$ for the mean and standard deviation of the average minimum temperature of the example in Fig. 4.1. The sample space of the (prior) posterior distribution can be restricted by setting bounds to the range of values of the individual parameters. Each restriction on the value range of a parameter defines a slice on the distribution. All the MCMC samples that

belong to the subset of the restricted sample space determine the distribution of the model within this subset of the sample space. In a conjunctive query, if we restrict the value range of more than one parameters, then the resulted subset of the sample space of the (prior) posterior joint distribution is defined by the intersection of all the restricted value ranges of the parameters. For example, Fig. 4.1(b) presents the MCMC samples that lie in the intersection of two value ranges restrictions in orange and Fig. 4.1(c) presents the re-estimated posterior distribution within the restricted sample space.

Such queries could be specified and reported visually if the user could interactively set value ranges and get the updated uncertainty visualizations for the *entire* model's joint distribution within the defined subset of the sample space. At the moment, it does not seem that tasks like these are supported by existing tools and libraries for representing and visualizing the outputs of Bayesian inference.

Box 4.1 Average Minimum Temperature in Scotland

This example presents a two-parameter Bayesian model modeling the average minimum temperature in Scotland for the month November. The average minimum temperature in Scotland in month November for the years 1884-2020 is available in a dataset. PyMC3 is used for the inference (see Appendix A.1 for definition of model in PyMC3 and link to the dataset). The inference results are explored by setting two value range restrictions; $\mu \in [1.6, 2.0]$ and $\sigma \in [1.0, 1.4]$. The posterior marginal distribution of μ (in orange) in the subset of the posterior space became slightly tighter and shifted towards lower values of μ , which lead us to less uncertainty about expecting lower average temperature in Scotland for November given this particular conditioning.

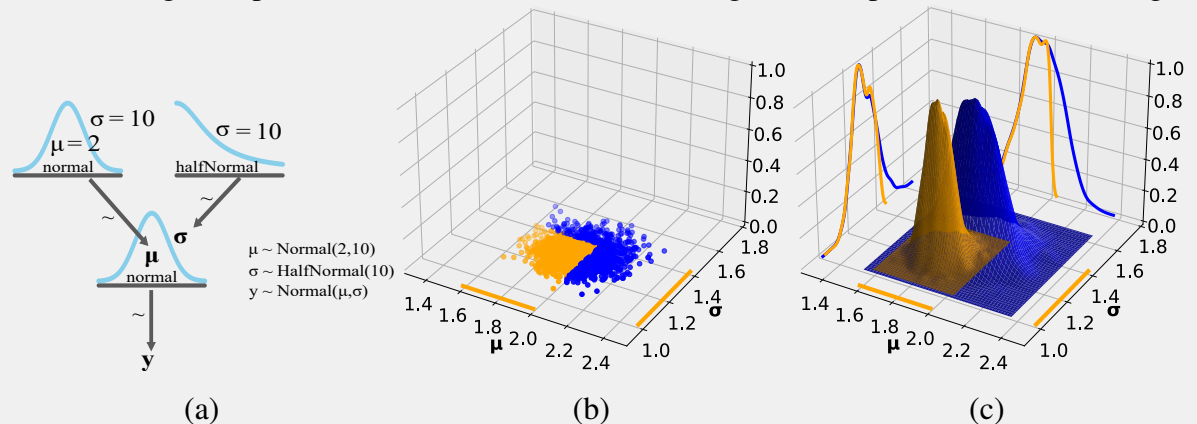


Figure 4.1: (a) Kruschke-style diagram of the two-parameter Bayesian model. (b) Posterior MCMC samples of the model's parameters. The samples that lie within the intersection of the value range restrictions are colored in orange. (c) The estimated posterior distributions and marginal distributions based on the entire posterior MCMC sample set (dark blue) and the restricted sample set (orange).

Inference Results Comprehension. Comprehending simply the distributional information of the posterior distribution might not suffice to achieve comprehension of the inference results.

For example, some users might want to investigate the sensitivity of inference results to the prior distributions chosen. This task would require a comparison between the prior and posterior distribution; the users would need to observe both in parallel and in a direct and explicit way. This kind of comparison between the prior and posterior distribution could also help users to get a profounder understanding of how their prior knowledge changes in the light of the data: e.g., does the posterior distribution become wider or shift in the posterior space given the actual observations? (for example see example in Box 1.6 in Chapter 1). These tasks could be facilitated by presenting the uncertainty visualizations or tables of summary statistics of the distributions of model's variables side-by-side. The complexity of the model could be a hurdle, though, in this case, as well.

Model Comprehension. In many tasks like decision-making in crucial areas like healthcare or stock market investments, going a step further and acquiring a profounder understanding of the structure of the model might be critical. Understanding whether and how one parameter might affect another in a probabilistic model could enable decision-makers to account for any possible risk entailed by the produced predictions of such models. Aspects of the structure of such models that are usually useful to be well-understood are the parameters' associations (e.g., which parameters are used to set the parameters of the distribution of another variable; that's, the parent-child relations of the variables in a model), the hierarchies (e.g., which parameters are hyper-parameters and how deep in the model's hierarchy they lie), and the dimensions of the parameters (e.g., which parameters are global or model groups of samples).

The parent-child relations of variables in a probabilistic model, and the hierarchy and dimension of the parameters are glanceable in the definitions of the models in probabilistic statements or any existing graphical representation of the model (in the form of a DAG). Nevertheless, understanding the effect that one parameter might have on the distribution of another parameter in a probabilistic model (e.g., how does the distribution of one parameter change if another parameter takes a certain value? how sensitive is one parameter to changes in the value of another parameter?) given these means of communicating the model's structure relies on the ability of the user to interpret the mathematical details of the model's definition.

Existing visualization libraries or tools do not seem to offer an alternative representation that could facilitate the communication of the effects one parameter has on others within the context of a probabilistic model in ways to alleviate users' potential inability to deal with statistical/mathematical details. One way to explore the sensitivity of parameters in the model could be setting range restrictions on one parameter and observe how the uncertainty visualization of others is affected. By restricting the value range of one or more parameters the influence on the remaining parameters can be observed, because only the MCMC samples of the prior or posterior joint distribution that satisfy the constraints are included. For example, setting parameter μ in increasing value ranges and observing how the posterior predictive distribution of y changes in the probabilistic model shown in Fig. 4.1(a) could help people interpret the effect of param-

eter μ on the mean value of the distribution of y . Through such exploration the parameters that are strongly coupled or, conversely, are wholly independent could be identified. This could be critical for decision-makers, who in the process of assessing the risk, would like to know how fragile the estimation of crucial parameters is.

Model Check. A common task especially for model builders is checking and validating the model. For example, one type of model validation is checking whether the prior (predictive) distributions capture effectively the prior knowledge (e.g., expected value ranges for parameters). Another type of model validation is checking whether the predictions of the model capture sufficiently aspects of the observed data that are of interest. These kind of checks are called *predictive checks* [Lambert, 2018a, Chap. 10.3; Sinharay and Stern, 2003]. Through predictive checks statistics of the observed data are tested against the predictions of the model. The aspects of the observed data that are usually investigated could be defined as statistical metrics over the data. For example, the extreme observed values could be interpreted as the `min` and `max` value of the observations. Other common aspects of the data that could be checked are the `mean` and `std` (standard deviation) values.

A common metric for checking how well these statistics of the observed data are represented by the predictions of the model is the *predictive (Bayesian) p-value* [Lambert, 2018a, Chap. 10.3; Sinharay and Stern, 2003]; the probability $\Pr(\text{metric}(y_j) \geq \text{metric}(\text{obs}))$, where $j \in \{0, 1, \dots, N\}$ and N is the number of coordinates of the indexing dimension of the observed variable y , y_j indicates the set of predictive samples of the model for the j -th coordinate of the indexing dimension and `obs` the set of actual observed data, and `metric` $\in \{\text{min}, \text{max}, \text{mean}, \text{std}\}$ or any other statistical metric considered [Sinharay and Stern, 2003]. The p-value represents the proportion of the N replicated datasets from the model (resulting from the sampling happening in the loop in line 3 of Algorithm 1 shown in Section 2.2.5) for which $\text{metric}(y_j) \geq \text{metric}(\text{obs})$. If none of the p-values of the test statistics is too high or low, the model is considered to generate “replicate data” similar to actual observed one based on the criteria of the provided test statistics.

Validating the priors of a model could be conducted by visualizing the priors with any typical uncertainty visualization, but conducting predictive checks visually is not that straight-forward. Histograms of the N calculated values of each statistical metric for the predicted samples are commonly used for conducting predictive checks [Lambert, 2018a, Chap. 10.3; Sinharay and Stern, 2003] to compare the distribution of the metric’s values for the predicted samples with the value of the metric for the observed data. Existing visualization tools do not generally support automatic creation of such visualizations from the definition of a model or its PPL output.

There is need for new tools to support users in probabilistic models-related tasks. These tools could support:

1. decision-makers seeking to interpret inference results or make predictions;
2. experts seeking to effectively express their prior beliefs and the implications of their be-

lies on inference;

3. data-scientists and statisticians seeking to refine and validate an inference process (debugging a PPL program such that it runs effectively).

The next section presents the design and implementation of one such tool, as proposed in this thesis, the *interactive probabilistic models explorer (IPME)*.

4.4 Interactive Probabilistic Models Explorer

This section discusses the design and implementation of the IPME tool. The discussion is split into two parts. The first part presents the challenges of encoding any PPL model and associated inference results into a coherent PPL-independent structure. This structure could be given as an input to graphical tools like IPME. The realization of a pipeline to create such a structure is shown. The second part focuses on the design and implementation of the IPME itself.

4.4.1 PPL Model Encoding

The automatic transformation of a PPL model into an interactive graphical representation is not technically straight-forward. There are two types of information required:

- **Model-related information:** some transformation of the PPL source code (i.e. the lines of code) that defines the model into model-related information;
- **Inference-related information:** the *traces* of the model's inference results (see Section 2.3.3).

This research restricts to MCMC approaches to inference, and so it always deals with collections of definite, (hopefully) independent samples representing possible model configurations. The traces can include samples from the prior, posterior, prior predictive and posterior predictive distributions.

A pipeline to encode a PPL model and its inference results into a unified PPL-independent structure is shown in Fig. 4.2 (this pipeline was suggested and implemented by Williamson [2019]). Fig. 4.2 presents the steps that a modeler should take to export a PPL model and its inference results into a form that can be accepted as input by the IPME tool. The following subsections describe the model- and inference-related information encoded in this structure and how this structure is generated by the pipeline shown in Fig. 4.2.

4.4.1.1 Model-related Information

Model-related information is necessary for the construction of the graphical representation (DAG) of the model. The construction of a DAG requires, at minimum, the names of the nodes and

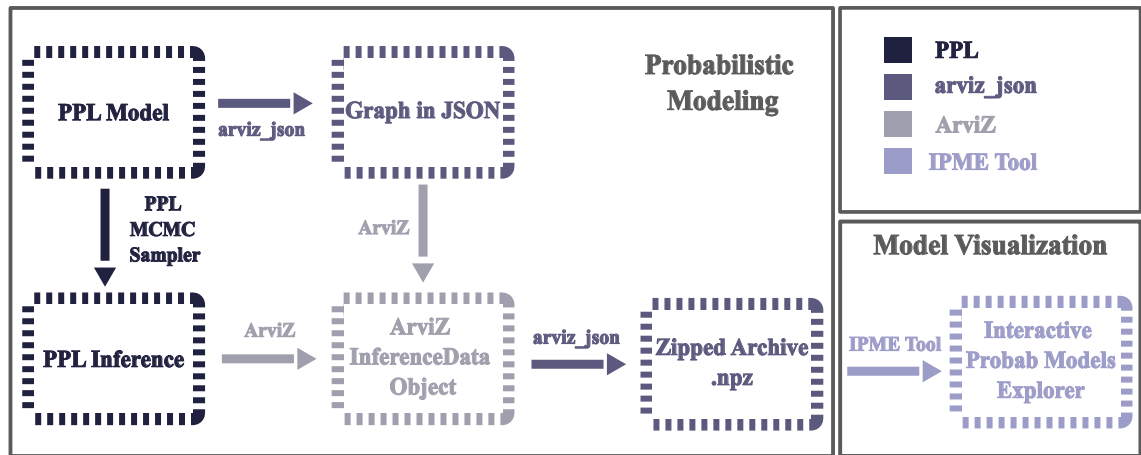


Figure 4.2: Flowchart showing the steps that a modeler should take to export the PPL-specified probabilistic model and the inference data into a standardized output, which will then be used as an input to the IPME module. The presented pipeline adds three new steps to the typical model specification and inference running routine; the export of the PPL model graph into a JSON structure; the export of the inference data into an ArviZ `InferenceData` object, where the structured description of the graph from the previous step is attached; and finally, the export of all these into a collection of `.npz` arrays and metadata in a zipped file. A concrete implementation of this pipeline is provided by the `arviz_json` module [Williamson, 2019], which was created for providing a standardized output for the PyMC3 models and inference data.

their associations that define the edges. Therefore, the input of the IPME tool should include the names of the model’s variables and a list of the parent nodes for each variable. We would also expect to be able to extract annotations for each parameter, including the distribution type (Gaussian, Poisson, binomial, Dirichlet etc.), data type (floating point, integer), tensor shape (univariate, N-d vector, MxN matrix, etc.), and inferential type (observed, free, deterministic).

In a tree-like structure like the DAG of a Bayesian network, the lowest level of nodes consists of observed variables (see Section 3.3.1.1). The DAG is “rooted” at the top with non-stochastic nodes with known, fixed values (e.g., constants used in specifying prior distributions). Unobserved variables could either be *free* parameters defined by prior distributions or they could be *deterministic* variables constituting transformations of other parameters or observed variables.

Each node in the DAG should provide some minimum information about the corresponding variable. This could include the name of the variable and the type of the variable’s distribution; this is sufficient to arrive at a graph similar to those of Kruschke [2015]. The last model-related information needed is the dimensions and coordinates of the probabilistic model’s variables; i.e. the tensor shape and its semantic structure. Interpreting tensor shapes in probabilistic programming is subtle [Ma, 2019] and may require additional annotation. The dimensions of a probabilistic model’s variables usually correspond to dimensions of the observed data and they can be used to model groups of parameters (like in the case of a heterogeneous models as explained in Section 2.2.4.1). A variable might have multiple dimensions because it is distributed

according to a multivariate distribution. IPME supports variables with univariate distributions, depicts their distribution using 2D visualizations, and enables the user to define the coordinates of the variables' dimensions corresponding to observed data.

PPLs usually provide an API for users to access this model-related information, though the form this information is provided in, varies significantly. A prototype (JSON) structure that encodes the model-related information has been used to formulate the input of IPME (see details about the JSON structure in Appendix B). The fields of this structure are vastly determined by the model-related information that IPME needs to create the graphical representation of the model. A modeler could access the model-related information using the corresponding API of any PPL and produce the prototype structure of the model's graph. A Python package called `arviz_json` was developed by Williamson [2019] to provide an API for the transformation of PyMC3 [Salvatier et al., 2016] model objects into these structures.

4.4.1.2 Inference-related Information and Data

PPLs store inference results in different ways, which are usually backend-specific. For example, PyMC3 stores the MCMC samples in PyMC3 `MultiTrace` [MultiTrace] objects, whereas PyStan stores inference results in PyStan `Fit` objects [Fit]. Thus, there is a need for a standardized backend-agnostic way of organizing and storing the inference-related information and results before forwarding it to a tool that would synthesize an interactive graphical representation of the model. The samples for estimating the prior and posterior (predictive) distributions of the model's variables should be stored in standardized structures and linked to the PPL model's variable names.

One solution to this problem is provided by the ArviZ library [Kumar et al., 2019], which provides an API for transforming the inference data of different inference back-ends and programming languages (PyMC3, PyStan, CmdStanPy, Pyro, NumPyro, emcee, and TensorFlow Probability objects) into ArviZ `InferenceData`¹ data structures. These are standardized data structures for storing MCMC-based inference results that are dependent on `xarray`'s² multi-dimensional array structures that introduce "labels in the form of dimensions, coordinates and attributes on top of raw NumPy-like multidimensional arrays". The `InferenceData` objects group various data sets that could be produced by a Bayesian analysis (prior or posterior samples, prior or posterior predictive samples, sample statistics etc.). ArviZ provides an API for exporting the `InferenceData` data structures in `netcdf`³ files. The `arviz_json` package is used for exporting ArviZ `InferenceData` objects into a zip file containing the model's DAG structure and inference results in a collection of `numpy`⁴ format arrays corresponding to those of the `InferenceData` object. JSON metadata that link the model's variables to the data ar-

¹<https://arviz-devs.github.io/arviz/schema/schema.html>

²<https://xarray.pydata.org/en/stable/why-xarray.html>

³<https://www.unidata.ucar.edu/software/netcdf/>

⁴<https://numpy.org/devdocs/reference/generated/numpy.lib.format.html>

rays and provide information about the type of the samples (prior, posterior), the dimensions and the coordinates are also included (see details in Appendix B).

4.4.2 Design and Implementation of IPME

IPME was created as a Python module that takes as an input the model- and inference-related information in the standardized format presented in the previous section, and more concretely, the output of the `arviz_json` module. IPME creates the interactive graphical representation of the model. For the development of the IPME's components `Bokeh`⁵, a Python interactive visualization library for modern web browsers, and `Panel`⁶, a Python library for interactive web apps and dashboards were used. The IPME tool provides a web-browser-based visualization of the interactive probabilistic models explorer and thanks to Bokeh that affords high-performance interactivity over large datasets, it provides a low-latency interactivity with the MCMC sample set. The code of the IPME tool and videos demonstrating the tool can be found in [Taka, 2020a,c]. I recommend the reader watches these videos as they read through the following descriptions of the tool because the interactive aspects of the tool cannot be easily demonstrated in the static form of figures shown.

Box 4.2 presents a hierarchical model that is used as a reference example in this chapter. Fig. 4.3 presents the IPME representation of this model.

Box 4.2 The Eight Schools' Hierarchical Model

This example presents a hierarchical model for predicting the effect of coaching programs on the scholastic aptitude test (SAT) for the admission to college in the US (see Appendix A.2 for details about the model, and its specification in PyMC3). Observed data from eight different schools is available. The eight schools' hierarchical model is described by the following set of probabilistic statements:

$$\mu \sim \text{Normal}(\mu = 0, \sigma = 5) \quad (4.1)$$

$$\tau \sim \text{Half-Cauchy}(x_0 = 0, \gamma = 5) \quad (4.2)$$

$$\theta_i \sim \text{Normal}(\mu = \mu, \sigma = \tau) \quad (4.3)$$

$$y_i \sim \text{Normal}(\mu = \theta_i, \sigma = \sigma_i) \quad (4.4)$$

where $i \in \{1, \dots, 8\}$, y_i are the observed changes in the SAT scores, and σ_i are the standard errors of the observed scores' changes.

IPME presents the model's DAG in a tree-like structure. The tree is simplified into rows of nodes ordered vertically from hyperpriors down to observed values. This format is a vertically ordered representation which orders nodes such that constant values (which are not shown)

⁵<https://docs.bokeh.org/en/latest/>

⁶<https://panel.holoviz.org/index.html>

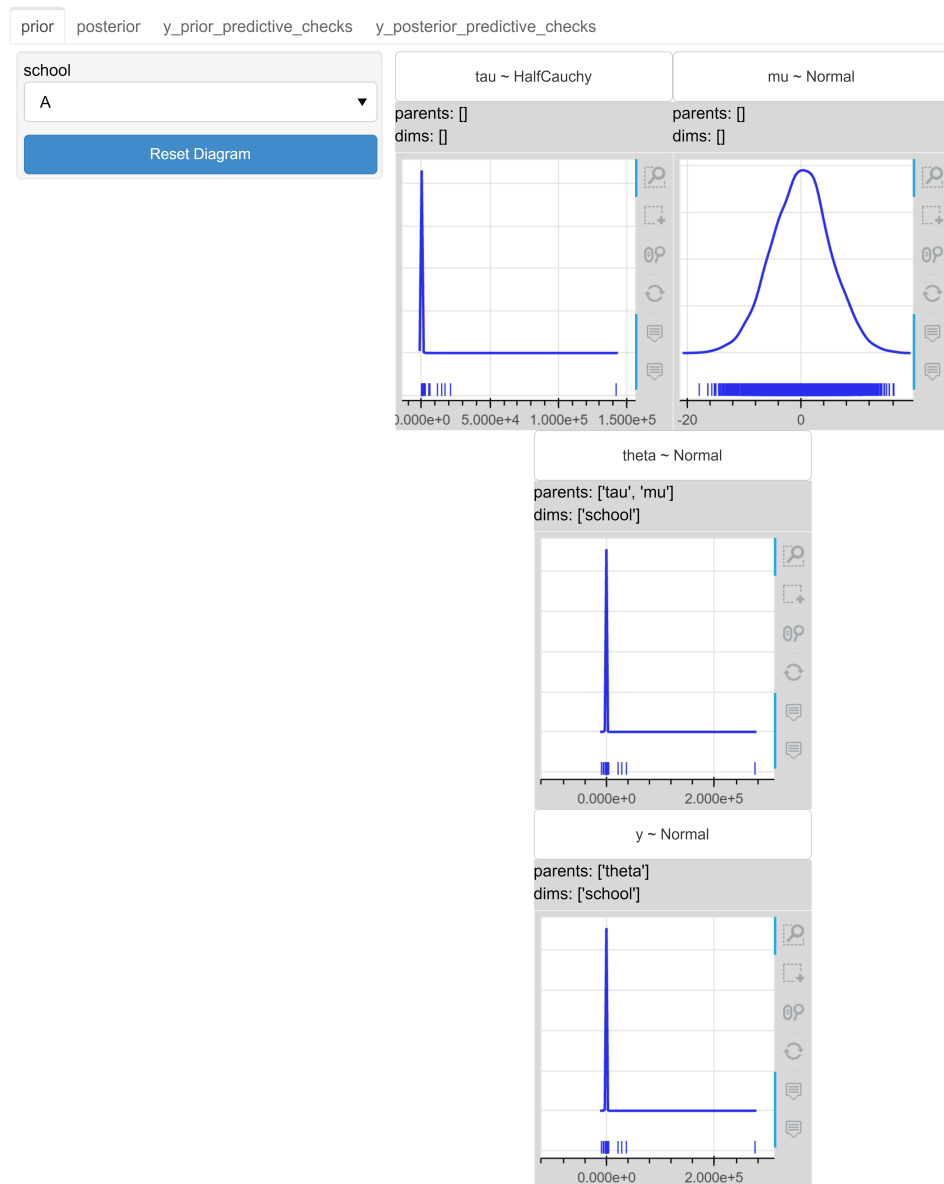


Figure 4.3: The IPME representation of the hierarchical model of the eight schools problem presented in Box 4.2. The KDE plots shown present the prior marginal distributions of the model’s variables. PyMC3 was used for the model’s specification and inference.

would be placed at the very top of the graphical representation and the lowest nodes in the graph are observed variables. Parameter nodes are organised such that child nodes appear on rows lower than their parent prior distributions. In most hierarchical models, which have well-separated hyperpriors, this leads to a neat separation of the graph into rows and an obvious reading of the graph from top to bottom. There are, of course, graphs which are hard to arrange well in this format. Each node is a cell in a `Panel GridSpec`⁷ object. Each node of the graph has a toggle button, some text and a KDE distribution plot. The toggle button label presents the variable’s name, the tilde (\sim) symbol and the variable’s distribution name. The text below the

⁷https://panel.holoviz.org/user_guide/Components.html

toggle button states the parent nodes and the dimensions of the variable. The KDE curve can be hidden with the toggle button (Fig. 4.4).

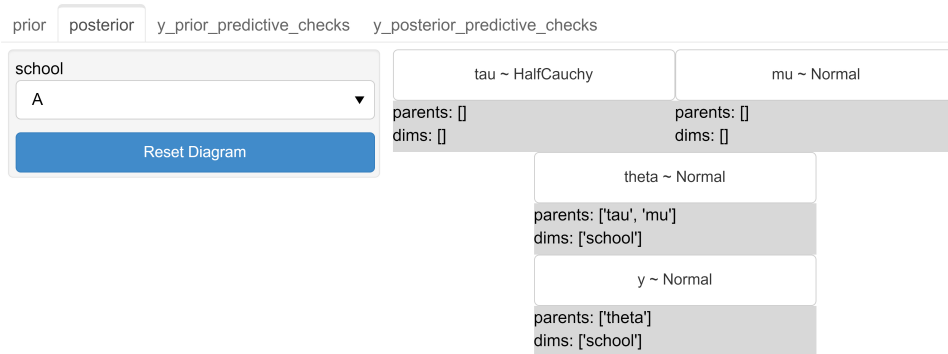


Figure 4.4: The collapsed IPME representation of the hierarchical model of the eight schools problem presented in Box 4.2. The toggle buttons were used to hide the KDE plots.

The variables' distribution is presented using a kernel density estimation⁸ (KDE) algorithm, computed based on the corresponding variable's (MCMC) samples, using Silverman bandwidth estimation [Silverman, 1986]. The KDE curve of the non-observed variables is estimated based on the (MCMC) samples of the (posterior) prior marginal distribution of the variable, while the KDE curve of the observed variables is estimated based on the (posterior) prior predictive (MCMC) samples. A rug plot is presented below the KDE curve to display the corresponding (MCMC) samples as tick markers. Two views of the DAG are presented, one showing the model's distribution in the prior and one in the posterior sample space, in separate tabs so that users can compare both quickly.

IPME uses the dimensions' and coordinates' information to automatically create a widget box on the left-side of the graphical representation. Each indexing dimension in the model is converted into a drop-down menu presenting the semantically meaningful coordinates of the indexing dimension. Users can select the value of the coordinate for each indexing dimension and get a different view of the data. The range of the x-axis of each variable's KDE plot is fixed across all coordinates of the indexing dimensions in each one of the prior or posterior sample space to enable immediate comparisons among the various coordinates of an indexing dimension. These ranges are calculated by taking the widest distribution in each sample space (prior or posterior).

Users can apply a value range condition to any variable in the model interactively by drawing a variable-width and fixed-height selection box on the corresponding KDE plot of the variable (Fig. 4.5). The part of the KDE curve that is enclosed into the selection box is highlighted in green and a second KDE curve that is computed based on the restricted (MCMC) sample set is drawn in orange. The rug plot is updated accordingly to give a better perception of the conditioning process on the sample space. The color palette of the `arviz-darkgrid` style

⁸https://en.wikipedia.org/wiki/Kernel_density_estimation

was used because it was designed to be color-blind friendly and it would add a tone of familiarity for the users of the ArviZ library. The user can update their initial selection by drawing a new selection box or can draw additional selection boxes on other distributions to add constraints to the query.

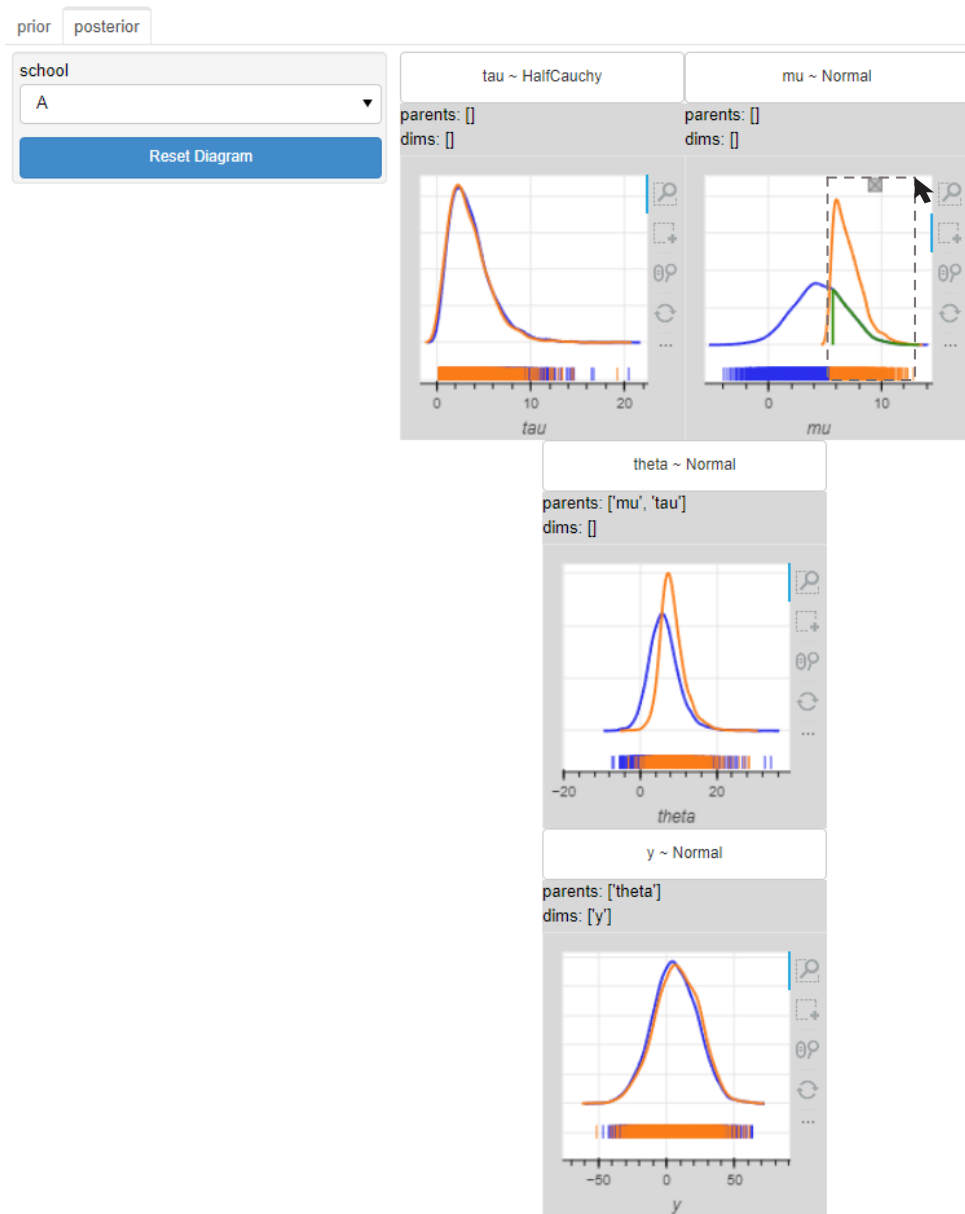


Figure 4.5: The IPME representation of the hierarchical model of the eight schools problem presented in Box 4.2. The KDE plots shown present the posterior marginal distributions of the model’s variables.

Every time the user draws a selection box, both the prior and the posterior spaces are restricted to include only samples that lay within the selected subset. The user can restrict the value range of any parameter at any coordinate of any indexing dimension and the restriction to the sample space will be reflected in both the prior and posterior, as well as to the prior and posterior predictive histograms. This kind of interaction allows the user compare the changes in

the uncertainty in the restricted space between the prior and posterior beliefs about the model parameters. The user can also remove the restriction of a parameter by clicking on the “x” button that accompanies the updated KDE curve. Finally, the user can reset all the uncertainty visualizations by globally removing all the restrictions by clicking the “Reset Diagram” button on the left-hand side of the graph.

IPME provides an interface for predictive model checking with *predictive p-values* (Fig. 4.6). The prior and posterior histograms of four statistical metrics (min, max, mean, std) over the prior and posterior predictive samples, respectively, are presented in two extra separate tabs. The actual observed value of each test statistics is indicated by a vertical black line on the corresponding histogram. The (Bayesian) p-values are also noted in the legend above each histogram. For the predicted data to capture sufficiently the test statistics of the observed data we expect the probability of the corresponding test statistic of the observed data to have a sufficiently big probability given the distribution of the test statistic of the predicted data. Usually the application might set a specific requirement as to the size of this probability.

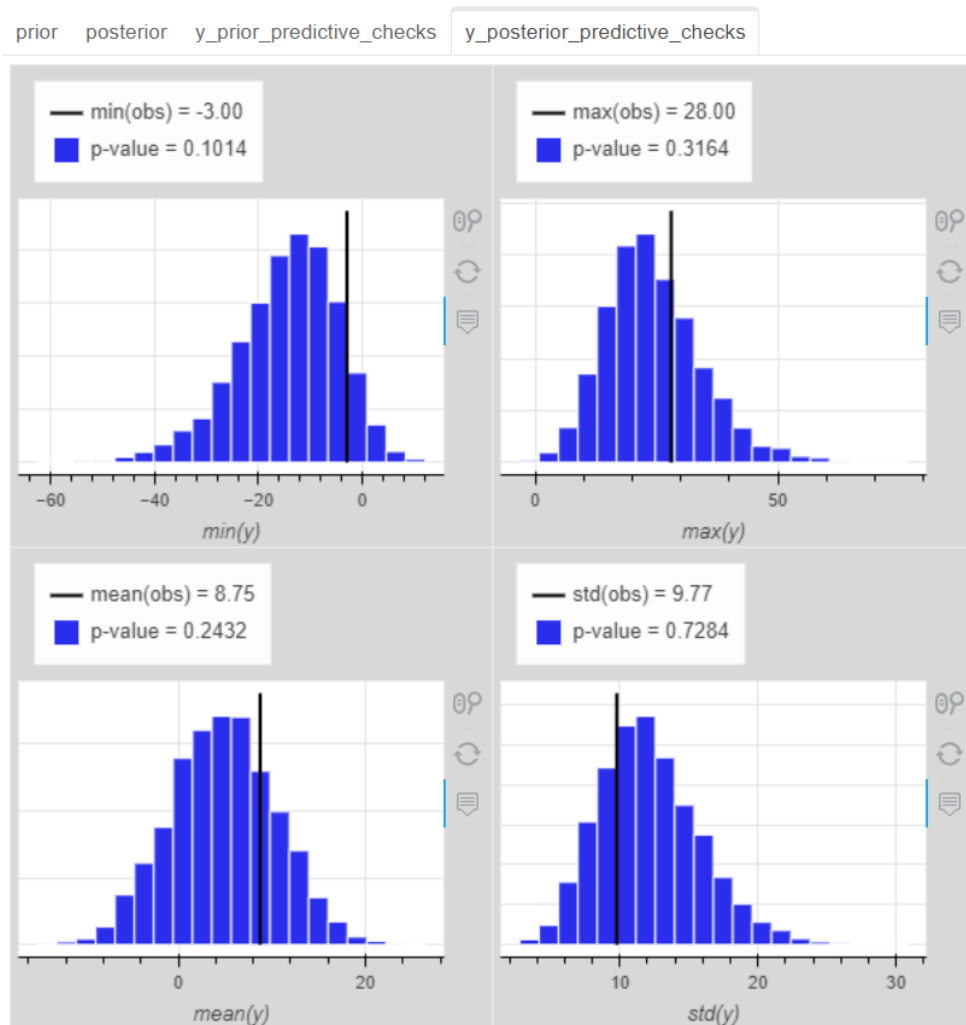


Figure 4.6: The posterior predictive test statistics (min, max, mean, std) of the model of the eight schools problem presented in Box 4.2.

4.4.3 Objectives of IPME

IPME was designed to fulfil a set of objectives that were identified to support the tasks that users of Bayesian analysis typically undertake. These objectives were summarized in Section 4.2 and are repeated here:

1. automatic synthesis of a graphical representation of a model independently of the PPL used;
2. seamless integration of uncertainty visualization into the graphical representation of the model;
3. inclusion of both prior and posterior distributions of the model’s parameters, and the corresponding predictive ones for the observed variables;
4. interactive exploration of inference MCMC sample space;
5. interactive sensitivity analysis of model’s parameters;
6. granularity in the presented visual information according to user’s choices and needs;
7. inclusion of predictive checks.

Let us discuss now how IPME meets these objectives.

Objective 1. The transformation of a PPL model – some lines of code of a PPL, potentially alongside some observed data – into a graphical representation of the model is the purpose of this objective. The intention is to transform PPL models independently of the PPL and structure of the specific model implemented. IPME was designed to accept as an input the model- and inference-related information encoded in a standardized format as explained in Section 4.4.1. This allows the automatic synthesis of a graphical representation of a model independently of the PPL used. By the term “graphical representation” a coherent graph-like representation of the probabilistic model is meant, which reveals the internal structure of dependencies among the model’s parameters. IPME presents the model’s graph in a tree-like structure with each node representing a variable in the model. The nodes contain annotations to indicate the parent-child relationships of the model’s variables.

Objective 2. The integration of parameters’ distribution into the graphical representation of the model is the purpose of this objective. In IPME the distributional information of each model’s variable is integrated into the graphical representation of the model by including an uncertainty visualization (i.e., a KDE plot) into the node of the variable. A Bayesian model can simulate drawing samples in the space of observations, which is sometimes the ultimate task (e.g., in prediction), is sometimes essential for validation and calibration (e.g., in prior predictive checks), and is sometimes the most relevant way to explain consequences to users. The

prediction of an observed variable's value by Bayesian probabilistic models is a distribution over possible (prior or posterior) observations and not just a single point estimate (the MCMC predictive sampling was explained in Section 2.2.5). Thus, nodes that represent observed variables also include an uncertainty visualization to present the predictions' uncertainty in IPME.

Objective 3. The integration of both prior and posterior distributional information of model's variables into the graphical representation of the model is the purpose of this objective. Bayesian probabilistic models can provide estimations of the model's uncertainty both before and after seeing the observed data. Two visual representations should be incorporated in each node of the graphical representation, one for the prior and one for the posterior distribution of the associated variable. In IPME there is a tab bar at the top which allows the user to switch between the prior and posterior view of the graphical representations.

Objectives 4 - 5. The interactive exploration of the variables' (prior and posterior) sample space and sensitivity analysis of model's parameters is the purpose of these two objectives. IPME enables users to apply value range conditions to model's variables by dragging a fixed-height selection box on the KDE plot of the variable. Appropriate highlighting in the form of brushing-and-linking is applied on the KDE curves and rug plots of all variables of the model both in the prior and posterior sample space to indicate the sample subset belonging to the restricted sample space. The KDE curves of all variables in this restricted sample space are drawn.

Objective 6. Adjustable granularity is the purpose of this objective. Some users may wish to have a simplified summary view; while others may be involved in tasks like validating sampling and require detailed interactive visualisations. It therefore makes sense the nodes of the model's graphical representation to be collapsible in some way, and the user is able to interactively select the information to be revealed. In IPME the toggle buttons used as headers in the nodes of model's variables serve this purpose: hide the variable's KDE plot when clicked.

Objective 7. The inclusion of prior and posterior predictive checks is the purpose of this objective. In IPME users can use the tab bar to switch between the prior and posterior predictive checks view where the prior and posterior histograms, respectively, have been calculated for the test statistics of the predictive samples and shown along with the test statistic of the observed data and the corresponding predictive p-value.

4.5 What Makes IPME a Unique Tool

This section discusses how other PPL-linked available tools display the graph structure of probabilistic models and compares this with IPME (Section 4.5.1). It also compares the way that inference results are presented using typical presentation practices with the way IPME achieves this, and discusses how IPME was designed to provide a more compact and flexible presentation of the inference results (Section 4.5.2). This section presents also a comparison of IPME with

existing visualization libraries for Bayesian analysis (Section 4.5.3) based on a set of indicative visualization features.

4.5.1 Presentation of the Probabilistic Programming Model’s Graph

Fig. 4.7(a) presents the graphical representation of the hierarchical model of the eight schools problem (shown in Box 4.1) created with DoodleBUGS [Spiegelhalter et al., 2003]. Fig. 4.7(b) presents the graphical representation of the same model using the PyMC Graphviz interface (`pymc.model_to_graphviz`). Although both ways present the structure of the probabilistic model as a graph providing an at-a-glance representation of the model’s parameters and dependencies derived from the PPL specification of the model, they both lack the presentation of the parameters’ distribution. Fig. 4.7(c) presents (a manually created) Kruschke-style diagram of the same model, which is more informed since it presents the prototypes of the prior distributions of the model’s parameters. But still, this representation of the probabilistic model does not provide any indication of the actual distribution of latent parameters.

Using the framework presented in Fig. 4.2, the IPME representation of the eight schools’ hierarchical model was created. Fig. 4.7(d) presents the IPME representation for the posterior space as an expanded DAG (Fig. 4.4 presents its collapsed form). IPME seamlessly integrates the actual uncertainty estimations into the graph’s nodes and provides variable granularity by allowing users to collapse certain elements. The widget box on the left-hand side contains one widget per indexing dimension and allows selecting different views of the inference data. Finally, the tab bar at the top enables switching between prior and posterior.

4.5.2 Presentation of the Inference Results

The most common practice in reporting of inference results in Bayesian analysis is the creation of tables that present summary statistics of the posterior distributions. As the number of model parameters or coordinates of indexing dimensions increases, these tables become unwieldy. For example, Silva et al. [2015] use a rather massive table of summary statistics when analyzing data of wildfires in Portugal between 1990–1994. The limited capacity of human cognition could be a hurdle for users like decision-makers to grasp the uncertainty presented in tables of this sort and assess the risk appropriately. The numerical data presented are usually statistics like mean, standard deviation or confidence intervals, which could mislead or overwhelm unfamiliar users.

A static representation of the inference data in summary tables cannot communicate sensitivity of the parameters, which would allow decision-makers to assess the impact of parameter inter-dependencies and associated risks. Communicating the prior would imply communicating a second table of similar complexity, which is often omitted in reports of Bayesian probabilistic models. For example, Table 4.1 presents the summary statistics of the posterior of the eight schools’ hierarchical model. The ArviZ API (`arviz.summary`) was used to produce this

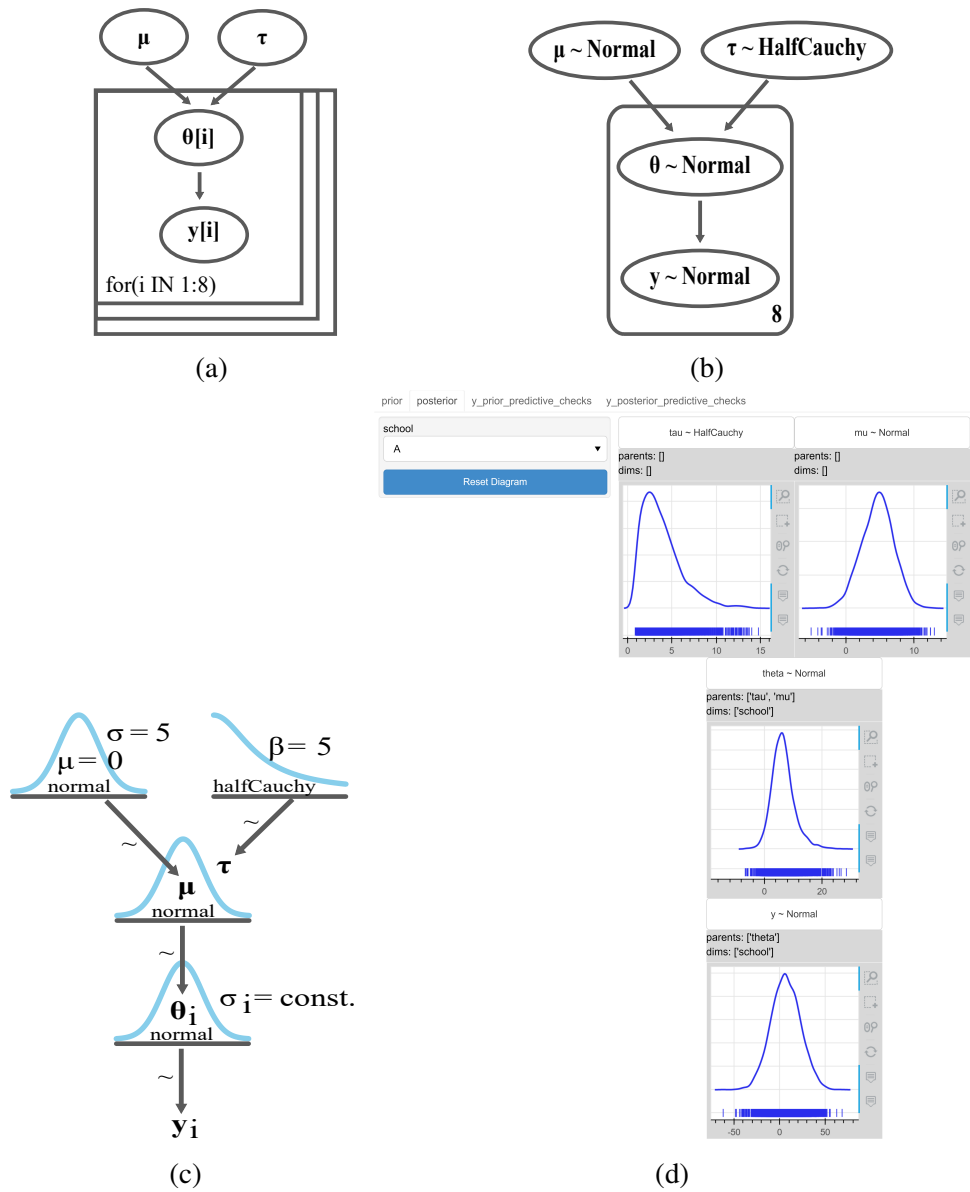


Figure 4.7: Graphical representations of the eight schools model.(a) DoodleBUGs graph, (b) PyMC graph using the Graphviz interface, (c) Kruschke-style diagram, (d) IPME representation.

table.

Another way of communicating inference results is by generating an uncertainty visualization to show the marginal distribution for each model’s parameter independently. This solution is not very common, although it is more informative than the tables of summary statistics. The reason for this is that it leads to a large number of visualizations in the case of many-parameter models, which are difficult to communicate in a concise way. For example, Fig. 4.8 presents the posterior densities for the parameters of the eight schools’ hierarchical model. Although, this is a rather simple model, we see that it produces 10 different uncertainty visualizations. This number could rise even more if the parameters of the model had more indexing dimensions or the indexing dimensions more coordinates.

Table 4.1: Posterior statistics in a tabular format for the eight schools' hierarchical model. This is a rather simple model and the table only consists of ten rows. This number could rise immensely if the model had more parameters or the parameters had more (multi-valued) indexing dimensions.

	mean	std	HDI_5%	HDI_50%	HDI_95%
mu	4.479	3.392	-1.090	4.562	10.012
theta[0]	6.524	5.849	-1.590	5.937	16.632
theta[1]	5.128	5.001	-2.892	5.104	13.255
theta[2]	3.991	5.589	-5.179	4.226	12.165
theta[3]	4.923	5.094	-3.503	4.914	13.160
theta[4]	3.519	4.934	-5.189	3.895	10.993
theta[5]	4.142	4.973	-4.137	4.311	11.903
theta[6]	6.697	5.294	-0.998	6.197	16.067
theta[7]	5.026	5.667	-3.804	4.877	13.716
tau	4.127	3.144	0.937	3.281	10.151

IPME presents an overview of the model's parameters in two similar tree-like structures for the prior and posterior that are as big as the number of the model's parameters. The number of indexing dimensions or coordinates does not affect the size of the graphical representation, in comparison to the summary statistics tables or uncertainty visualizations where each extra coordinate results in an extra row in the table or an extra graph, respectively. The presentation of inference results with IPME becomes more compact and concise. Users have more flexibility in the exploration of inference results with IPME; they can define the coordinates of the indexing dimensions, expand the nodes of interest and collapse the rest, compare priors to posteriors.

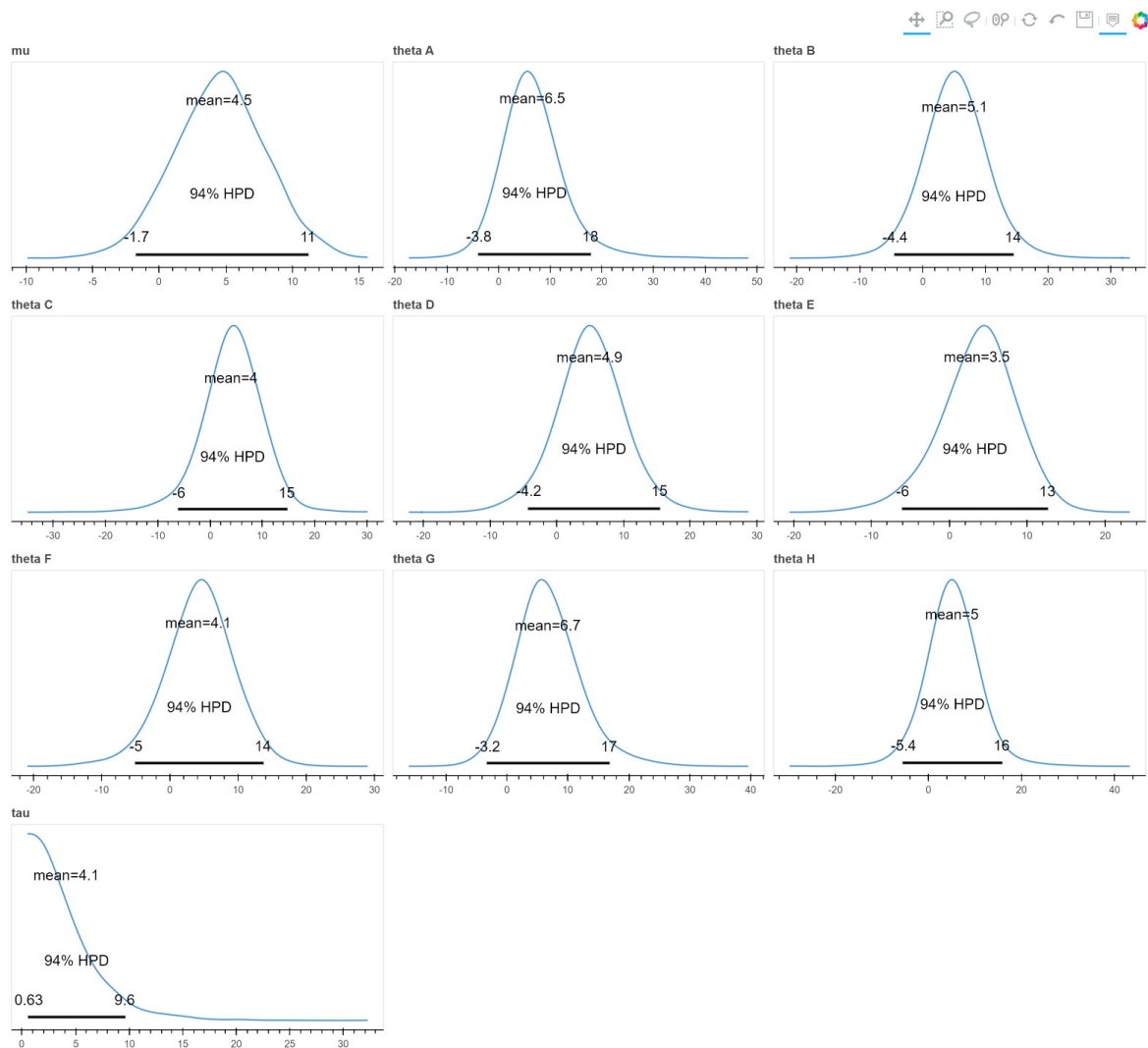


Figure 4.8: The posterior densities of the eight schools’ hierarchical model presented in the style introduced by Kruschke [2015]. Although, this is a rather simple model, we can see that it can produce 10 different uncertainty visualizations. This number could rise even more if the parameters of the model had more (multi-valued) indexing dimensions.

4.5.3 Comparison of IPME with Existing Visualization Libraries for Bayesian Analysis

There are various existing visualization libraries specialised for presenting the outputs of Bayesian analysis. The aim of this section is to present the main visualization features the most commonly used such libraries offer and compare them to the proposed IPME tool. Table 4.2 provides a comparative summary presentation of these tools including the IPME, based on a set of visualization and analysis means for inference results in Bayesian analysis.

Kumar et al. [2019] created *ArviZ*, a unified Python tool for exploratory analysis, processing and visualization of the inference results of probabilistic programming models. *ArviZ* integrates

Table 4.2: Comparative presentation of existing Bayesian analysis visualization libraries including IPME. IPME offers unique features that are not encountered in any other of the existing tools; the interactive exploration of the (MCMC) sample space and the graphical analysis of the model.

	ArviZ	bayesplot	tidybayes	shinystan	IPME
Visual summaries of parameters	yes	yes	yes	yes	yes
Numerical summaries of parameters	yes	no	yes	yes	no
PPL/MCMC-algorithm independent input	yes	yes	no	no	yes
MCMC diagnostics	yes	yes	no	yes	no
Predictive test statistics	yes	yes	no	yes	yes
Semantic definition of indexing dimensions and coordinates	yes	no	yes	no	yes
Interactive customization of visual/numerical summaries	no	no	no	yes	yes
Interactive exploration of MCMC sample space	no	no	no	no	yes
Graphical analysis of model	no	no	no	no	yes

seamlessly with various established PPLs, which makes it a very powerful tool in the field of Bayesian data analysis. While it has sophisticated tools for visualising trace statistics and model diagnostics, it does not analyse the model graph, nor does it offer interactive visualisation tools. A corresponding tool in R is *bayesplot* [Gabry and Mahr, 2020], which provides a variety of plotting functions and MCMC diagnostics for users working with a variety of R packages for Bayesian modeling, such as RStan or packages powered by RStan, such as *rstanarm* or *brms* packages. Another visualization tool of Bayesian analysis in R is *tidybayes* [Kay, 2020]. Tidybayes extracts, manipulates, and visualizes prior and posterior samples from Bayesian models of a range of PPLs and Bayesian analysis packages (JAGS, Stan, *rstanarm*, *brms*, *MCMCglmm*, *coda*) in a tidy data format. Common visualization primitives (ggplot geometries) are exploited for the visualization of priors and posteriors like quantile dotplots, eye plots, point/interval sum-

maries, and fit curves with multiple, arbitrary uncertainty bands.

Shinystan [Stan Development Team, 2017] is a web-based interactive Bayesian exploratory tool in R, which is PPL-agnostic and offers “customizable visual and numerical summaries of model parameters and convergence diagnostics for MCMC simulations”. This tool, although it exploits interactivity to customize the visual presentation of the data, it does not offer the possibility of exploring prior and posterior sample space through interactive conditioning or any analysis of the model’s graph.

4.6 Use Case Scenarios

This section presents use case scenarios, where IPME is used in realistic modeling problems. IPME could assist modelers in model checking and validation. IPME provides two possibilities for checking the model. The first arises through the prior interactive graphical representation, where users could explore and observe the prior beliefs that were set during the model definition process and the prior predictive distributions along the various coordinates of the indexing dimensions. The consistency of the model’s priors with prior knowledge and experience could be investigated in this way. The second arises from the prior and posterior predictive model checking with predictive p-values. Users could observe how well aspects of the observed data are represented in the predictions of the model.

IPME could also help users acquire a more intuitive comprehension of various aspects of the model and inference results. This could be achieved through interactivity. IPME offers two types of interactivity; the interaction with the indexing dimensions that allows the exploration of the data from different viewpoints, and the interactive conditioning on the sample space that allows the exploration of the prior and posterior sample spaces. The first type of interactivity could reveal similarities or differences between groups of data. The second type of interactivity could reveal associations between parameters, changes in the parameters and predictions uncertainty under certain circumstances (conditions), or the effects of priors on posteriors. The following use case scenarios illustrate all the above.

4.6.1 Drivers’ Reaction Time

The first use case is based on the drivers’ reaction time problem used in Chapters 1, 2, and 3. The data scientist of the logistics company created a homogeneous and hierarchical probabilistic linear regression model to predict the reaction times of the drivers on each day of driving (see models’ definition in Box 4.3). The bigger the slope of the regression line is, the more tired (bigger reaction times) the drivers get after consecutive days of driving. The data scientist used PyMC3 for the specification of the model and the inference, `arviz_json` to export the models in a standardized format, and IPME to visualize the models and their outputs (Python code can be found in Appendix A.3).

Box 4.3 The Drivers' Reaction Time Models Defined by the Data Scientist

The data scientist of the logistics company specified the following probabilistic regression models for the drivers' reaction time problem. The *homogeneous* model is described by the following probabilistic statements:

$$a \sim \text{Normal}(\mu = 100, \sigma = 250) \quad (4.5)$$

$$b \sim \text{Normal}(\mu = 10, \sigma = 250) \quad (4.6)$$

$$\sigma \sim \text{Half-Normal}(\mu = 0, \sigma = 250) \quad (4.7)$$

$$y_{\text{pred}_i} \sim \text{Normal}(\mu = a + t \cdot b, \sigma = \sigma), \quad (4.8)$$

where $i \in [0, 17]$ indexing the drivers, $t \in [0, 9]$ indexing the days of driving, and y_{pred_i} denotes the predicted reaction time of driver i .

The *hierarchical* model is described by the following probabilistic statements:

$$\mu_a \sim \text{Normal}(\mu = 100, \sigma = 250) \quad (4.9)$$

$$\sigma_a \sim \text{Half-Normal}(\mu = 0, \sigma = 250) \quad (4.10)$$

$$\mu_b \sim \text{Normal}(\mu = 10, \sigma = 250) \quad (4.11)$$

$$\sigma_b \sim \text{Half-Normal}(\mu = 0, \sigma = 250) \quad (4.12)$$

$$\sigma_{\sigma} \sim \text{Half-Normal}(\mu = 0, \sigma = 200) \quad (4.13)$$

$$a_i \sim \text{Normal}(\mu = \mu_a, \sigma = \sigma_a) \quad (4.14)$$

$$b_i \sim \text{Normal}(\mu = \mu_b, \sigma = \sigma_b) \quad (4.15)$$

$$\sigma_i \sim \text{Half-Normal}(\mu = 0, \sigma = \sigma_{\sigma}) \quad (4.16)$$

$$y_{\text{pred}_i} \sim \text{Normal}(\mu = a_i + t \cdot b_i, \sigma = \sigma_i) \quad (4.17)$$

where $i \in [0, 17]$ indexing the drivers, and $t \in [0, 9]$ indexing the days of driving, and y_{pred_i} denotes the predicted reaction time of driver i .

4.6.1.1 Model Check

The data scientist uses the IPME first to check the models. Both models predict a priori negative slopes (Fig. 4.9 and 4.10) meaning that drivers could have faster reaction times as days pass by. This is not realistic, but could happen with a small probability. The models also predict negative reaction times meaning that drivers could react before even they see a stimulation. Very big reaction times, close to tens of seconds, are also predicted as we move closer to the 10th day. It seems that the priors might not fit well the prior knowledge about the problem.

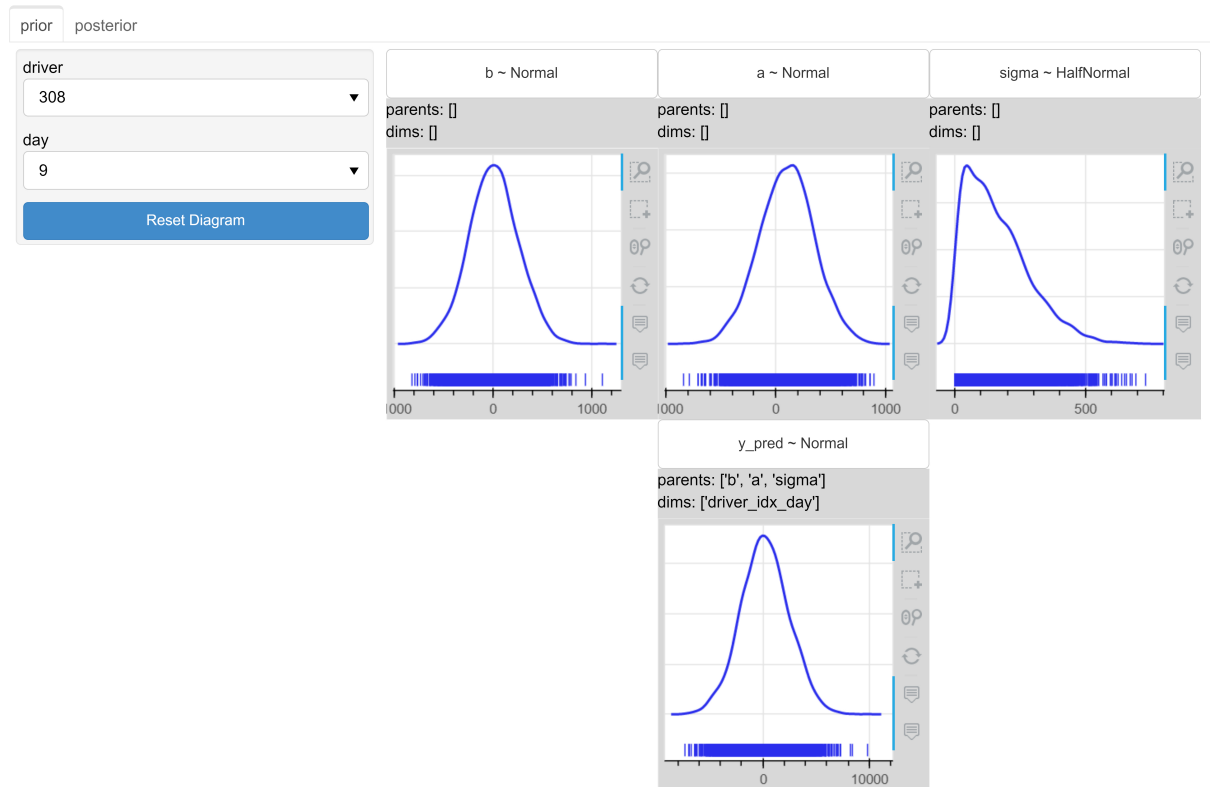


Figure 4.9: The prior IPME representation of the homogeneous drivers' reaction time model. The model predicts negative slopes, negative and very big (tens of seconds) values of reaction times a priori, which indicates that the priors might not fit well with the prior knowledge about the problem.

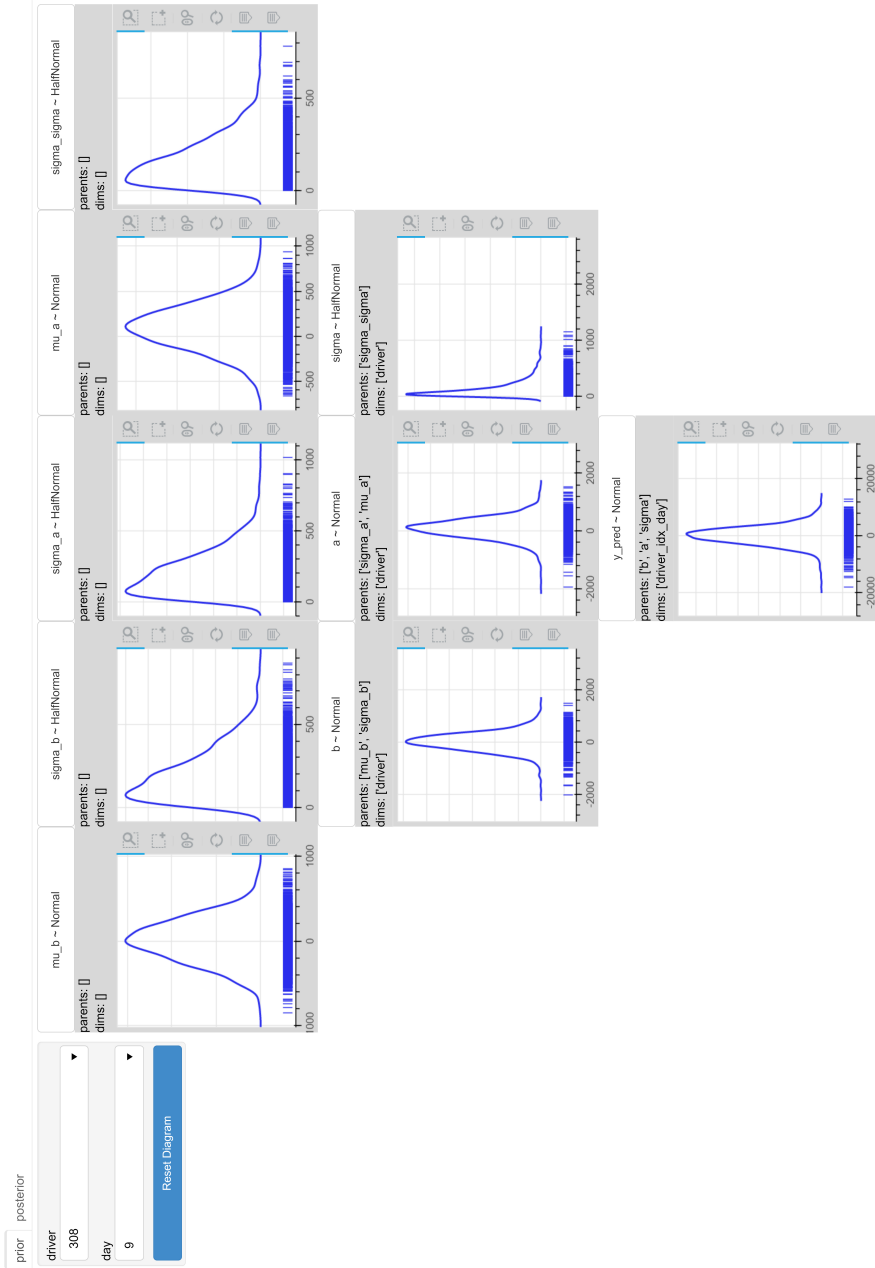
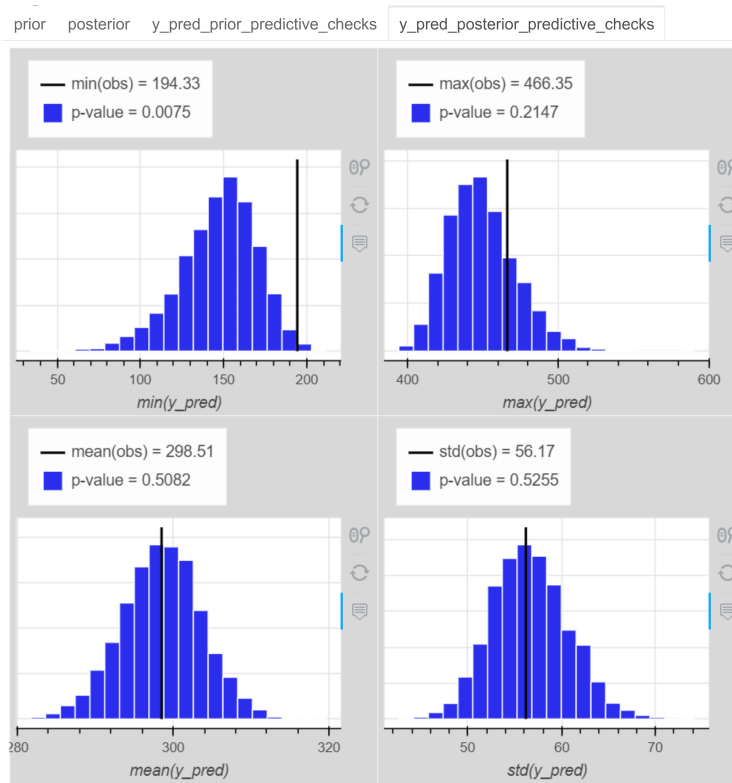
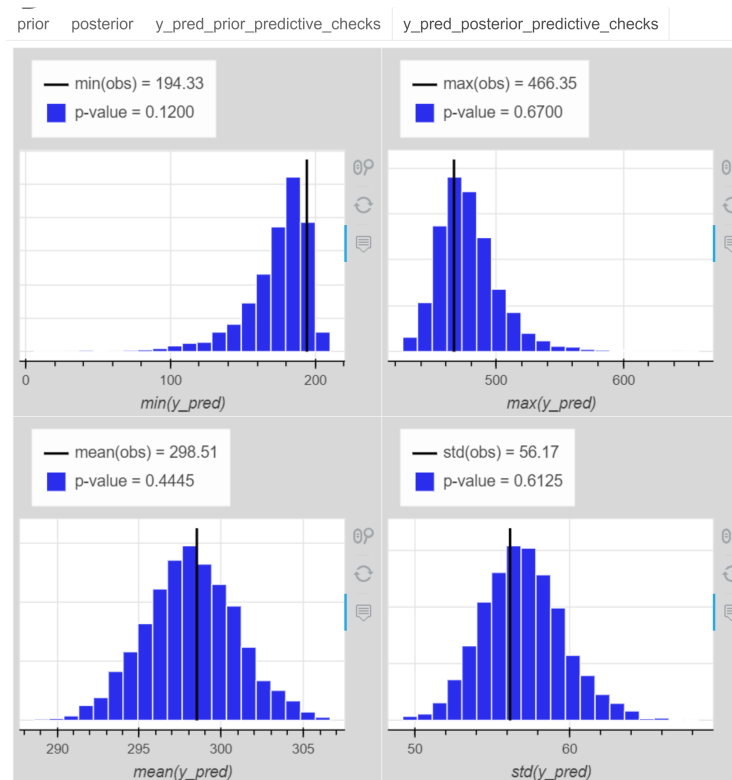


Figure 4.10: The prior IPME representation of the hierarchical drivers' reaction time model. The model predicts negative slopes, negative and very big (tens of seconds) values of reaction times a priori, which indicates that the priors might not fit well with the prior knowledge about the problem.

The data scientist wonders which model of the two is the most appropriate, and thus, observes the posterior predictive test statistics of both models. Fig. 4.11(a) and (b) present the four histograms for the homogeneous and hierarchical model, respectively. The homogeneous model gives a very low posterior p-value for the `min` test statistics, which is improved in the hierarchical model. The hierarchical model improves the p-values of the `min`, `max`, and `std` test statistics of the observations in the predictions.



(a)

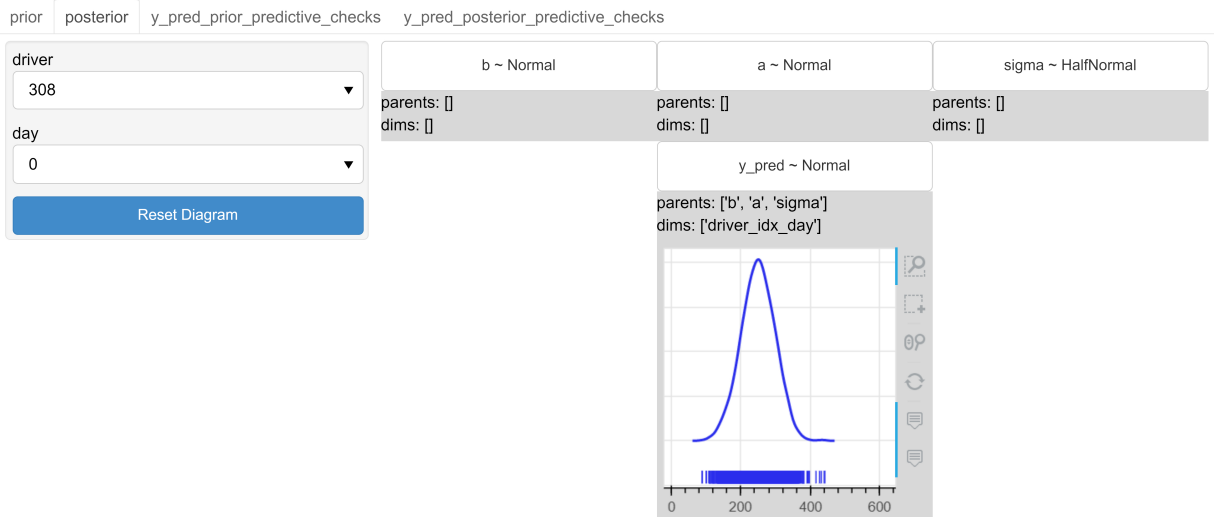


(b)

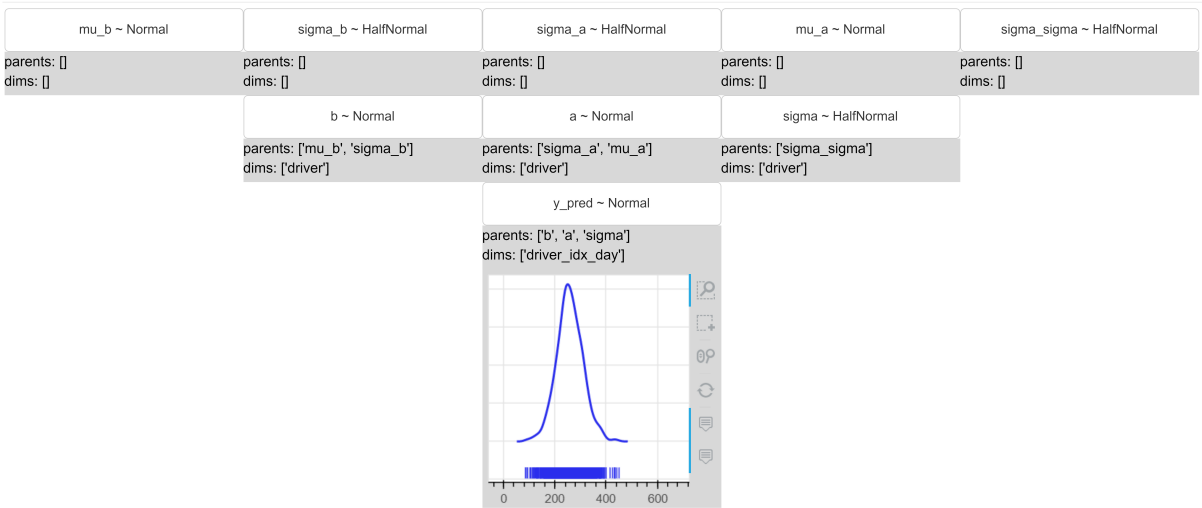
Figure 4.11: The posterior predictive test statistics of the (a) homogeneous and (b) hierarchical drivers’ reaction time model. The hierarchical model improves the representation of the `min`, `max`, and `std` test statistics of the observations in the predictions.

4.6.1.2 Interactivity

Using the interactive drop-down menus, the data scientist observes the posterior predictive distribution of each driver on the same day of driving and realizes that the homogeneous model does not present significant differences in the uncertainty of the predicted reaction times among drivers in contrast to the hierarchical one, which actually does (Fig. 4.12 and 4.13) (this topic was discussed in Sections 2.2.4.1 and 2.2.6). The data scientist chooses the hierarchical model and passes it over to the logistics' manager.

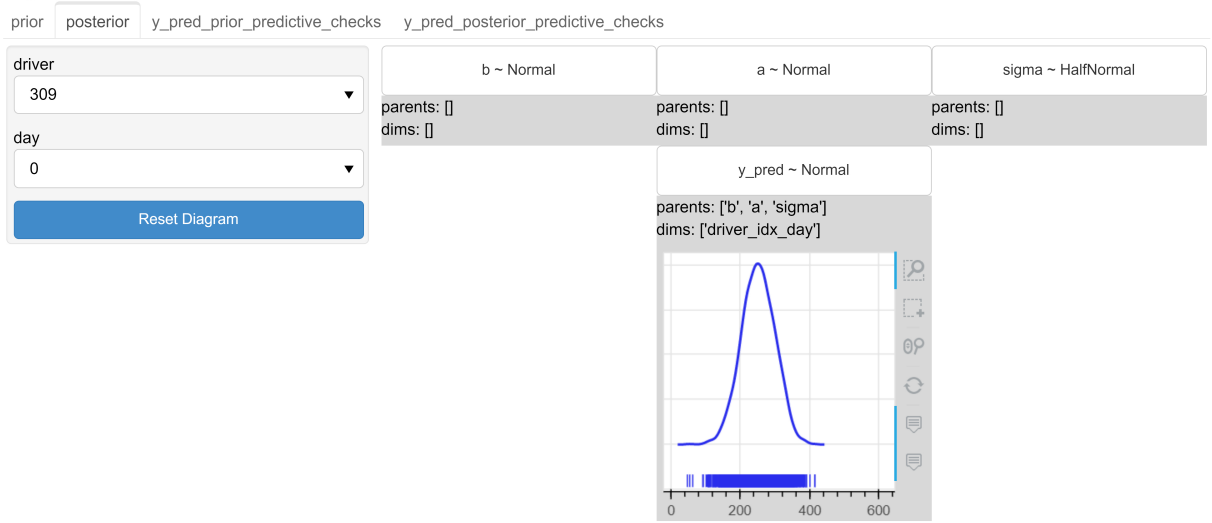


(a)

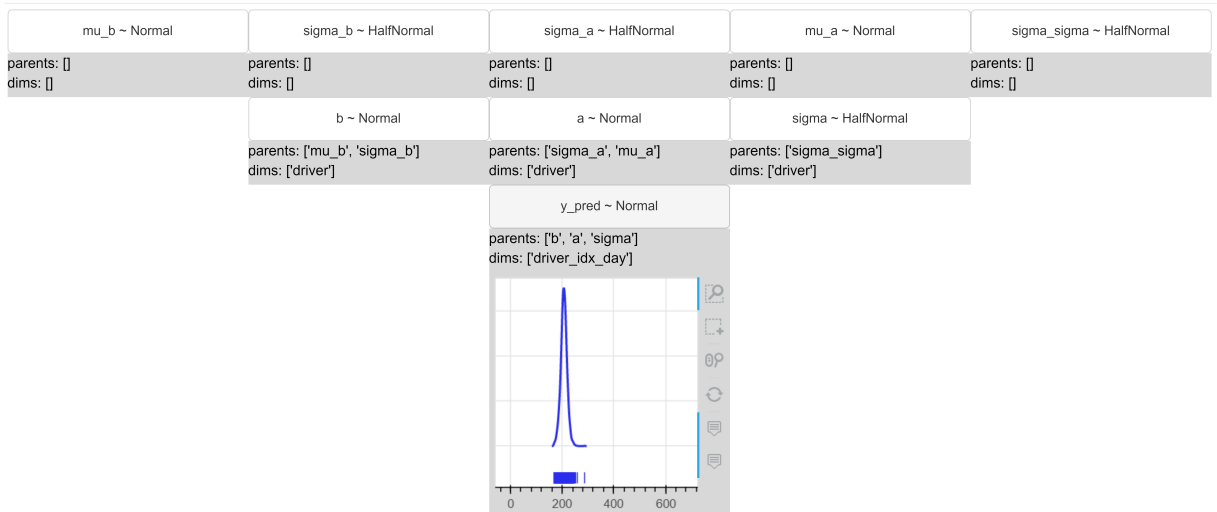


(b)

Figure 4.12: The IPME representation of the drivers' reaction time models, where we set the driver to **308** and observe the posterior predictive distribution of the reaction times for the homogeneous and hierarchical model. The posterior predictive distribution of the reaction times for driver **308** and the (a) homogeneous and (b) hierarchical model. The homogeneous model does not reveal significant differences among drivers.



(a)



(b)

Figure 4.13: The IPME representation of the drivers’ reaction time models, where we set the driver to **309** and observe the posterior predictive distribution of the reaction times for the homogeneous and hierarchical model. The posterior predictive distribution of the reaction times for driver **309** and the (a) homogeneous and (b) hierarchical model. The homogeneous model does not reveal significant differences among drivers.

The logistics' manager has to choose between two available drivers, driver 310 and 335, who would take over an urgent shipping of a cargo that had to be delivered in 6 days, although it would normally require 9. Using the drivers drop-down menu, the logistics' manager observes the uncertainty of the predicted reaction times for both drivers on the 6th day of driving (Fig. 4.14 and 4.15); driver 310 has a wider distribution (more uncertainty), but centered to lower reaction times, whereas driver 335 has a tighter distribution (less uncertainty), but centered to higher reaction times.

The logistics' manager would like to see how the uncertainty of the predicted reaction times would look like in the worst case of the model's predictions; the bigger values of slope. He sets a condition on the hyperprior of the mean value of the slopes to restrict the posterior sampling space to higher values of slopes (Fig. 4.14 and 4.15). The distribution of driver 310 becomes wider with a slight shift to lower reaction times, whereas the distribution of driver 335 becomes again wider with a slight shift to higher reaction times. It seems that driver 310 is more robust to the worst case conditioning, although initially he was having more uncertainty over his predictions in comparison to driver 335.

A video demonstrating a similar scenario can be found in the talk Taka [2020c].

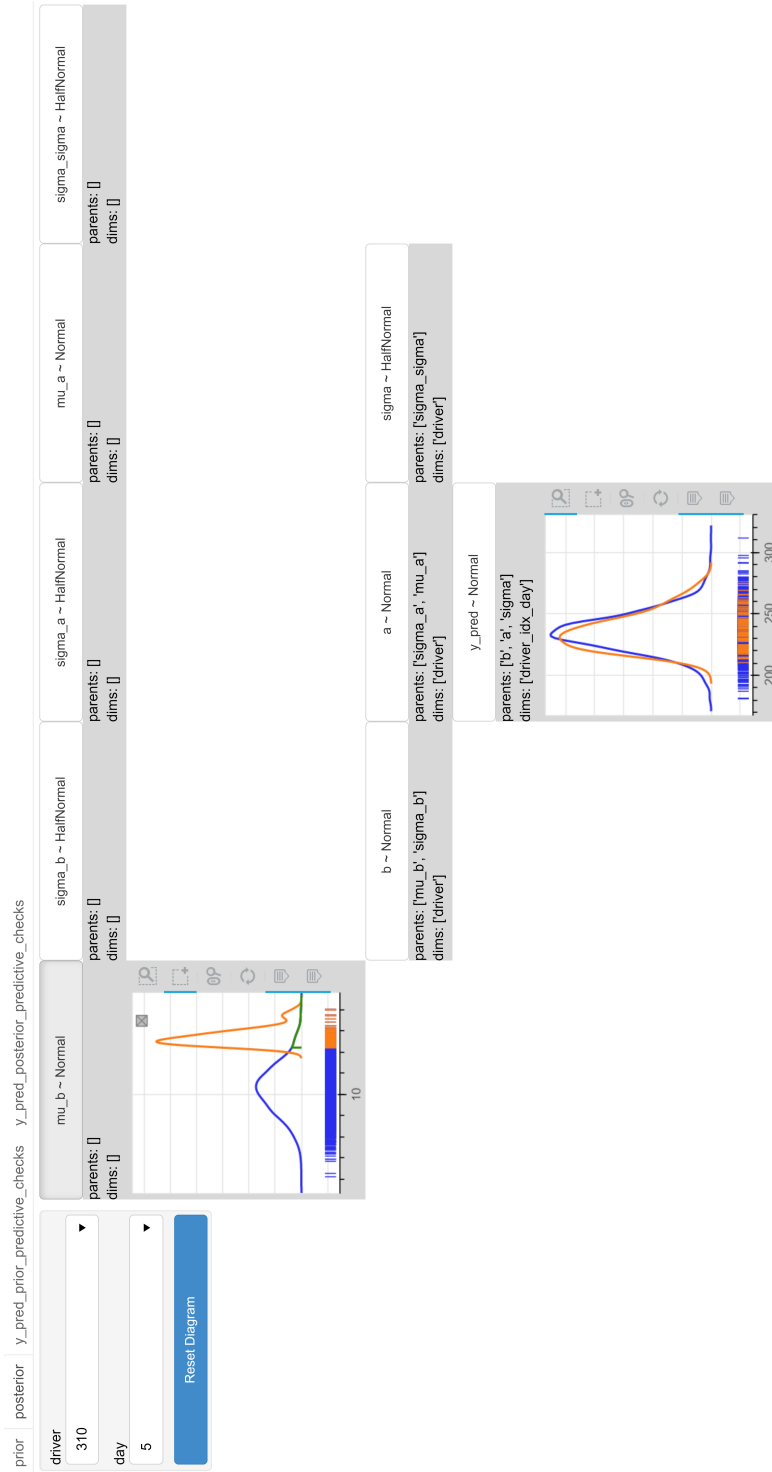


Figure 4.14: The IPME representation of the drivers' reaction time hierarchical model. We set a condition $\text{mu}_b > 12.1$. The initial and re-estimated posterior predictive distribution of the reaction times for driver 310 on day 6 are shown.

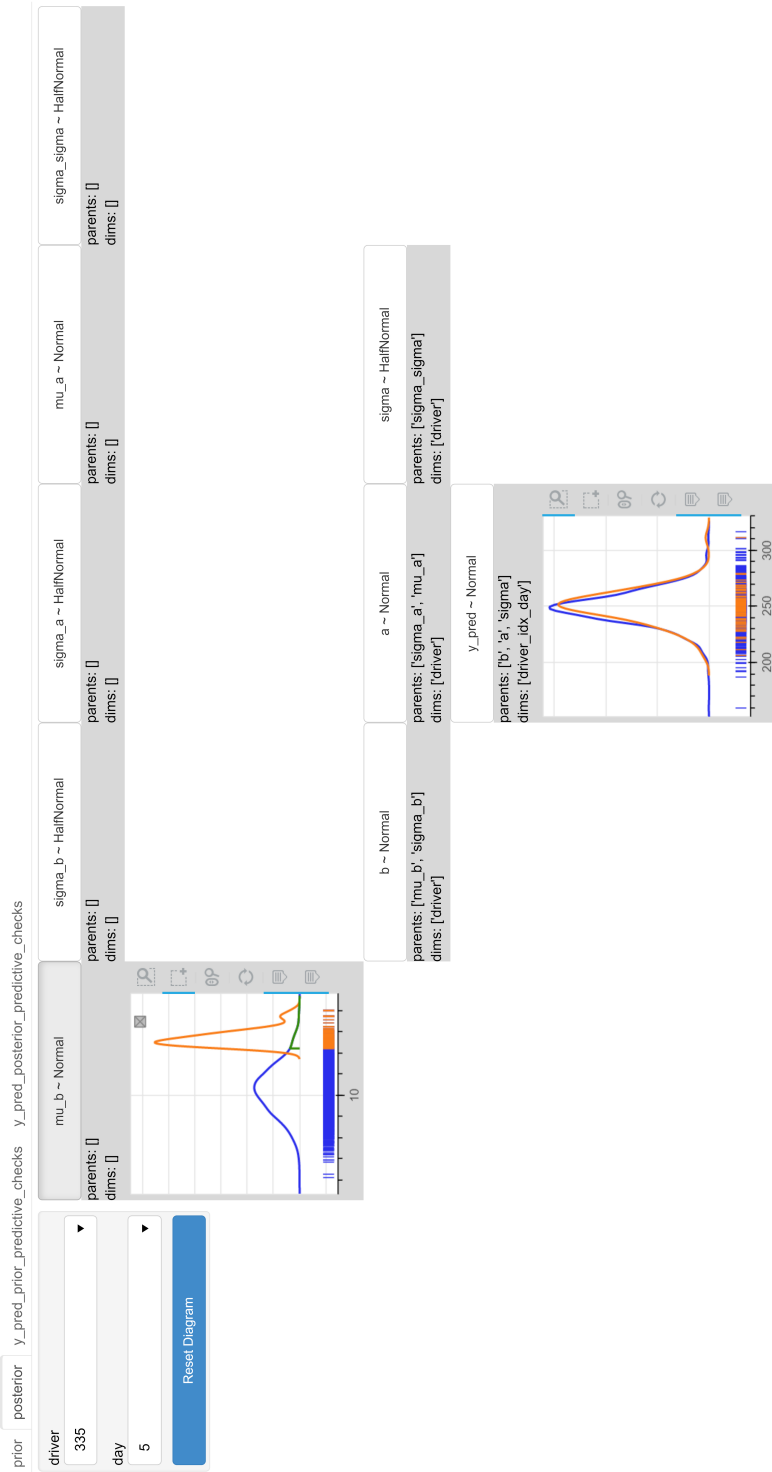


Figure 4.15: The IPME representation of the drivers' reaction time hierarchical model. We set a condition $\text{mu}_b > 12.1$. The initial and re-estimated posterior predictive distribution of the reaction times for driver 335 on day 6 are shown.

4.6.2 Stochastic Volatility

There are periods of time when assets' returns are highly variable, which implies greater uncertainty about the actual value of the returns of the assets the next day. The day returns of assets is defined as the difference in asset's price at the end of the day (P_{t-1}) in comparison to the start of the day (P_t) over the initial asset price at the start of the day, $R_t = (P_t - P_{t-1})/P_{t-1}$. Box 4.4 provides the definition of a probabilistic model to predict the returns of assets.

Box 4.4 The Stochastic Volatility Model

The models that are used to predict the assets' returns use a latent variable `volatility`. The following probabilistic statements relate the `returns` variable to the `volatility` parameter based on the specification of such a model (see more details in Appendix A.4):

$$\text{step_size} \sim \text{Exp}(\lambda = 50) \quad (4.18)$$

$$v \sim \text{Exp}(\lambda = 0.1) \quad (4.19)$$

$$\text{volatility}_t \sim \text{Normal}(\mu = \text{volatility}_{t-1}, \sigma = \text{step_size}^{-2}) \quad (4.20)$$

$$\text{returns}_t \sim \text{StudentT}(v = v, \lambda = \exp(-2 \cdot \text{volatility}_t)) \quad (4.21)$$

4.6.2.1 Model Check

By looking at the range of x-axis of the prior predictive distribution of the `returns` variable in Fig. 4.16, we realize that the model produces extremely large values of prior predictive returns for some coordinates of the indexing dimension. These values are not reasonable, especially if we think that “the total value of all goods and services the world produces is $\$10^9$.”⁹ We might need to change our prior distributions because they seem to fail to capture the prior knowledge well.

⁹https://www.pymc.io/projects/docs/en/v3/pymc-examples/examples/case_studies/stochastic_volatility.html

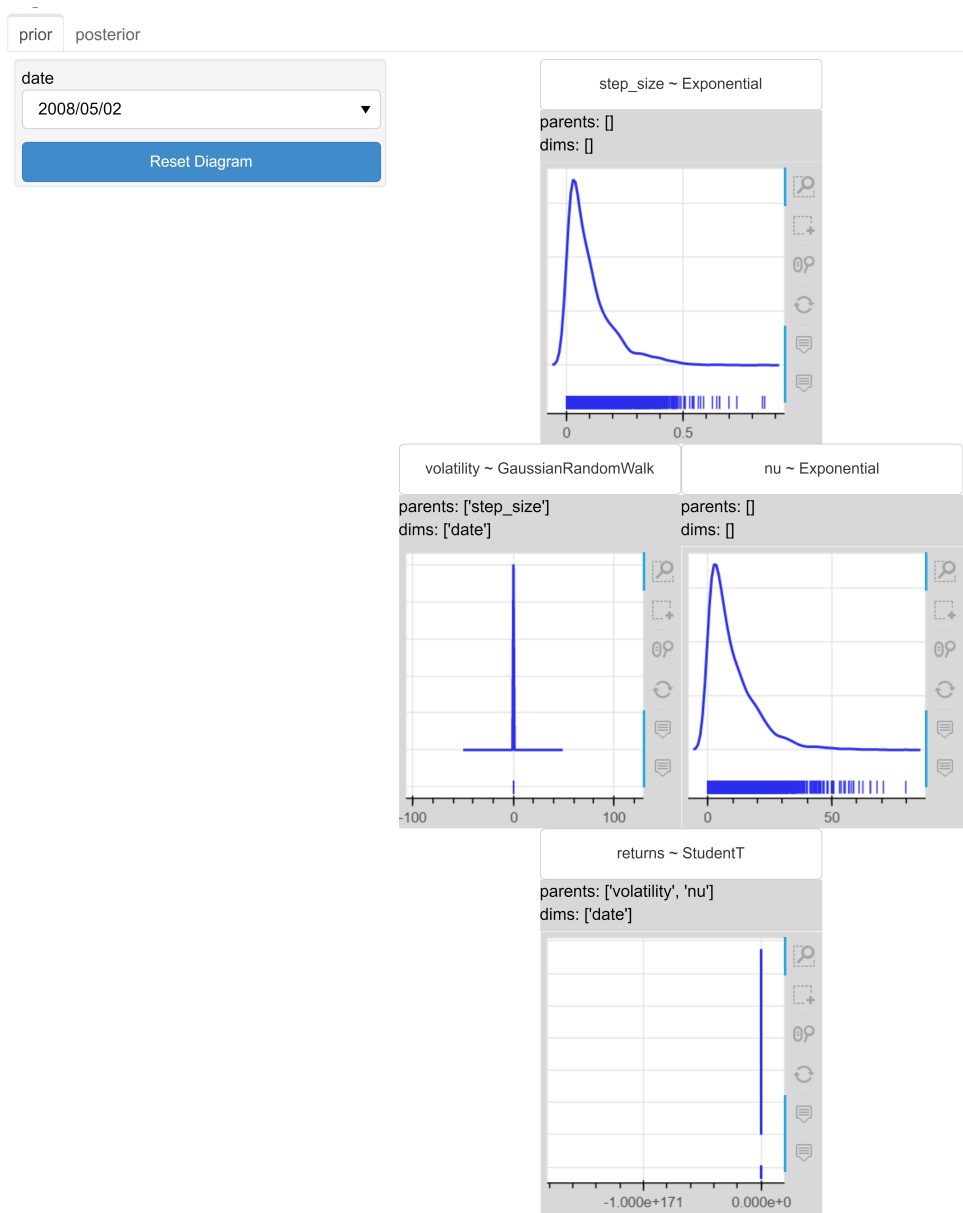


Figure 4.16: The prior IPME representation of the stochastic volatility model.

4.6.2.2 Interactivity

It is difficult to explain how `volatility` affects the predictive distribution of `returns` based solely on the model specification. For example, does increasing `volatility` mean that `returns` will increase, too? To reason about this, users would need to be familiar with the λ parameter of the student-t distribution and with the exponential transformation. Using IPME we can answer such questions by using the interactive conditioning option. When we restrict the `volatility` to values higher than -4.5 (Fig. 4.17), the predictive distribution of the `returns` variable becomes wider and therefore, there is more uncertainty about the values of the `returns`. On the other hand, when we restrict the `volatility` to values lower than -5.1 (Fig. 4.18), the predictive distribution of the `returns` variable becomes tighter and there is less uncertainty about the values of the `returns`.

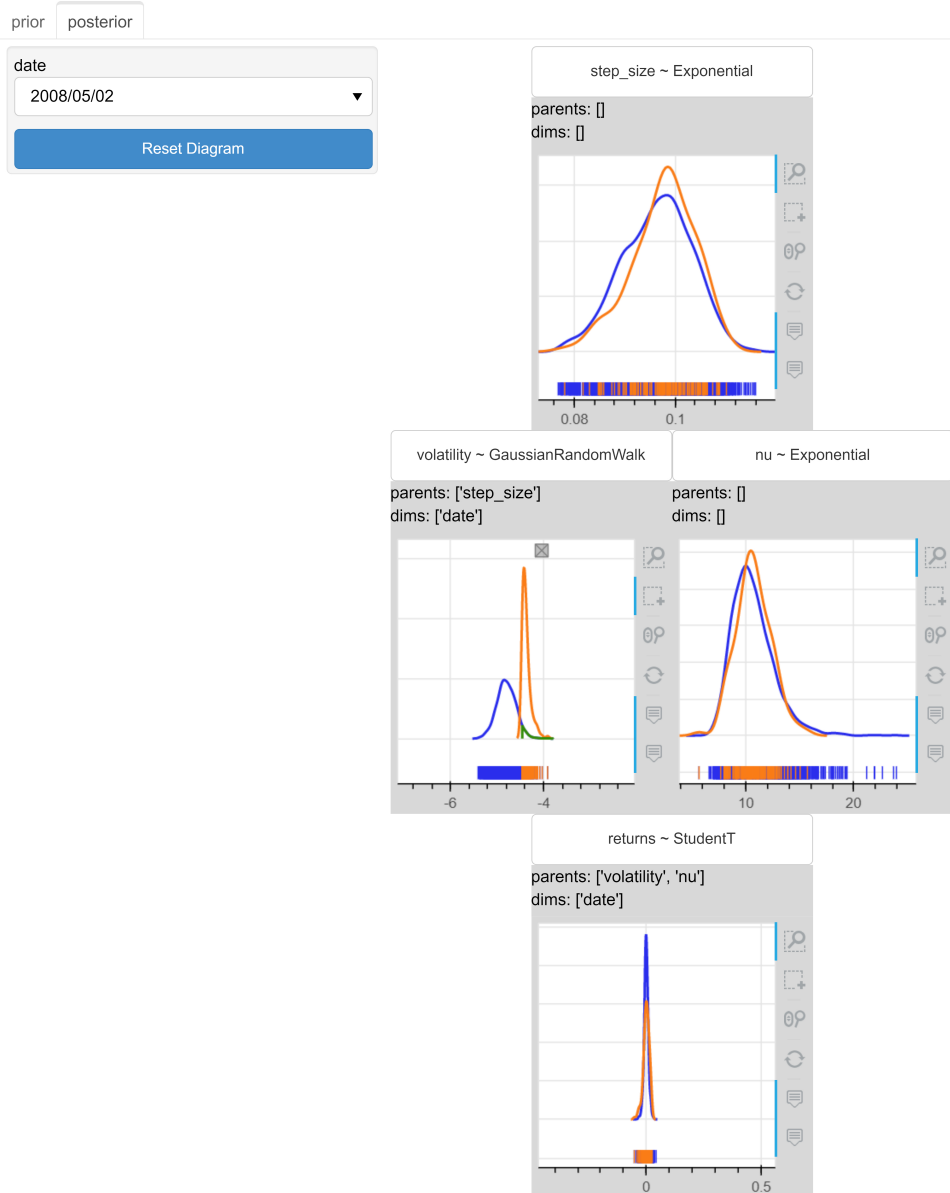


Figure 4.17: The posterior IPME representation of the stochastic volatility model. The updated posterior graphical representation after restricting the volatility to values **greater than -4.5** is shown.

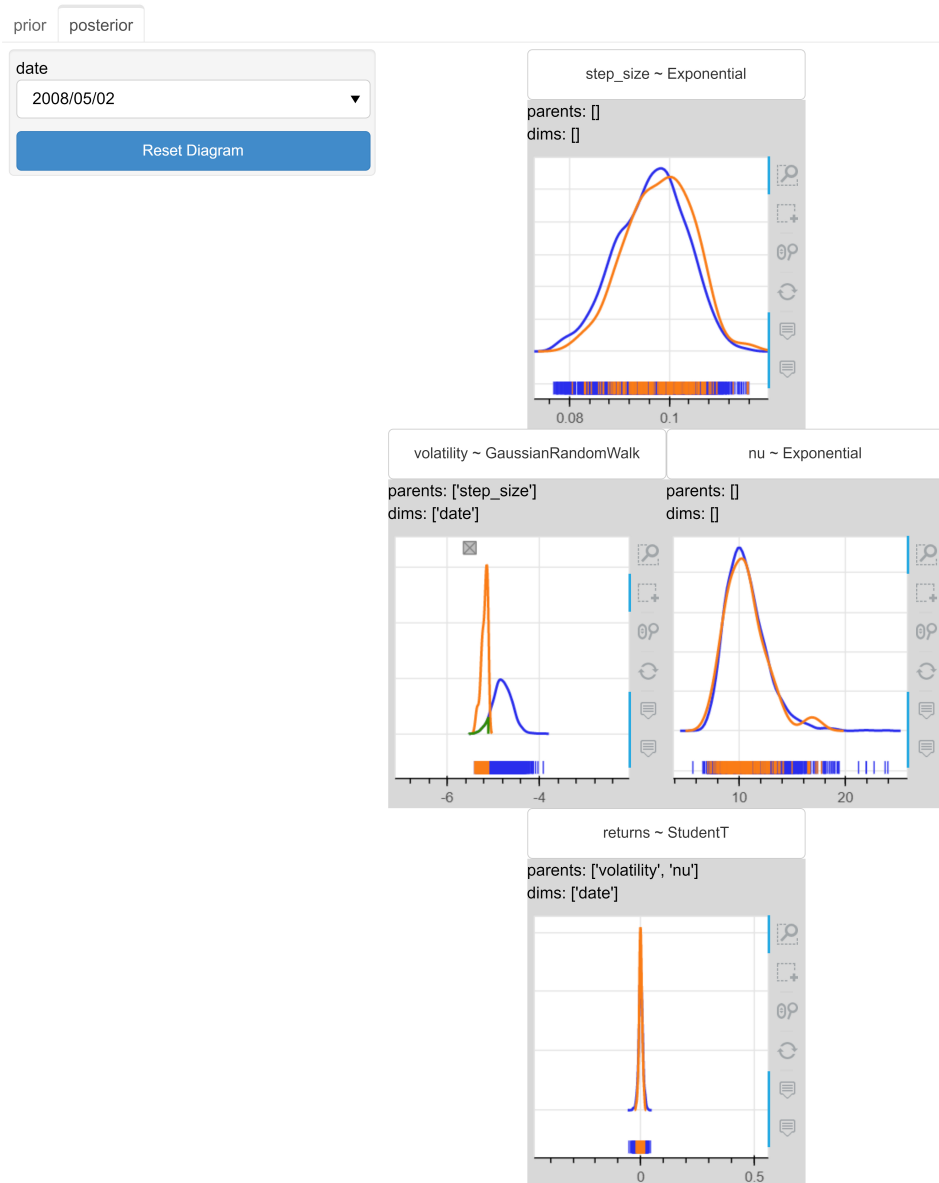


Figure 4.18: The posterior IPME representation of the stochastic volatility model. The updated posterior graphical representation after restricting the `volatility` to values **smaller than** -5.1 is shown. The greater the value of the `volatility`, the more uncertain the model becomes about the predicted values of the `returns`.

In this example we could also use interaction to observe the effect of priors' informativeness on the posteriors. The model uses a quite wide exponential distribution (with $\lambda = 0.1$) as a prior for the degrees of freedom parameter (ν) of the student-t distribution. We could set a condition on the prior distribution of (ν) to make it tighter and see how the posteriors will change. We restrict $\nu \in [21.0, 30.0]$ in Fig. 4.19(a). The posterior distribution of `volatility` and the posterior predictive distribution of `returns` on date 02/05/2008 become tighter in Fig. 4.19(b). It is obvious that the more informative a prior is, the less uncertainty the model gives to its predictions.

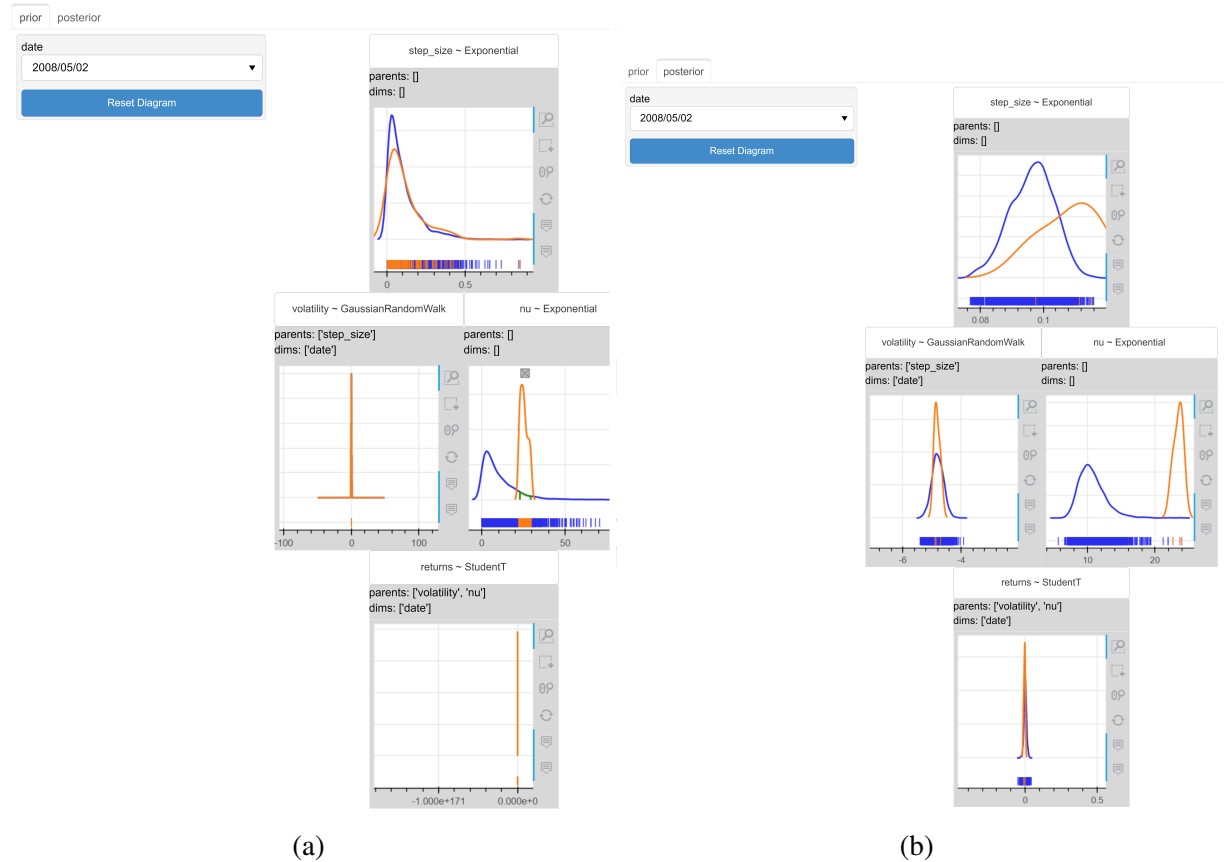


Figure 4.19: The IPME representation of the stochastic volatility model where we set the condition $\nu \in [21.0, 30.0]$. The (a) prior and (b) posterior graphical representation. The posterior distribution of `volatility` and the posterior predictive distribution of `returns` become tighter. More informative priors make the model to be more certain about its predictions.

4.6.3 Coal Mining Disasters

The coal mining disasters model in Box 4.5 models the recorded coal mining disasters in the UK from 1851 to 1962 [Jarrett, 1979]. During this period changes in the safety regulations are thought to have influenced the frequency of disasters. Thus, the model tries to predict a switchpoint, when the disasters seemed to start declining.

Box 4.5 The Coal Mining Disasters' Model

The following probabilistic statements define the probabilistic model that was created to infer the period of time when most probably a change in the safety regulations took place (see more details in Appendix A.5):

$$\text{disasters}_t \sim \text{Poisson}(\lambda = \text{rate}_t), \text{rate}_t = \begin{cases} \text{early_rate} & \text{if } t \leq \text{switchpoint} \\ \text{late_rate} & \text{if } t > \text{switchpoint} \end{cases} \quad (4.22)$$

$$\text{switchpoint} \sim \text{Uniform}(a = t_l, b = t_h) \quad (4.23)$$

$$\text{early_rate} \sim \text{Exp}(\lambda = 1) \quad (4.24)$$

$$\text{late_rate} \sim \text{Exp}(\lambda = 1) \quad (4.25)$$

where disasters_t is the number of disasters in year t , rate_t the rate parameter of the Poisson distribution in year t , switchpoint is the switchpoint, namely the year, when the change in the safety regulations occurred and there was a switch in the rate rate_t , early_rate is the rate parameter before the switchpoint, and late_rate is the rate parameter after the switchpoint.

Using IPME we can see that the posterior PMF of the switchpoint variable indicates that the switch most probably took place at some point around the year 1890 (Fig. 4.20).

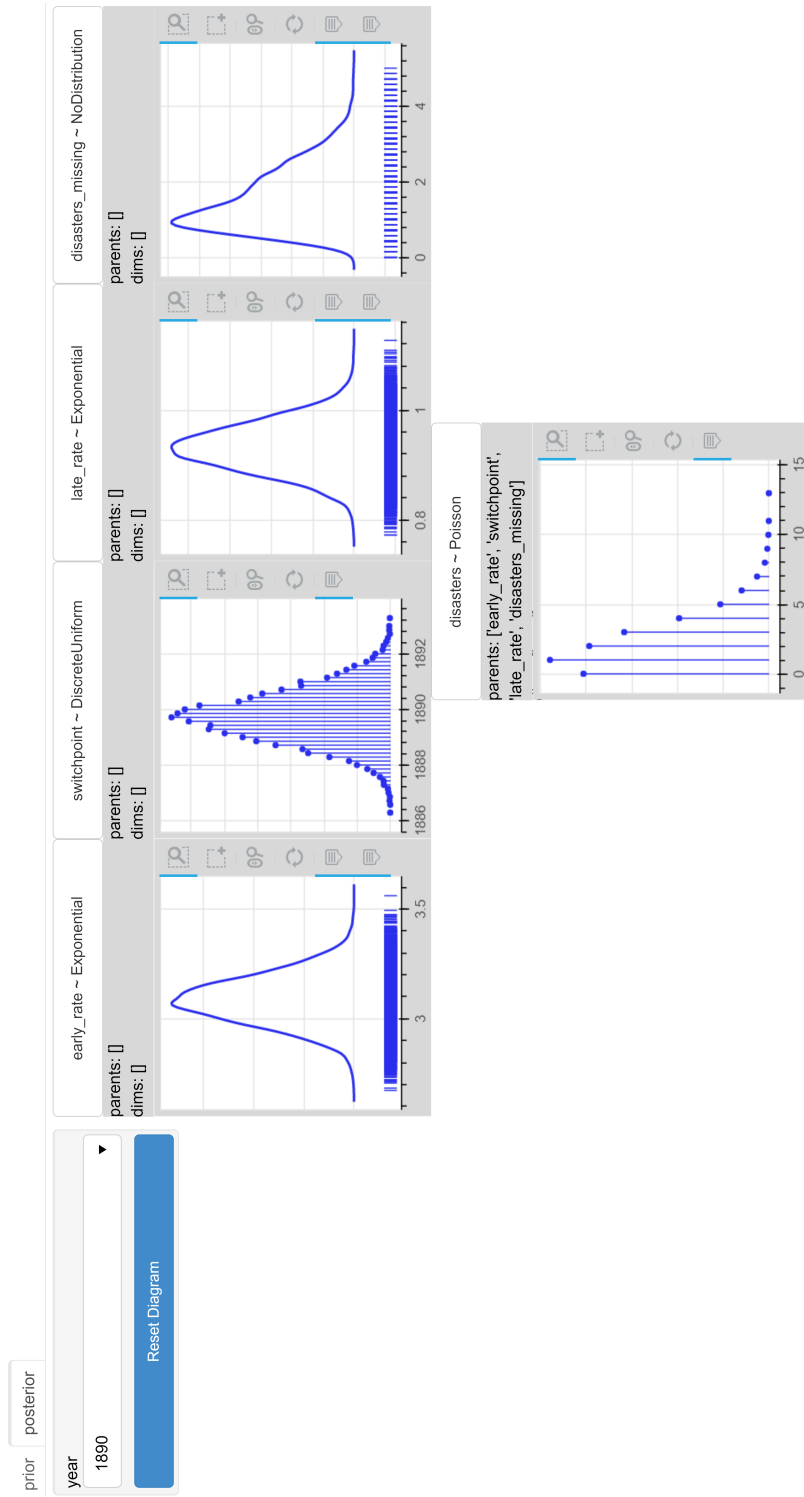
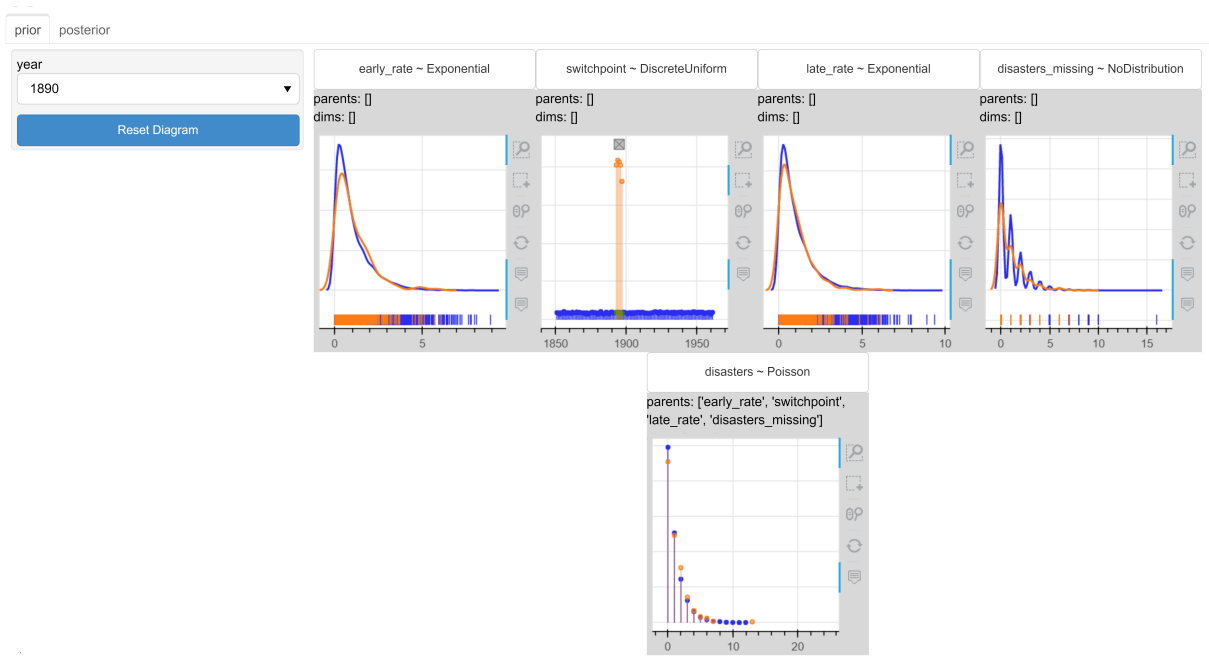


Figure 4.20: The posterior IPME representation of the coal mining disasters model.

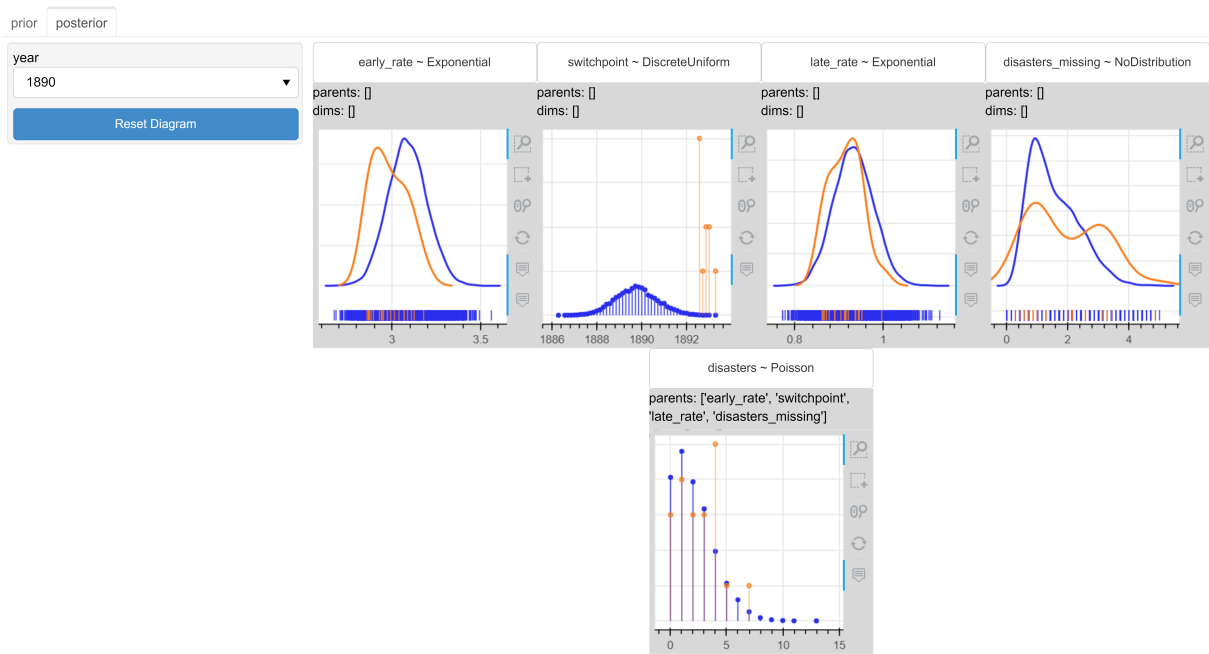
4.6.3.1 Interactivity

We could select the year 1890 from the year drop-down menu so that we can observe the uncertainty of the posterior predictions about the number of disasters in this year. Most of the probability mass is concentrated around the values of 1 or 2 disasters. What would happen to the uncertainty of the model about the number of disasters in year 1890 if the change in the regulations had happened between 1893-1897?

Fig. 4.21(a) presents how the uncertainty about the disasters' number changes a priori given this condition is applied on the `switchpoint` variable. The prior predictive PMF of the `disasters` variable becomes more dispersed, but the model does not seem to believe that the number of the disasters in year 1890 could be increased given the change in the regulations happens after 1890. Fig. 4.21(b) shows the posterior predictive PMF of `disasters` becomes more dispersed and the probability mass gets slightly shifted towards 4. The model becomes more uncertain about the number of disasters and the mean value increases slightly. This is reasonable, because if the regulations changed after 1890, the number of disasters in year 1890 is expected to increase.



(a)



(b)

Figure 4.21: The IPME representation of the coal mining disasters model. The updated (a) prior and (b) posterior graphical representation after restricting the `switchpoint` to values within the interval 1893 – 1897. The posterior predictive PMF of the number of disasters in year 1890 shifts towards 4. This is reasonable as we assumed with our conditioning that the regulations changed at a later time.

4.7 Discussion

4.7.1 Contributions

This section provides an overview of the contributions the proposed novel representation of probabilistic models, IPME, could have in the field of Bayesian modeling, reasoning, and visualization. IPME was mostly inspired by the work presented in Section 3.4. Hence, it is discussed here how IPME could enhance the *interpretability* of Bayesian probabilistic programming models in terms of *informativeness*, *transparency*, and *explainability* of the model to humans, and finally, the potential of more *trust* into the model.

IPME could be seen as an extension of the Kruschke-style diagram [Kruschke, 2015], where MCMC samples from the prior and posterior distributions are integrated into a visualisation of the parameters' distribution. This seamlessly integrates the inference results with the structural representation in a single compact representation with varying levels of abstraction to provide modelers and decision-makers with a powerful visualization tool. This increases the *informativeness* of a Bayesian probabilistic model.

A probabilistic programming model, expressed as PPL source code, can be too abstract for a user to understand or too difficult for conducting model validity checks during the model specification process. The proposed interactive graphical representation provides a broader overview of a model's structure. Modelers can at a glance observe the structure of the model and the specifications of its parameters. Variable granularity makes IPME adaptable to users' needs, skills and experience. IPME increases *transparency* of the model's structure by balancing levels of visual abstraction to avoid potentially overwhelming users.

The most important feature of IPME is interaction. This includes sub-setting, indexing/plate dimension selection and prior/posterior comparisons. Users can restrict the prior or posterior space by selecting ranges for individual parameters, can slice inference across indexing dimensions, and can quickly compare prior and posterior beliefs. Users can interactively create projections on the 2D plane of a high-dimensional entity that consists of the model's parameters, their indexing dimensions, and the prior and posterior MCMC samples. Users can explore this "multiverse" of different views of the model's inference results [Steege et al., 2016]. In this way, the proposed interactive graphical representation increases the *transparency* of the model's inference results and observed data.

IPME could be seen as a Bayesian "scrubbing calculator" [Victor, 2011c]. Users can iteratively update the restrictions of the sample space based on one or more parameters/indexing dimensions until they reach certain levels of uncertainty for a parameter or prediction of interest. This explorability allows the investigation of the sensitivity of the distribution about any parameter or the predictions of the model. Although Bayesian reasoning requires reasoning about joint distributions and conditional probabilities, which are usually prone to misconceptions and biases [Koehler, 1996; Tversky and Kahneman, 1974], IPME avoids the abstraction of distribu-

tions and numerical probabilities by allowing “explorable explanations” [Victor, 2011a]. Users develop an intuitive understanding of the mechanics of the Bayesian inference through active engagement. The proposed interactive graphical representation increases the *explainability* of the model for the users.

IPME integrates visualizations to present prior and posterior predictive checks as part of a Bayesian workflow. Prior predictive checks could provide evidence of the consistency of the model with the domain expertise and posterior predictive checks could provide evidence of the degree that the model is rich enough to capture the essentials of the true data generating process according to Betancourt [2018]. This is very important in cases where *trust* in models’ inference is questionable and modelers or decision-makers need evidence that the assumptions made led to a model that works as expected. IPME could enhance the *trust* of the users in the model’s inference results.

The last contribution is the development of a Python package that automates the creation of interactive probabilistic models’ explorers from PPLs. This is a unified tool that could become a valuable tool for users of any PPL. This tool is offered to the research community with the hope of fulfilling an existing gap in the field of Bayesian probabilistic programming models and with the hope of raising valuable feedback and possibly contribution for improving its design and features, and extending its functionalities.

IPME aims at shedding light to the abstraction of the PPL code and the conditional probabilities of Bayesian inference by actively engaging users in exploring the inference results and developing an intuition about the inherent probability space of the model’s parameters and predictions. IPME enables users to move around the “ladder of abstraction” [Victor, 2011b] starting from an abstract representation of PPL code and climbing up to a “multiverse” of concrete fully-expandable views of inference results in the pursuit of a profounder understanding of the model and the inference.

4.7.2 Limitations

This section discusses the limitations of IPME’s implementation. First of all, feeding back re-specified prior distributions and rerunning the inference is something that would be useful, but requires additional engineering to implement the feedback path to the sampler. For this reason, there might be cases where restrictions on the sample space exhaust the available samples and result in very small or empty sample sets. IPME currently does not handle these cases well. Only a single restriction in the value range per parameter is supported currently.

There are two issues connected to the presentation of the tree-like structure of the graph and the KDEs that arose during the implementation of the tool. First, because a tree level with many nodes is not uncommon, the number of nodes in each row should be restricted to avoid horizontal scrolling. If a tree level has more than some maximum number of nodes, the row is broken into two or more rows. Because this presentation spoils the tree-like feeling of the structure, the

rows that correspond to the same level of the tree were left-aligned to give the feeling of the continuation for these rows. This is analogous to typesetting text: nodes are words, which are broken at the end of lines for visual presentation, and separated into paragraphs which are semantically distinct levels, offset with vertical spacing and indentation.

The second presentation issue had to do with the scaling of the KDEs. The KDEs were not normalized so that the prior and posterior distributions are presented as valid probability distributions, namely the area under the KDE graph sums to 1. If normalization was applied, this would obscure changes in the uncertainty in the subsets of the (MCMC) sample set. However, this lack of normalization could lead to significant differences in the maximum a posteriori (MAP) estimation¹⁰ of the initial and updated KDE in cases of restricting the sample space to highly dense subsets. In these cases the overlap of the KDEs retaining their scales will present a very tiny initial KDE in comparison to the updated one and the user might need to use the “zoom wheel” tool to inspect the initial KDE.

A limitation of the work presented in this chapter is that the design of IPME was not evaluated with human users. Although Section 4.5 discusses the uniqueness of IPME in comparison to existing tools and Section 4.6 provides a series of illustrative use cases for IPME, the evaluation of IPME’s effectiveness in real user’s comprehension of probabilistic models would provide evidence about the importance of such visual representations of probabilistic models in facilitating and making Bayesian inference accessible to a wider audience. Also, although the design of IPME was guided by a set of objectives identified to support typical Bayesian inference-related tasks, the details of the design were not informed by the feedback of real users as to whether or which individual features they think are more useful or preferable.

4.7.3 Future Work

Future work in the field of creating representations of probabilistic models could involve improvement of IPME’s design or development of alternative designs to better support users in their tasks. Future work towards this direction is definitely encouraged through this work. Some indicative examples of future work that could constitute extension of the work presented in this chapter are discussed below.

IPME’s design as presented in this work meets the minimum requirements to serve its objectives. This design could be improved or elaborated as to various aspects like the navigation of the DAG or users’ flexibility in customization of visual presentation of inference results and predictive test statistics. For example, interaction could be used to highlight parent nodes when the user interacts with or hover over a specific node. Facilitating users to sub-set the coordinates of interest could be useful in case indexing dimensions are of big sizes. This would narrow down the options in the drop-down menus of indexing dimensions and make navigation of data easier. More visual primitives (Box plots, CDF plots, quantile dotplots etc.) or more predictive

¹⁰https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation

test statistics could be included in a list of pre-implemented features for users to select from and enable customization according to users' preferences and scenario requirements. Future extensions of this tool could incorporate recommendations of sets of "interesting" parameters based on the scenario to provide some sort of guidance through decision-making tasks.

The design of IPME was determined at a considerable degree by the limited flexibility offered by the existing visualization libraries that could be used as a backend (e.g., Bokeh, Panel). An alternative and possibly more natural design of the view of model's graph could use the typical presentation of nodes and edges (arrows). Although visualizing a graph in this way is possible by using various existing visualization libraries, integrating various annotations and (interactive) uncertainty visualization plots into the nodes of such graphical representations is not supported by the existing tools. Future work on developing a graphical representation of probabilistic models that is more compatible with the typical representation of graphs (with nodes and edges) would be worthwhile. The presentation of the parent-child relations would be visually conveyed directly through the arrows (the user now gets this information from the tree-like structure and the textual information included on the header of each node). Such designs could be benefited from further interactivity to adjust the granularity of the presented information: e.g., making the nodes clickable to hide or expand the sub-tree rooted at the node that is clicked. Such design options would facilitate the scalability of the representation for increased complexity of the model.

Evaluation user studies could inform the design of IPME and lead to improved and more effective representations based on feedback provided by real users or the evaluation of their performance in various tasks given different possible alternative designs.

The presented design of IPME in this chapter can only support querying the precomputed results of Bayesian inference. An alternative design could support closing the loop between the user interface and the rest of the components in a typical inference system; the observed data, model, and inference. Such an alternative design could enable users to retrieve more inference data possibly in subsets of the sampling space, update the priors, or even to set disbelief upon the presented data. In this way, we would move from exploring static inference results to exploring dynamic inference results.

For example, in IPME a user conditional query in the sample space of the model leads to the calculation of the conditional distributions in the precomputed restricted sample space and not to rerunning the inference. In a closed-loop tool each user conditional query in the sample space of the model could lead to rerunning the inference and the tool could enable exploring how the posterior is updated when the users' constraints are applied to the model as a new (prior) knowledge. This could convert IPME into a tool of "interactive Bayesian inference" in the notion of "interactive machine learning" [Kulesza et al., 2015]. Work towards this direction would most possibly require the possibility of running the inference with close to real-time performances.

4.7.4 Conclusions

A generic pipeline was presented to transform a probabilistic programming model and associated sample-based MCMC inference results into a standardized format which can then be automatically translated into an interactive probabilistic models explorer, a novel representation of Bayesian probabilistic models that fuses structural and distributional display. An initial tool was developed to render these interactive graphical representations from standard PPL definitions. This representation aims at a threefold representation effort; the representation of a collapsible tree-like structure of the model's variables and parameters to reveal internal levels of statistical or mathematical dependencies among them; the representation of each inferred parameter as a node that presents graphically the prior or posterior marginal distribution of the inferred parameter's MCMC samples; the representation of the observed variables through prior or posterior predictive distributions of the model's predictive samples. Appropriate uncertainty visualization techniques are used to graphically represent the marginal distributions.

The added value of this representation lies in the interaction. Slicing on indexing dimensions or forming conjunctive restrictions on parameters by interacting with distribution visualizations are supported. Each user interaction with the explorer triggers the reestimation and visualization of the model's uncertainty based on the users' preferences. This closed-loop exchange of responses between the user and the explorer allows the user to gain a more intuitive comprehension of the Bayesian probabilistic model.

IPME provides at-a-glance communication of a probabilistic program's structure and uncertainty of latent parameters, and allows interactive exploration of the multi-dimensional prior or posterior MCMC sample space. The representation was designed with the principles of enhancing informativeness, transparency and explainability and ultimately, the potential of increasing trust in models. PPLs have made sophisticated statistical methods available to the mainstream, but the user interface lags behind. There is enormous potential in interactive exploratory tools for probabilistic models that support the elicitation, validation and presentation of Bayesian probabilistic models.

Chapter 5

Using Interactive Conditioning for Supporting Users' Understanding of Probabilistic Models

5.1 Summary

This chapter presents the second part of this research. The focus of this part is on interactive conditioning of a probabilistic model's distribution and what effect this could have on users' comprehension of probabilistic models' structure (i.e., how the variables are related). A hybrid design of a visualization of a model's distributions is explored: the interactive pair plot (IPP). This is a scatter plot matrix of the distribution of a probabilistic model allowing IPME's interactive conditioning on the model's variables. A user study was conducted to investigate whether interactive conditioning in a scatter plot matrix of a model helps users better understand variables' relations in comparison to static scatter plot matrices and without conveying the mathematical details of these relations.

Section 5.2 discusses the purpose of this work, Section 5.3 describes the structural aspects of variables' relations and presents existing ways to visualize them, Section 5.4 presents the details of the design and implementation of IPP, Section 5.5 presents the concrete research questions investigated by the user study, while Section 5.6 details the design of the user study. Finally, Section 5.7 presents the analysis of the collected data and the results that occurred, and Section 5.8 discusses the findings, practical implications, and limitations of the user study.

5.2 Purpose

The emergence of PPLs made probabilistic modeling accessible to a broader audience. Despite growing interest, analysis methods based on probabilistic models are not widely adopted. Non-experienced researchers who conduct experiments and analyze data do not feel confident to use

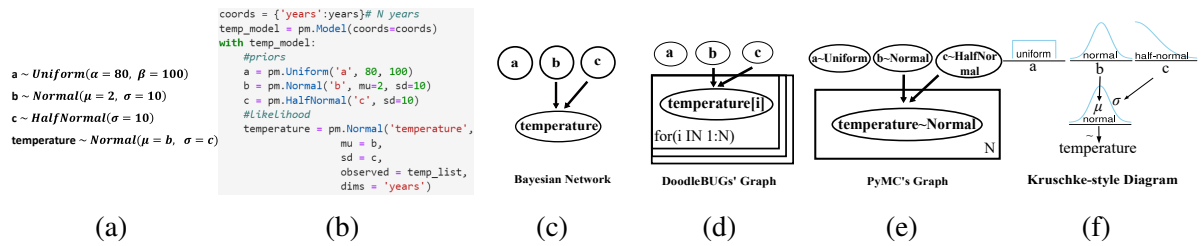


Figure 5.1: Visual representations of Model 1 of user study. **Definitions in Textual Languages:** (a) Probabilistic statements. (b) PPL code (PyMC) of model for Bayesian inference. A likelihood is defined for the observed variable `temperature` to account for the list `temp_list` of N observed temperatures for a set of `years`. **Graphs:** (c)-(f) Transcriptions of model in various graph representations.

such methods for the analysis [Hullman et al., 2019; Kay et al., 2016b] even when they have access to learning and exploration tools [Phelan et al., 2019]. Users, who need to rely on such models to do their job, might find it difficult to understand their structure. Decision-makers with moderate statistical background might make uninformed and potentially risky decisions because they cannot understand the effect of intervening on a variable upon other variables in a model. The mathematical definition of probabilistic models can be complex, unintuitive and hard to understand even for more experienced users [Sarma and Kay, 2020].

Understanding the relations among variables in a probabilistic model given definitions of the model in textual languages or graphs [Ellson et al., 2004; Koller and Friedman, 2009; Kruschke, 2015; Spiegelhalter et al., 2003] (Fig. 5.1) is very much dependent on users' statistical knowledge. For example, variable `b` in Model 1 (Fig. 5.1(a)) controls the mean value of variable `temperature`. Increasing `b`'s value would increase `temperature`'s mean value. In models where relations are governed by more complex statistical or mathematical associations, it requires good statistical knowledge to tell what the effect of a variable on others would be.

There is a need for tools to communicate variables' relations in probabilistic models more intuitively and help users to conduct their tasks without having to delve into the mathematical details of the models' definitions. For example, such tools would be useful for a clinician, who needs to understand how the dose of a medicine might affect the treatment of a disease and might need to understand how the parameters of the treatment's distribution are set by the variable of the medicine's dose. Or such tools would be useful even for model builders who might want to check and validate their models.

Variables' relations in a probabilistic model can be visualized through visualizations of variables' distribution. Scatter plot matrices present variables' pairwise distributions conveying existing correlations. IPME uses interactive conditioning implemented as a brushing-and-linking interactivity on KDE plots to enable a form of "sensitivity analysis" of the variables and reveal their relations.

Various visualization designs were explored in the literature to facilitate reasoning about unintuitive mathematical concepts like uncertainty [Correll and Gleicher, 2014; Hullman et al.,

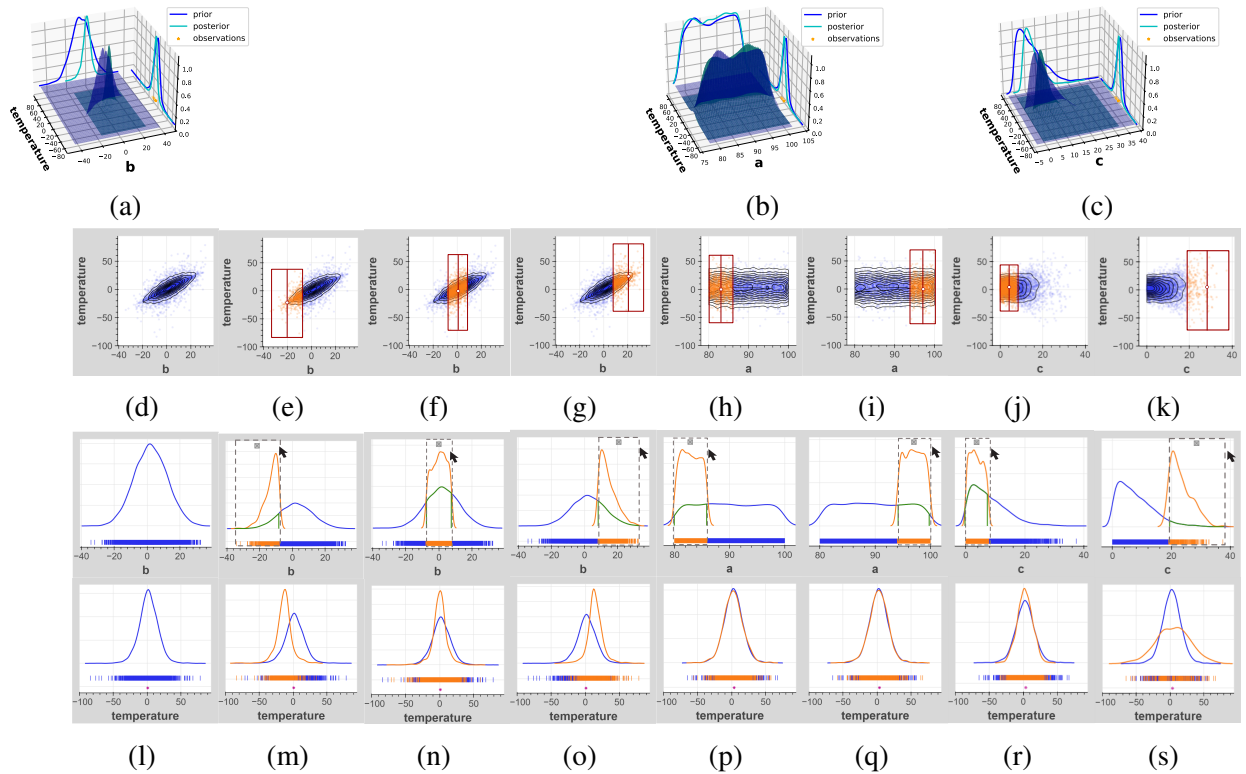


Figure 5.2: Visualizations of variables’ relations in Model 1 of user study. **Joint & Marginal Distributions:** (a)-(c) The prior and posterior joint (3D surface plots) and marginal distributions (line plots on cube faces) of variables `temperature`, and `b`, `a`, or `c`, respectively. The yellow stars represent the observations in `temp_list`. **Scatter Plots:** (d)-(k) Samples and contours of variables’ pairwise prior joint distributions. Conditioning facilitates the interpretation of scatter plots’ shape. For example, conditioning on `b` in sequential increased ranges in (e)-(g), increases the mean value (white dot) of `temperature`’s distribution. **Interactive Conditioning with IPME:** (l)-(s) IPME-like representation. Interactive conditioning is applied on the prior marginal distributions of `b`, `a`, or `c` and the conditional marginal distributions are drawn (in orange).

2015; Kay et al., 2016a] and Bayes’ rule [Brase, 2009; Cole, 1989; Micallef et al., 2012; Mosca et al., 2021; Ottley et al., 2012, 2016; Tsai et al., 2011]. Brushing-and-linking [Becker and Cleveland, 1987; McDonald, 1982; Newton, 1978], although it is often present in scatter plot matrices to help with the exploration of multidimensional data, is rarely evaluated for its efficiency with real test subjects [Martin and Ward, 1995; Sankaran and Holmes, 2018].

This work explores a “hybrid” design as to its ability to communicate variables’ relations in a probabilistic model; the *interactive pair plot (IPP)* (the code can be found in [Taka, 2020b]). This is a classical scatter plot matrix that integrates IPME’s interactive conditioning with some additional sample highlighting (i.e., brushing). IPP is different from IPME in that it presents the pairwise distribution of a model’s variables so that any correlations become explicitly glanceable through the scatter plots. IPP is also different from the classical scatter plot with brushing-and-linking in that it enables users to explore the “sensitivity” of the parameters’ or predictions’ distribution to the values a certain parameter in the model might take (the conditional marginal

distributions of the model's variables are calculated and drawn as it will be explained in Section 5.4). This is achieved through the feature of IPME's interactive conditioning.

IPP offers more flexibility in the exploration of a probabilistic model's distribution than the classical scatter plot matrices, which constitute the most common way of presenting the pairwise distributions (and correlations) of a probabilistic model's variables. This is due to the interactive conditioning option. In existing literature there is contradictive evidence in regard to the benefits of interaction for users when this is used in visualization. To investigate whether there is added value to scatter plot matrices attributed to interactive conditioning an evaluation user study with human participants was conducted in which IPP was used as a visualization instance.

More concretely, the user study investigated whether users make better inferences about variables' relations presented in a scatter plot matrix when they use interactive conditioning in comparison to simply observing a static scatter plot matrix. The focus was on which levels of detail of variables' relations (Section 5.3.1) and which probabilistic model designs (e.g. hierarchical structures, complex parameterizations of variables' distributions) (Section 5.6.1) interactive conditioning can be more beneficial for. The concrete research questions that were investigated are the following: **RQ1: Does interactive conditioning when used on pair plots help users understand the structure of probabilistic models more accurately, faster, and with more confidence? RQ2: Are there levels of detail of variables' relations or model designs for which interactive conditioning is beneficial?** Participants' accuracy, response times, and confidence was measured.

This user study provides useful evidence as to what extent (level of detail of relations or model designs) interactive conditioning of a probabilistic model's distribution could facilitate people's comprehension of model's structure without conveying details about the mathematical definition of the model. These types of investigations are important to guide research in visualization of probabilistic models and inference results towards a direction to first, support users in tasks in which they cannot or do not need to comprehend the mathematics, and second, make the mathematics more intuitively explainable through the visualization. In the user study presented in this chapter, IPP is evaluated as to its ability to convey variables' relations in a probabilistic model through visual means like graphs and interactive conditioning independently of the complexity that might govern these relations.

5.3 Relations of Probabilistic Models' Variables

5.3.1 A Hierarchy of Variables' Relations

Two random variables in a probabilistic model are related when one is used for setting the distribution's random-valued (non-fixed) parameters of the other. In this work three levels of detail, from lowest to highest, are identified to characterize such relations among variables in a

model. These levels are used in the user study conducted to evaluate users' understanding in the different levels of detail as it will be explain in Section 5.5.

- L1 **Existence.** Are variables related? Variable `temperature` is related to `b` and `c`, but not to `a` in Model 1 (Fig. 5.1(a)).
- L2 **Polarity.** What is the sign (positive or negative) of the polarity of the effect one variable has on another? The μ parameter of `temperature`'s distribution increases when the value of `b` increases in Model 1 (positive polarity), while it would decrease if it was set equal to $-b$ (negative polarity).
- L3 **Quantification.** How are variables related? A relation is quantified by the specific statistical associations (which parameters of a distribution are affected by a variable) and formula (mathematical transformation or equation) that sets it. Variable `b` sets `temperature`'s μ parameter through a simple assignment. A transformation $\exp(b)$ or an equation $5 - 2 \cdot b$ could be more complex ways to do so.

The following subsection discusses how variables' relations could be represented through the existing visual representations, and explains which of these levels of detail could be retrieved from these representations, and how.

5.3.2 Visualization of Variables' Relations

Textual Language Definitions. Variables' relations (L1) and their quantification (L3) are retrievable through an at-a-glance observation of the model's definition in probabilistic statements (Fig. 5.1(a)) or PPL code (Fig. 5.1(b)). Retrieving the polarity of variables' effects (L2) is dependent on the ability of the user to interpret the mathematical details.

Graphs. Graphs (Fig. 5.1(c)-(f)) can hide the mathematical details of probabilistic models, while preserving some structural information. The *edges* (directed arrows) from one variable to another indicate the direction of the relation between the two variables. Given a graph, users could view relations among variables (L1) at a glance (through the existence or absence of edges). In the case of the more informed graphs like Kruschke diagrams, users could even observe the exact statistical associations or mathematical equations (L3). But inferring the polarity of the effect of a variable on other variables (L2) is still very much dependent on the ability of the users to understand the mathematical details.

To convey relations' polarity (L2) visually, we need to incorporate representations of the model's real-data uncertainty.

Joint & Marginal Distributions. A model's joint distribution is multivariate. The pairwise joint or marginal distributions of the variables could be represented (Fig. 5.2(a)-(c)). While KDE plots are a common way of representing marginal distributions, 3D surfaces are rarely used for

representing the pairwise joint distributions especially in the context of probabilistic modeling. Contour and scatter plots are more commonly used for this instead. There are various existing visualization libraries to create such representations for Bayesian analysis (ArviZ [Kumar et al., 2019], bayesplot [Gabry and Mahr, 2020], tidybayes [Kay, 2020], shinystan [Stan Development Team, 2017]).

Variables' relations (L1), their polarity (L2) and aspects of their quantification (e.g. statistical associations) (L3) are conveyed by the shape of scatter and contour plots. Conditioning could help interpreting the shapes of these plots in regards with these three aspects of variables' relations, and retrieving them from KDE plots alone, as it is explained below.

Scatter Plots. The shape of a scatter plot of samples and contours representing a 2D distribution can reveal relations between the two variables. For example, the well-elongated elliptical shape of the scatter plot of `temperature` and `b` in Fig. 5.2(d) implies the existence of a relation, while the rectangular shape of the scatter plot of `temperature` and `a` the absence of a relation (L1). In the first case, increasing values of `b` lead to higher mean value of the distribution of `temperature`, while in the later, increasing values of `a` do not affect the distribution of `temperature`. The shape of the scatter plot reveals the polarity of the relation (L2) and the statistical associations (`b` controls the μ parameter of `temperature`'s distribution) (L3). The effect becomes more evident if the sample set is divided into subsets of samples for sequential increasing ranges of `b` (Fig. 5.2(e)-(g)).

Interactive Conditioning with IPME. Interactive conditioning could also be used to convey information about variables' relations through KDE plots. In IPME the interactive conditioning of the marginal distributions and the presentation of the conditional marginal distributions of the variables are possible. For example, comparing the marginal distribution of `temperature` (drawn in blue) with its three sequential conditional marginal distributions (drawn in orange) in Fig. 5.2(m)-(o) while conditioning on `b` in three increasing and sequential ranges, we could infer a relation such that increasing values of `b` lead to higher mean value of the distribution of `temperature` (reveals L1, L2, and statistical associations in L3). Conditioning is applied by the user by dragging a fixed-height and variable-width selection box in KDE plot corresponding to the conditioning variable.

Scatter Plot Matrix. A Scatter plot matrix (or pair plot) presents the pairwise joint and marginal distributions of a model's variables. ArviZ offers the ArviZ Point Estimate Pairplot (APEP) [ArviZ Point Estimate Pairplot], which presents variables' joint samples and contours of the pairwise distributions on the bottom corner of the matrix and the KDE plots of the marginal distributions on the diagonal. Scatter plot matrices usually offer selection tools for applying data filtering (conditioning). In this work the interactive pair plot (IPP) is introduced; that is an interactive scatter plot matrix like APEP that incorporates IPME's interactive conditioning on the KDE plots to present the conditional marginal distributions. Section 5.4 presents IPP in more detail.

IPME and IPP implement interactive conditioning as a *brushing-and-linking* effect. Brushing-and-linking is an interactive approach usually used on static visualizations of multivariate data like scatter plot matrices. This method is useful in many tasks like analyzing subsets of multivariate [Elmqvist et al., 2008; Martin and Ward, 1995; Nguyen et al., 2016, 2020] or hierarchical [Sankaran and Holmes, 2018] data which could not easily be conducted through static visualizations. The added value of brushing-and-linking to static visualizations is rarely the main focus of evaluation user studies. For example, Nguyen et al. [Nguyen et al., 2020] found that an interactive version of a scatter plot visualization improved the accuracy of users in data exploration compared to static versions of it. The effect of brushing-and-linking alone could not be evaluated through this study design because the provided interactivity ranged from choosing the number of plot panels to brushing-and-linking.

A user study was designed as part of this work to investigate whether users can infer structural information about probabilistic models presented in scatter plot matrices and whether there is an effect of interactive conditioning in scatter plot matrices on users' performance on inferring this information. A visualization-only design was followed for a clearer experiment design; the questions included the visualization and no textual or mathematical (e.g. probabilistic statements) description of the model to exclude any effect on participants' performance (to identify structural relations) from other sources of information other than the scatter plot matrix and interactive conditioning.

Through this user study, the intention is to collect evidence about how well the distributional information of probabilistic models shown in scatter plot matrices and explored through interactive conditioning is understood and interpreted by participants. These types of experiments are especially important to understand to what extent such designs of exploratory visualizations could support tasks without necessarily to communicate the mathematical details of the model and expect lay users to understand them. For example, a decision-maker with poor statistical background, who needs to have a good understanding of how one parameter affects another or how sensitive one parameter is to various values of another parameter, does not have to know the mathematical details of the model as long as they have an informative representation of it.

5.4 Interactive Pair Plot

IPP (Fig. 5.3) was designed by combining elements from the designs of APEP and IPME visualizations, and adding some extra highlighting. The design of the APEP was replicated in terms of the outlook of the pair plot. The scatter and contour plots of variables' pairwise joint samples and distribution are presented on the columns and rows of the matrix's lower triangle, and the KDE plots of the variables' marginal distributions on the diagonal.

IPP was built on IPME's framework inheriting its design elements (e.g., plot's style and attributes like the grey background, color palettes, rug plots' inclusion of variables' samples

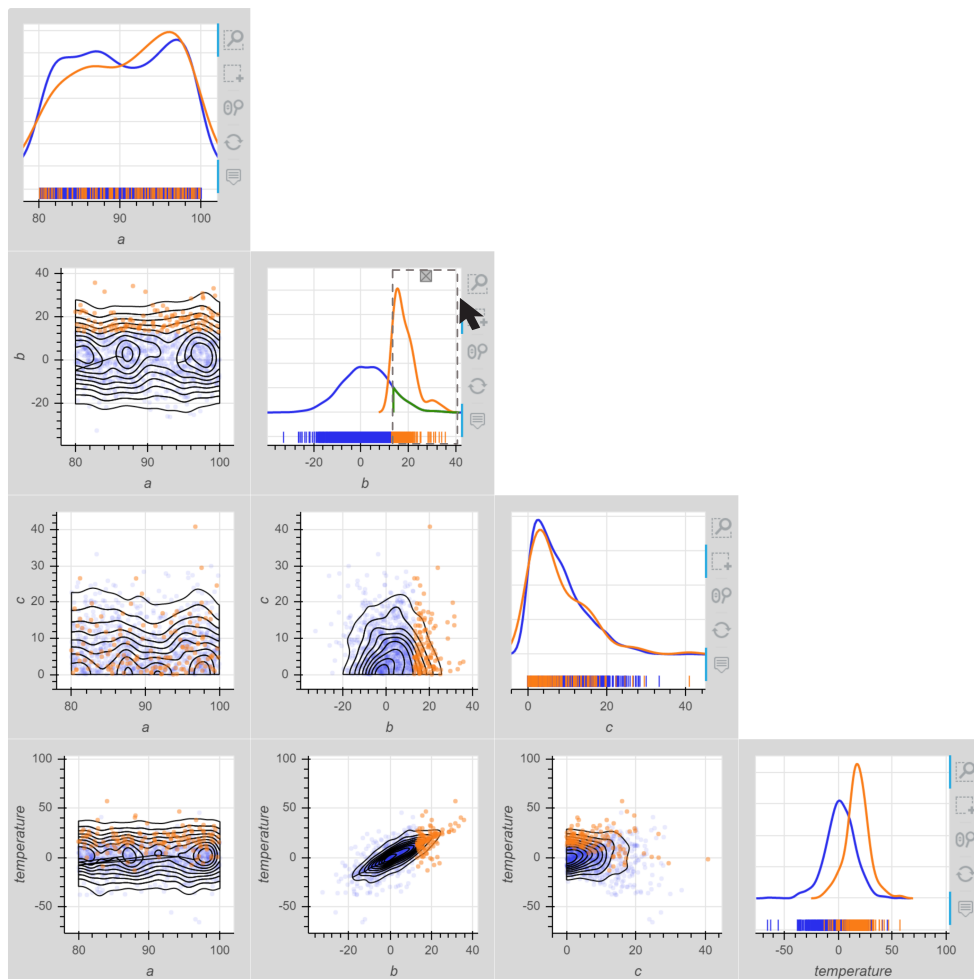


Figure 5.3: IPP of Model’s 1 variables. A selection box is dragged and drawn on the KDE plot of variable b restricting its range to $[12 - 40]$. The conditional marginal distributions of the variables are drawn (in orange) in the KDE plots on the diagonal and the samples in the restricted sample space are highlighted (in orange) in the scatter and rug plots.

on the KDE plots, side interactive toolbar, and the interactive conditioning’s mechanism and design) and limitations (e.g., inflexibility in rerunning online (prior) sampling or inference to get more samples in sub-ranges of model’s sample space with few or no samples, or applying multiple conditions on a single variable).

The interactive conditioning’s design in IPP replicates IPME’s corresponding brushing-and-linking interactivity. The KDE plots (on the diagonal) can be interactively conditioned by dragging and drawing a fixed-height and variable-width selection box (brushing) and the KDE and rug plots are updated (linking) with the KDEs of the conditional distributions being drawn and samples in the restricted sample space being highlighted in orange exactly like in IPME. The choice of the selection box was limited by the offered options of the Bokeh visualization library used in the backend. The interactive conditioning’s design in IPP was enhanced by adding the highlighting of the joint samples in the scatter plots (linking) in orange color for consistency. The intention was to replicate the typical linking effect encountered in scatter plot matrices. A

video demonstrating IPP’s interactivity can be found in [Taka, 2020a].

Both the discrete (scatter plots, rug plots) and continuous (KDEs, contours) representations of model’s distribution encountered in APEP and IPME were kept in IPP to complement each other. The contours and KDEs illustrate how the density of the samples change with the value of the variables. Without these plots, identification of high probability density value ranges would be difficult especially in cases of high-density sampling that creates visual overlaps of samples. The scatter plots and rug plots of samples provide a discretized form of the continuous representations, which according to existing literature could better support people’s reasoning for uncertainty [Hullman et al., 2015; Kay et al., 2016a].

IPP’s API considers subsets of variables of interest to deal with the quadratic scaling in area of the scatter plot matrix with the number of variables. This would be especially useful, for example, when users might want to inspect only an aspect of complex models with many variables or deep-hierarchy structures.

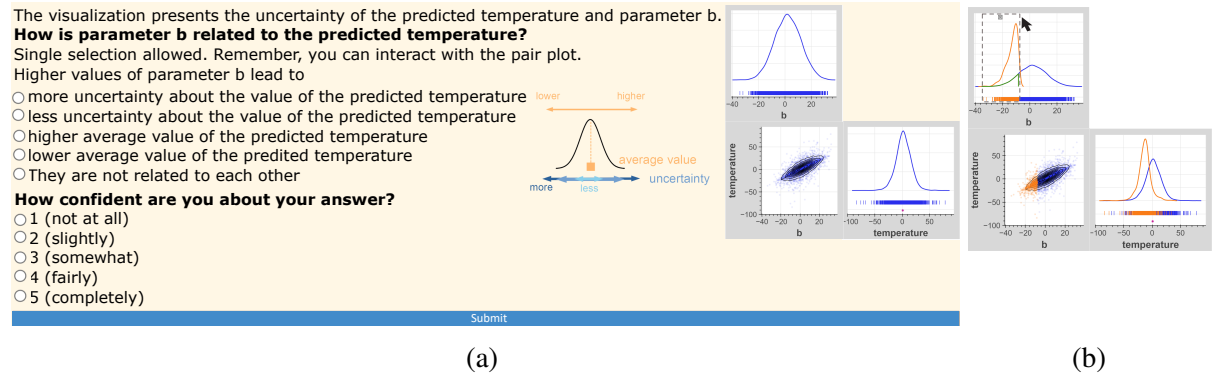


Figure 5.4: (a) Task τ_3 (Model 1 - T2) of user study. Participants in SG were shown a static pair plot. (b) The interactive pair plot participants in IG were shown instead. Both pair plots showed the minimum necessary subset of model’s variables.

5.5 Research Questions, Tasks & Conditions

RQ1: Does interactive conditioning when used on pair plots help users understand the structure of probabilistic models more accurately, faster, and with more confidence? The aim was to investigate how efficient interactive conditioning is in the comprehension of probabilistic models and when it can be beneficial. **RQ2: Are there levels of detail of variables’ relations or model designs for which interactive conditioning is beneficial?** Three types of tasks were determined in the user study, T1, T2, and T3, each accounting for the L1, L2, and L3 level of detail, respectively. Various model designs were explored as described in Section 5.6.1. Table 5.1 summarizes the study’s tasks and models.

A between-subject user study with two conditions was designed. The two conditions were the static pair plot and the interactive pair plot (IPP). Participants in both conditions were view-

Table 5.1: Summary of probabilistic models and tasks used in the user study.

Model	Graph	Task id	T	Question
<p>Model 1</p> <p>$a \sim \text{Uniform}(\alpha = 80, \beta = 100)$ $b \sim \text{Normal}(\mu = 2, \sigma = 10)$ $c \sim \text{Half-Normal}(\sigma = 10)$ temperature $\sim \text{Normal}(\mu = b, \sigma = c)$</p>		t1	T1	Which of the parameters a, b and c are related to temperature?
		t2	T2	How is parameter a related to temperature?
		t3	T2	How is parameter b related to temperature?
		t4	T2	How is parameter c related to temperature?
		t5	T3	How would you describe the effect of parameters a, b and c on temperature?
<p>Model 2</p> <p>$a \sim \text{Normal}(\mu = 0, \sigma = 10)$ $b \sim \text{Half-Normal}(\sigma = 10)$ $c \sim \text{Half-Normal}(\sigma = 20)$ random_number $\sim \text{Uniform}(\alpha = a - c, \beta = a + c)$</p>		t6	T1	Which of the parameters a, b and c are related to random_number?
		t7	T2	How is parameter a related to random_number?
		t8	T2	How is parameter b related to random_number?
		t9	T2	How is parameter c related to random_number?
		t10	T3	How would you describe the effect of parameters a, b and c on lower_bound?
		t11	T3	How would you describe the effect of parameters a, b and c on upper_bound?
<p>Model 3</p> <p>$c \sim \text{Normal}(\mu = 100, \sigma = 150)$ $e \sim \text{Half-Normal}(\sigma = 150)$ $f \sim \text{Normal}(\mu = 10, \sigma = 100)$ $g \sim \text{Half-Normal}(\sigma = 100)$ $h \sim \text{Half-Normal}(\sigma = 200)$ $a_i \sim \text{Normal}(\mu = c, \sigma = e)$ $b_i \sim \text{Normal}(\mu = f, \sigma = g)$ $\text{sigma}_i \sim \text{Half-Normal}(\sigma = h)$ $d \sim \text{Normal}(\mu = 0, \sigma = 10)$ reaction_time_i $\sim \text{Normal}(\mu = a_i + \text{day} \cdot b_i, \sigma = \text{sigma}_i)$</p>		t12	T1	Which of the parameters a, b, c and d are related to reaction_time?
		t13	T1	Which of the parameters b, c and d are related to a?
		t14	T2	How is parameter a related to reaction_time?
		t15	T2	How is parameter b related to reaction_time?
		t16	T2	How is parameter c related to reaction_time?
		t17	T2	How is parameter d related to reaction_time?
		t18	T3	If reaction_time, a and c lie on a graph, what is the structure of the graph?
		t19	T3	How would you describe the effect of parameters a, b and day on reaction_time?

ing the same pair plot designed as described in the previous section, but participants in the static condition were not able to use the interactive conditioning. Fig. 5.4(a) presents a T2 task of the user study with a static pair plot shown to participants in the **static group (SG)** and an interactive pair plot (Fig. 5.4(b)) shown to participants in the **interaction group (IG)** instead.

Participants had to look at the static scatter plot matrices (in SG), or interact with them and

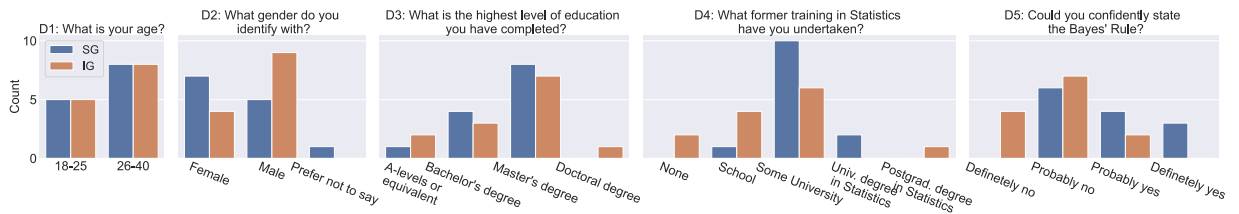


Figure 5.5: **Demographic statistics** of participants in the user study. Both groups of participants (SG and IG) comprised of more older participants (D1). There was a slight gender imbalance between the groups with IG having more males and SG more females (D2). The educational background was generally well-balanced between the groups (D3), while participants in SG had a slightly higher former training in Statistics (D4, D5).

look at the additional highlights (in IG) to perform the tasks. For example, participants in SG could determine the direction of relation between `b` and `temperature` in task `t3` (Fig. 5.4(a)) based on the shape of the scatter plot. Participants in IG could use interactive conditioning on increasing ranges of `b` and observe the changes in the highlighted visual elements of variable `temperature`. The fourth training video used in the user study (more about training in Section 5.6) presents a demonstration of task examples and how they could be answered by each group based on the presented visualization.

This design of the pair plot allows a fair comparison of the two conditions in regards with the amount of presented information. The changes in the interactive condition refer to two new types of visual elements added to the static visualization; firstly, a highlighting of visual information already existing in the static case (of dots in the scatter plots or vertical lines in the rug plots), and secondly the inclusion of the conditional marginal distributions in the KDE plots. These visual additions are informationally equivalent to scatter plots as the first does not insert new information, and the second represents existing information (in scatter plots' shape) in a different format (marginal distribution).

5.6 User Study's Design

The study was approved by the institution's ethics review board (approval number: 300200319) and conducted online. It consisted of three parts; **training**, **tasks**, and **demographic questions**. The training consisted of 4 videos (find links in Appendix C.1) presenting the aim and structure of study, an introduction to basic probabilistic concepts, an explanation of assigned version (static or interactive) of visualization, and some example tasks.

The study tasks were split into three parts, each corresponding to a different probabilistic model of increasing complexity. A set of questions of all three task types (Ts) was created for each model. Table 5.1 presents a summary of the 3 models and 19 tasks used in the study. Appendix C provides a detailed description of the models and screenshots of all tasks. The models were presented in increasing complexity and the tasks in increasing level of structural

detail, and in the same order to all participants (from t_1 to t_{19}). The version of the scatter plot matrix was only varying among participants.

All questions were multiple-choice. Multiple selections were allowed for the T1 tasks, and single selection for the rest. Each available option was graphically illustrated in the cases of T2 and T3 questions. Participants' confidence was input in a five level Likert scale.

At the end participants recorded their age, gender, highest educational level completed, former training in Statistics and knowledge of Bayes' rule. Fig. 5.5 presents participants' demographic statistics. 26 participants were randomly assigned in the IG and SG (13 participants in each). Participants were recruited through mailing lists and social media of the institution and personal contacts without any requirement regarding their statistical background. They each received a £10 worth online shopping voucher as a compensation.

5.6.1 Task Models' Design

The intention was to include different models of increasing complexity in the study. That was achieved by increasing the number of variables used for setting parameters of the observed variable's distribution in each model, combining various mathematical operations (+, -, ·) for the assignment of distributions' parameters, and the use of hierarchy (Model 3). The aim was to include both typical (Model 1 and 3) and more exotic (Model 2) model designs to account for any possible prior familiarity of participants with certain statistical structures, and the use of a variety of distribution types.

Model 1 was the simplest probabilistic model used in the user study and is a typical one; a normal distribution for the observed variable with the mean and variance being directly set by two other (latent) parameters of the model.

Model 2 used a parameterization to set the bounds of the observed variable's uniform distribution through a deterministic transformation: $\alpha = a - c$ and $\beta = a + c$. This parameterization broadens the visual effects that can be explored. The combination of a uniform (temperature) and normal (a) or half-normal (c) distribution creates unusual shapes of the pairwise scatter plots. The interpretation of the changes in the conditional marginal distribution while interacting is different in this model in comparison to previous model, because it is the bounds of the distribution that change here.

Model 3 was the most complex model representing a typical hierarchical linear regression model with a normal distribution for the observed variable, an often encountered structure in probabilistic modeling. The mean of the distribution was set as $\mu = a + b \cdot \text{day}$ and there were hyper-priors set for the priors of the a and b parameters. The hierarchy of the latent parameters in this model is an added complexity in comparison to previous models.

The observed random (or deterministic) variable in each model had a semantically meaningful name (temperature, random_number, reaction_time, day (deterministic)). The unidentified parameters were named with letters a, b, c etc. to avoid revealing information

about variables' relations through their names (e.g., `sigma`, μ). Each model had an unidentified parameter which was *unrelated* to the rest of variables. A variety of prior distributions for the unrelated parameters was used; a uniform for parameter `a` in Model 1, a half-normal for parameter `b` in Model 2, and a normal for parameter `d` in Model 3.

Models' prior distributions were used in the study. Prior distributions reflect directly models' structure. As observations come into a model and the prior beliefs are updated, the initial structure of the model can be overwhelmed in the posterior distribution. For a clearer experimental protocol, the tasks restricted to the *prior space*.

5.6.2 Implementation Details

Irrelevant interactive elements (zoom tools, hovering-over tooltips, tabs, drop-down menus) were removed from pair plots in study tasks to isolate the conditioning-related interactivity as the focus of the study was on that. The pair plot was showing the minimum necessary subset of model's variables in every task to avoid overwhelming participants with irrelevant information. The unidentified variables were appearing in alphabetical order across the diagonal of the matrix with the observed variable presented at the bottom to create a consistent view across participants and tasks and avoid any possible extra cognitive load of participants having to search for a variable in a randomized matrix.

The task models were specified and interpreted in PyMC3. The PyMC3's prior sampling method (`pymc3.sample_prior_predictive`) was used to generate prior samples for models' variables. For example, for Model 1 specified in PyMC3 in Fig. 5.1(b), samples were generated from the prior joint distribution of `temperature` and `b` represented by the blue surface in Fig. 5.2(a). This set of samples was used to create the scatter and KDE plots in Fig. 5.2(d)-(g) and (l)-(o), Fig. 5.3, and Fig. 5.4(a),(b).

5.7 Analysis and Results

5.7.1 Evaluation Measures

Participants' accuracy, response time and confidence were evaluated in the user study. Accuracy was measured as the number of correct selections in the multiple-choice input by each participant in every task. Participants' selections in the multiple-choice input were transformed into a binary representation with 0 indicating a wrong and 1 a correct selection. The binary representation of each response in T1's tasks (multiple selections were allowed) consisted of as many binary digits as the available options of the multiple-choice input, excluding the "none" option, while for T2 and T3 tasks (single selection was allowed) consisted of a single digit. Participants' performance in each task was computed as the number of occurrences of digit 1 in their response.

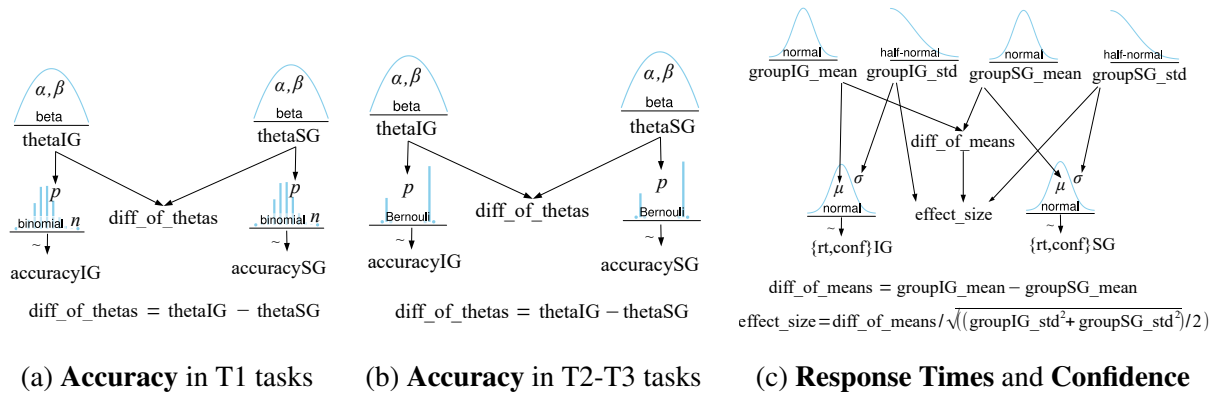


Figure 5.6: Probabilistic models used for the **Analysis** presented as Kruschke-style diagrams.

Participants’ response time was measured (in seconds) from the moment the task was displayed until the answer was submitted. Participants rated their confidence in each task on a 1-5 scale with increasing level of confidence (1: not at all, 2: slightly, 3: somewhat, 4: fairly, 5: completely). This was remapped to the $\{-2, -1, 0, 1, 2\}$ set to center the parameterization.

5.7.2 Data & Bayesian Modeling of Responses

Data. The collected data was split into subsets based on the condition (IG and SG). No participant or response was excluded. The multiple blank registrations of some participants, who accidentally clicked the “Register” button multiple times were omitted. Accuracy data consisted of numbers of participants’ correct selections in the multiple-choice input in every task. Response time data consisted of times (in sec). Confidence data consisted of ordinal values. A Bayesian analysis of the collected data was conducted on the level of the individual tasks. Fig. 5.6 presents the graphs of the three probabilistic models used for the analysis (see more details about these model in Appendix C.3).

Accuracy Analysis Models. Each group’s performance in every task was modelled by a binomial likelihood (that was reduced to a bernoulli likelihood for T2 and T3 tasks) (Fig. 5.6(a)-(b)). The posterior *probability of success* θ of the binomial distribution was estimated for each group. This probability expresses the propensity of a participant in the corresponding group to make a correct selection in each task. The two groups were compared in terms of accuracy by taking the difference of each group’s posterior distribution of θ .

Response Time Analysis Model. Each group’s response time in every task was modelled by a normal likelihood (Fig. 5.6(c)). The posterior distribution of *effect size* (*Cohen’s d*) was estimated for the comparison of the two groups to normalise for the varying duration (and thus typical variances) of the tasks.

Confidence Analysis Model. Each group’s confidence in every task was modelled by a normal likelihood (Fig. 5.6(c)). The simplifying assumption that the ordinal values could be treated as if they lay on a common continuous scale was made; hence the normal likelihood. A

more sophisticated analysis could have inferred a (potentially per-subject) monotonic relationship between ordinal responses and “true” confidence. The posterior *mean confidence level* was estimated for each group as confidence takes ordinal values and there was no need to normalise. The difference of the mean confidence posterior distribution of each group for every task were estimated to compare the two groups.

5.7.3 Results

Fig. 5.7 presents the results of inference in a set of forest plots. Comparing the two groups (IG and SG) based on the differences of the posterior distributions, an effect of the interactive conditioning is more likely given the data as the value 0.0 (reference line in columns 3-5 of Fig. 5.7 indicating no difference) becomes less likely under the difference of the posterior distributions (horizontal posterior highest density interval bars). That is the highest density intervals of the posteriors in the forest plots presenting the differences are pulled away from the reference value towards the right. The effect of interactive conditioning becomes less likely when the highest density intervals of the posteriors are pulled towards the reference value.

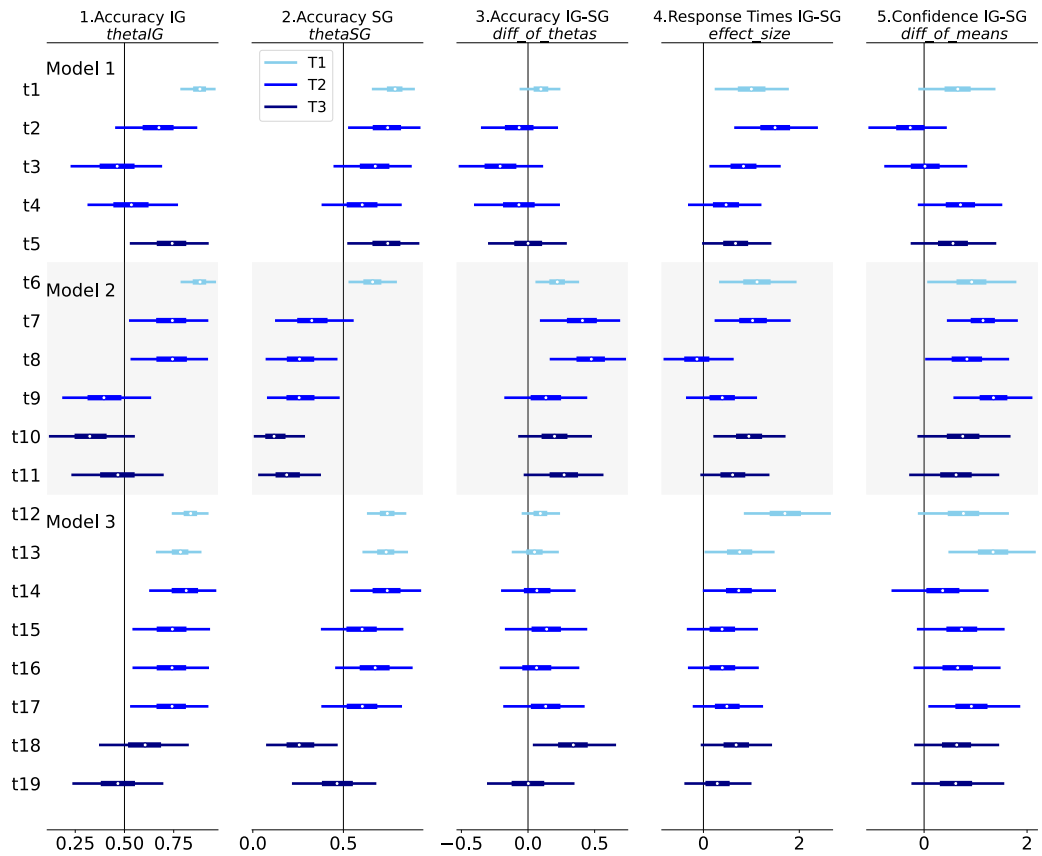


Figure 5.7: **Results.** Forest plot (94% HDI) of the posterior distributions of the probability of correct answer for IG (θ_{IG}) and SG (θ_{SG}), difference of θ s ($diff_of_thetas$), effect size of response times ($effect_size$) between IG and SG (normalised difference of duration), and difference of the estimated mean confidence of participants about their responses ($diff_of_means$). Tasks are presented vertically grouped per model.

Accuracy. Participants’ performance overall was good in both groups with the inferred probability θ of giving a correct answer being over 0.5 in most tasks with greater certainty for tasks of lower level of structural detail (T1-T2) (columns 1-2 of Fig. 5.7). Participants’ accuracy in tasks t_3-4 , t_{11} , t_{19} in IG and t_{19} in SG was around 0.5, while in tasks t_9-10 in IG and t_7-11 , t_{18} in SG was the lowest than other tasks (< 0.5). The lower accuracy concerned tasks with instances of more complicated statistical modeling; the parameterization of the bounds of a uniform distribution in t_7-11 (Model 2) and a hierarchical structure of a hyper-prior and prior in t_{18} (Model 3).

The effect of interactive conditioning is revealed by the differences of θ s. The effect is strong in t_6-8 (Model 2), and t_{18} (Model 3) and weaker in t_{10-11} (Model 2) (column 3 of Fig. 5.7) clearly showing the benefit of interactive conditioning in more sophisticated model designs.

Tasks t_2 (Model 1), t_8 (Model 2), and t_{17} (Model 3) concerned unrelated variables. A strong effect of interactive conditioning is observed in t_8 (Model 2) and not in the rest (columns 3 of Fig. 5.7). The square and full-gaussian shapes of the scatter plots in t_2 and t_{17} respectively

were more accurately interpreted as “absence of relation”, while the half-gaussian shape of the scatter plot in t_8 confused participants in SG regarding the existence of a relation (see relevant screenshots of tasks in supplemental material). This shows the benefit of interactive conditioning in identifying relations in cases of peculiar shapes of scatter plots resulting from unusual combinations of prior distributions.

Response Times. Participants using interactive conditioning needed more time to complete tasks overall with the effect being strong in tasks t_{1-3} (Model 1), t_{6-7} and t_{10} (Model 2), t_{12-14} (Model 3) (column 4 of Fig. 5.7). The differences in response time between groups are pulled towards the reference line (0.0) in tasks of middle or high level of detail t_{4-5} (Model 1), t_{8-9} and t_{11} (Model 2), t_{15-17} and t_{18-19} (Model 3) (column 4 of Fig. 5.7) implying that the extra exploration time interactive conditioning introduces tends to diminish in cases of more complex tasks.

Task t_8 was the one with the smallest mean difference (close to 0.0) in the response time and the greatest mean difference in accuracy between the groups (columns 3-4 of Fig. 5.7). This could imply that the observed effect on accuracy cannot be explained by possible extra exploration time in IG (see Section 5.7.4 for further analysis on this).

Confidence. Participants in IG are more confident than those in SG overall with the effect being strong in tasks of lower level of detail t_{6-9} (Model 2), t_{13} and t_{17} (Model 3) (column 5 in Fig. 5.7). Task t_{13} presents one of the strongest effects of interactive conditioning on confidence, while there is no corresponding effect on accuracy (columns 3, 5 in Fig. 5.7). This could imply that interactive conditioning makes participants with equivalent performance more confident (see Section 5.7.4 for further analysis on this).

5.7.4 Comparative Analysis

Do higher response times imply better accuracy or higher confidence? Does better accuracy imply higher confidence? The conduction of a causal analysis of the observed data is out of the scope of this analysis, but the existence of relations (correlations) between these pairs based on the inferred data will be investigated.

Fig. 5.8 presents the pair plot of the mean values of the differences of the posteriors for the accuracy, response times, and confidence between the IG and SG groups. There is a positive correlation between the differences in accuracy and confidence implying increased confidence with increased accuracy of the IG in comparison to the SG. Interactive conditioning when used in pair plots seems to support more accurate and certain decisions in the study tasks than the static pair plots. There is a slight negative correlation between the differences in accuracy and response time implying that any increase in the accuracy of the IG would not be attributed to increased response times. The negative correlation between the differences in confidence and response time is implied by the previous two correlations.

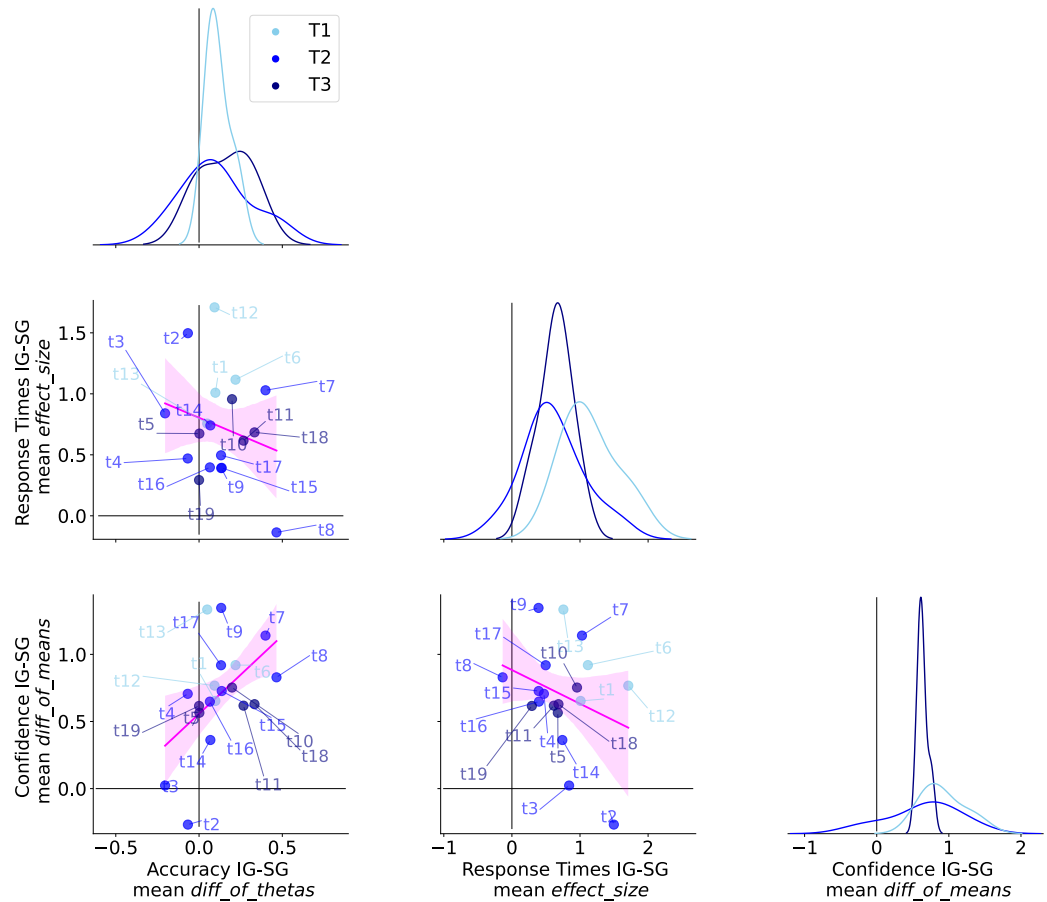


Figure 5.8: **Results.** Pair plot of mean values of the posterior distributions of *diff_of_thetas* for the accuracy, *effect_size* for the response times and *diff_of_means* for the confidence. The fitted linear regression line is drawn with a 90% bootstrap confidence interval in each scatter plot.

5.7.5 Analysis of Interaction Logs

The coordinates of the selection boxes drawn by the IG participants in each task were recorded. The proportion of participants having drawn at least one selection box was high in all tasks (10/13 in task t5, 12/13 in tasks t3–4, t6, t15, t19, and 13/13 in all other tasks). The (Q1,Q2,Q3) quartiles of the number of selection boxes drawn per task were (4.5, 9., 13.) and of the normalized (by the range of the corresponding variable) length of selection boxes were (0.11, 0.16, 0.24). The lengths of the selection boxes dragged and drawn by participants were between the 10% and 25% of variables' ranges being coherent with the shape of the distributions. Such sizes of selection boxes are big enough to capture part of the distribution with roughly constant curvature. Bigger sizes would capture several modes.

5.8 Discussion

5.8.1 When is Interactive Conditioning (not) Beneficial?

Model Designs. The analysis of the collected data (Section 5.7.3) showed that the use of interactive conditioning is beneficial for users' comprehension of probabilistic models in cases of more sophisticated model designs like the parameterization of the bounds of a uniform distribution, hierarchical structures, and unusual combinations of prior distributions (e.g. uniform and half-normal) (t_{6-8} , t_{18}). This finding is strengthened by the fact that participants in SG had a higher former training in Statistics than those in IG (Fig. 5.5D4-5) and by excluding the possibility that the higher accuracy could be attributed to the longer response times observed in IG due to more time spent in the exploration of the structures (Section 5.7.4).

In all other tasks (t_{1-5} , t_{10-17} , t_{19}) there was no strong effect of interactive conditioning on participants' performance (column **3** of Fig. 5.7). The possibility of low use of interactive conditioning being an explanation for this was excluded (Section 5.7.5). The low complexity and commonness of the statistical associations encountered in most of these tasks or the high complexity of others could possibly, at least partly, explain this observation. Tasks t_{1-5} and t_{12-17} concerned simple common statistical associations (e.g. a normal distribution setting the μ of another normal distribution) making both representations adequate enough for participants to achieve a similarly good performance ($\theta > 0.5$) in most of these tasks (columns **1-2** in Fig. 5.7). Task t_{19} was a very complex task concerning the determination of the mathematical formula (a linear regression) (T3) linking four variables (temperature, a, b, day) together. Participants' performance in any of the groups was close the random choice ($\theta \approx 0.5$) (columns **1-2** in Fig. 5.7).

Level of Structural Detail. Interactive conditioning did not seem to benefit participants' performance in a specific level of detail of variables' relations (column **3** of Fig. 5.7). Model design was more determining in the effect of interactive conditioning on users' performance than the complexity of the task. On the contrary, the level of detail seem to play a role in the differences of participants' response time and confidence between the two groups. The differences in response time diminish in tasks of higher levels of detail (Section 5.7.3). The burden of the extra exploration time required for interactive conditioning reduces in more complex tasks. The greatest advantage of interactive conditioning in participants' confidence in comparison to static pair plots appear with greater certainty in tasks of lower level of detail (Section 5.7.3). In simpler tasks, participants using interactive conditioning are more confident.

5.8.2 Practical Implications

Various types of users of probabilistic models could benefit from visualizations like interactive pair plots. Model builders could benefit from such visualizations when they encounter complex

model designs or have lack of statistical experience to conduct prior predictive checks and validate models (are the relations (effects) of variables as expected?). Decision-makers in crucial areas like healthcare or stock market could benefit from such visualizations when they should eliminate any risk of ignorance or misunderstanding of models' structure to decide on crucial interventions (what is the effect of a model's variable on another?).

Researchers could benefit from such visualizations when they need to tune some parameters in a model and need to (for)see the effect of doing so, or to communicate their models to a broader audience and provide a more intuitive overview of the model. Teachers and learners of Bayesian modeling could benefit from such visualizations when the former seek for ways to illustrate the effects of variables in various model designs, and the latter to gain a more intuitive understanding of the different model designs.

Visualizations like IPP could help users explore the effects of variables in the posterior space, as well. The relationships of variables under the posterior are governed by effects, although these effects usually cannot be expressed analytically in explicit mathematical or statistical associations as they can in the prior space. This happens because the exact form of the posterior distribution usually cannot be expressed analytically unless the priors are conjugate [Lambert, 2018a]. In these cases, posterior distributions can be estimated by sampling algorithms (e.g. MCMC), which does not allow us to know how exactly variables are associated (e.g. variable x sets the μ of the distribution of variable y). Although the effects of variables in the posterior space could be visualized and explored through such visualizations, most probably they could not be interpreted as specific analytical relations.

The findings of the user study are not restrictive to a specific PPL. Visualizations like IPP could be used to present the output of any probabilistic programming code that is being interpreted. Any programming language or library (including PPLs) that supports operations on probability distributions could be used for sampling from the prior and any PPL could be used for inferring the posterior.

5.8.3 Limitations of User Study

The analysis of the collected data suggests that interactive conditioning is beneficial for users' understanding, but it is not clear what aspect of it actually helps users. The recorded interaction data could not provide us with more insight into how participants in IG were exploiting the specific implementation of interactive conditioning. Were they combining information from both scatter plots and marginal distributions? Were they only looking at the conditional marginal distributions? Or were they only looking at the highlighted scatter plots? Such questions would require other experiment designs that would include one or combinations of open questions, think-aloud protocol, analysis of participants' micro-interactions [Breslav et al., 2014] or eye-tracking.

The participants' sample in this study present limited demographics with respect to age and

educational background. Further experimentation could be conducted on an expanded sample with broader demographics to investigate if the findings of this user study would replicate (as Ottley et al. [2012] did for Brase [2009] and Micallef et al. [2012]).

The presentations of the study's models was restricted to the prior space. This offers analytic descriptions of variable's relations (what controls what and how) as explained in Section 5.8.2, which could be used to validate participants' responses (ground truth). For a clearer experiment design, the case of posteriors determined by conjugate priors was not included. The types of distributions that could be explored was limited by the fact that prior sampling from heavy tail distributions (student-t, Pareto, Cauchy) gives a Dirac-delta-looking estimation of the probability density. Exploring variables distributed in such ways in IPP would not reveal any effect on their distribution while conditioning on them.

The number of questions had to be limited to ensure the completion of study by participants in roughly an hour. The user study was designed to include a variety of probabilistic model types (parameterized, linear regression, hierarchical), distributions (normal, half-normal, uniform), and statistical associations (setting the mean, standard deviation, or bounds of the distribution directly or through simple mathematical equations). There are many more model designs (logistic regression, GPs), distributions (discrete distributions like binomial and Poisson) and configurations that could be explored in the future in the context of a study like this one.

The user study presented in this chapter focused on a concrete implementation of interactive conditioning of a model's distribution; a brushing-and-linking approach in a classical scatter plot matrix. The findings of the user study concern this concrete implementation. Further experimental evaluation is required to investigate whether other designs of interactive conditioning (e.g., interactive conditioning on a parallel coordinate graph) would confirm the results of the analysis shown in this chapter.

The user study in this chapter was a controlled experiment where participants were not provided with any textual or mathematical description of the probabilistic models for the purpose of a clearer experiment design. Nevertheless, there are various tasks like building, validating, and refining a model in which presenting a textual or mathematical description of the model along with the interactive pair plot could be required, expected and helpful. A limitation of the work presented in this chapter is that such scenarios of tasks in which information about the model might be retrieved and/or combined from an explicit and implicit presentation of a probabilistic model have not been investigated. In such cases, further investigation is required as to whether the findings of this user study can be confirmed when these two sources of information about the model are combined.

5.8.4 Future Work

This chapter focused on a specific implementation of a visualization to communicate visually the relations of variables in a probabilistic model. The design of this visualization could be further improved or extended. For example, by interviewing actual users of the tool like experts in the field (analysts or statisticians who build models) or decision-makers, who would need to use such a tool in their everyday work routine, could inform the design of the tool and lead to various improvements.

For example, at the moment brushing can only be applied on the KDE plots of the marginal distributions. Brushing in the scatter plot matrices is typically applied on the data shown. A similar implementation of brushing could be implemented for IPP; users could apply either a univariate or conjunctive condition on the scatter plot showing the distribution of two variables of the model. This could offer more flexibility in conditioning the model's distribution. An alternative design of IPP could offer a conjunctive view of the scatter plot matrix and the definition of the model in probabilistic statements and create a link between the two. For example, every time a user hovers over a scatter or KDE plot the probabilistic statements of the corresponding variables could be highlighted. Such a design would be more informative as it would include the mathematical details of the model. It would be interesting to investigate whether a design like this would be useful for or preferred by actual users.

A completely different design approach to show the distributional information of the model could be an interactive parallel coordinate graph; the vertical axis of the graph would correspond to the model's variables and each line to a joint sample drawn from the (prior or posterior) model's distribution. In this design interactive conditioning could be implemented in various ways: for example, imposing a value restriction on the vertical axis, or selecting individual lines. This representation could possibly convey the distributional information more intuitively as it does not involve any graphs of distributions, which in existing literature do not seem to be as well understood as other frequency-framed forms. It would be interesting to investigate if the findings of the user study presented in this chapter would replicate when this alternative design is used by users.

5.8.5 Conclusions

Although there are various existing visualizations of probabilistic models and variables' relations, it is very little known about whether and when they support users' comprehension of the models. This work focused on interactive conditioning and investigated through a user study whether adding it to classical scatter plot matrices helps users better understand probabilistic models and if there are levels of structural detail and model designs for which it is beneficial. The analysis of the collected data showed that interactive conditioning is beneficial in cases of sophisticated model designs and the difference in response time between the interaction and

static group becomes less important in higher levels of structural detail. Participants using interactive conditioning were more confident about their responses overall with the effect being stronger in tasks of lower level of detail. These initial findings evoke the need for more research to understand how users can benefit from visual representations of probabilistic models and could pave the way for future investigation into the role of interaction to support more explainable Bayesian probabilistic models [Kulesza et al., 2015] and users' engagement with them.

Chapter 6

Visualizations of Simulated Data of Interventions to Support Users' Causal Reasoning and Decisions on Interventions

6.1 Summary

This chapter presents the third part of this research. This work focuses on proposing a pipeline to generate simulated probabilistic data from interventions applied on causal structures that are expressed in PPLs using probabilistic modeling and Bayesian inference. The creation of an automatic visualization tool for visualizing the simulated probabilistic data from causal structures and interventions as generated by this pipeline is an aim of this work. The purpose of this tool is to support causal reasoning and design of interventional experiments through the exploration of simulated data from interventions. The evaluation of the proposed visualizations in this tool is another aim of this work. The evaluation focuses on how people reason causally and make decisions on interventional experiments when the uncertainty in the simulated data of interventions is presented in static, animated, or interactive visualizations.

Section 6.2 discusses the purpose of this work, Section 6.3 describes the proposed pipeline for simulating interventions on causal structures using Bayesian inference, Section 6.4 presents the automatic visualization tool for visualizing the simulated probabilistic data from the interventions, Section 6.5 presents the details of the evaluation user study conducted, the analysis of the collected data and the results of the analysis. Finally, Section 6.6 discusses the contributions of this work and the conclusions drawn from it.

6.2 Purpose

6.2.1 Motivation

A common task in many fields in science (e.g., medicine, epidemiology, biology), or industry (e.g., marketing) is the identification of cause-effect relations between variables of interest. For example, a doctor needs to know what the cause of a disease is to be able to treat it, or a marketing manager needs to know which campaign locations or referral channels influence the purchases of a product the most to decide on a marketing plan.

The first and most intuitive thing to do to identify the cause of an observed effect is to follow a “trial-and-error” approach. Everybody applies this, many times even unconsciously, in their everyday life [Lagnado and Sloman, 2004]; click on buttons, icons, and other interactive elements on our mobile phone to see what does what, remove items from our diet to see what makes us overweight, etc. This trial-and-error is formally called *intervention* in the language of causality.

Interventions is a valuable tool in science and industry for inferring causality but they are not always possible or might be expensive to run. Thus, a current issue in existing literature is how to infer causality through observed data, which leads to the so widely-studied question “Is correlation causation?” [Pearl, 2009; Sekhon, 2008]. Various visualization tools have been proposed for the visualization, exploration, validation, and elicitation of causality based on observed data [Dang et al., 2015; Ge et al., 2020; Wang and Mueller, 2015; Wang and Mueller, 2017; Xie et al., 2020]. Nevertheless, it is very little known about how well people reason causally or how informative their decisions on interventional experiments are when they use such visual analytics tools [Kale et al., 2022; Yen et al., 2019]. On the other hand, the existing literature presents some evidence that users’ performance in identifying the causal relations of variables improves when simulated data of interventions is presented to them instead of the observed data [Lagnado and Sloman, 2004].

6.2.2 Purpose of This Work

This work focuses on proposing a visualization tool to support the exploration of various assumed causal structures of observed data by simulating interventions on them. The purpose of such a tool is to guide users in the design of interventional experiments. Users will be able to simulate interventions on structures that encode their causal assumptions about how the observed variables might be causally related and observe what the expected outcomes of possible interventional experiments could be under various assumptions.

Using such a tool, users would be able to test their expectations about an interventional experiment against the outcomes of simulations that incorporate prior knowledge or possible assumptions about the data generating mechanisms. This could possibly lead to more informed

decisions about the conduction of an experiment. Such a tool could also be useful in cases that interventional experiments might not be possible and some background knowledge about them is available.

The aim of a visual exploratory tool that relies on simulation like the one described above is not the conduction of causal inference, i.e., the identification of the true causal structure of data. Such a tool will not reveal the actual underlying data generating mechanism but will help determine what the most informative interventional experiments might be to reveal the true causal structure of data. Any conclusion users draw from their explorations with such a tool will depend on the specific causal model and dataset used for the simulation. If the causal model or dataset are not true, there is always the risk that the conclusions might not be valid in the true world. After all, true causal structures can only be retrieved through actual interventional experiments or causal inference tools.

6.2.3 Summary of This Work

The contribution of this work is mainly in three areas; first, the suggestion of a pipeline to exploit existing powerful tools for probabilistic modeling and Bayesian inference in simulating interventions on causal structures (Section 6.3), second, the creation of automatic visualization tools to present the simulated data from causal structures and interventions on them (Section 6.4), and third, the evaluation of these uncertainty-aware visualizations in causal reasoning and decision-making tasks (Section 6.5). The intention was to investigate how well people reason causally and how informed decisions on interventional experiments they make when they are presented with uncertainty-aware visualizations of simulated probabilistic data of interventions.

Existing PPLs are tools for probabilistic modeling that are not inherently built to model causality. Thus, they are quite inflexible in the manipulation of models for applying and simulating interventions. In this work the ways that these tools could be used for this purpose are demonstrated. The intention is to exploit the advantages of Bayesian inference in modeling uncertainty and accounting for prior knowledge in causal inference, and simulating interventions given the existing probabilistic modeling tools.

In terms of visualization, the classical scatter plot matrix is used for the parallel presentation of the simulated data before and after an intervention. This is enhanced with interaction or animation to present the simulated data of interventions in slices conditioned on the intervened variable. An automatic tool that generates this visualization is created.

This implementation of the scatter plot matrix has many similarities with IPP especially in the outlook but differs in various crucial aspects of the design to serve the purposes of this research: two separate sets of data are presented in parallel in the new tool, the simulated data before and after the intervention, while only the first was presented by IPP; the slicing of the interventional data is performed using a slider instead of the IPP's selection box; only the selected slice of the interventional data is shown each time in contrast to IPP in which the whole

dataset is shown and the slice is simply highlighted. All these design options will be explained in Section 6.4.1.

A user study was conducted to investigate the effect of interaction and animation on users' ability to identify the causal structures of the presented data and to make decisions on interventional experiments.

6.3 From Observed Data to Simulated Probabilistic Data of Interventions

6.3.1 Approaches to Causal Inference

An observed correlation between two variables in a dataset does not guarantee that there is a cause-effect relation between the variables. This correlation could be spurious resulting, for example, by the existence of a confounding (common cause) variable as explained in Section 2.4.1. This difficulty of interpreting the correlations in observed data as causal or non-causal relations leads to the need for methodologies of causal inference from observed data.

The two main existing frameworks for causal inference differ in the approach they follow; one is more data-driven, while the other is model-driven [Hernán and Robins, 2020; Pearl and Mackenzie, 2018]. The first framework is based on the idea of *potential outcomes* developed by Neyman and extended by Rubin [Sekhon, 2007]. In this framework a causal effect is approached as a difference between the effects of the potential outcomes of a treatment on the level of the individual subjects. Only one of the potential outcomes realizes for each subject. The other is counterfactual and is missing from the observed data. As a result, the causal effect of a treatment on each subject cannot be directly estimated due to the missing data. The causal inference is treated as a missing data problem using assumptions and statistics to estimate, for example, mean effects. In this methodology, causal inference is data-driven.

The second framework mainly proposed and developed by Pearl unifies all the pre-existing approaches to causation; graphical, potential outcome, structural equations, decision analytical, interventional, sufficient component and probabilistic approaches to causation with each of these viewed as a restricted version of the *structural causal model* [Pearl, 2010]. Pearl's structural causal model (or causal model) [Pearl, 2009] approaches the causal effects as influences of variables on each other within the context of a system or population. Pearl's structural causal model is a mathematical model that consists of two sets of variables, V and U , and a set of functions, F determining how values are assigned to each variable $V_i \in V$ [Bareinboim and Pearl, 2016]. The variables in V are endogenous variables whose values are determined by factors within the model. The variables in U are observed or unobserved exogenous variables whose values are determined by factors outside the model and influence but are not influenced by the endogenous variables [Pearl, 2010]. These variables describe background conditions

for which no explanatory mechanism is encoded in the causal model (i.e., it is not explained how they occur) because the modeler decides to keep them unexplained [Bareinboim and Pearl, 2016; Pearl, 2010]. Their existence is acknowledged to qualitatively assess how they relate to endogenous variables [Pearl, 2010].

The variables in U can induce (epistemic and aleatoric) uncertainty in the causal model. There might be lack of knowledge about their values because they might be unobserved (epistemic uncertainty is introduced) [Pearl, 2010]. They might be part of a “physical reality (e.g., genetic factors, socio-economic conditions)” which causes variations in the observed data (aleatoric uncertainty is introduced) [Pearl, 2010]. An example of this could be a natural variability appearing in terms of a feature in a general population [Bareinboim and Pearl, 2016; Stanford Encyclopedia of Philosophy, 2018]. Every instantiation $U_i = u_i$ of the exogenous variables uniquely determines the values of all variables in V and, hence, if probability distributions $p(U_i)$ are defined for the variables in U , a probability distribution function $p(V_i)$ will be induced on V .

Regarding the functions in F , the causal model consists of a system of functions, $f_i \in F$, with each function corresponding to one of the observed variables $V_i \in V$ in the model and expressing the dependencies of the observed variable on the other exogenous and endogenous variables. If each function is invariant to possible changes in the form of the other functions, the system of these functions is said to be *structural* [Pearl, 2010]. Let us assume $X, Y \in V$, and $U_X, U_Y \in U$ assumed to be jointly independent, and X causing Y . The structural equations for X , and Y are as following:

$$X = f_X(U_X) \tag{6.1}$$

$$Y = f_Y(X, U_Y) \tag{6.2}$$

According to Pearl [2010]:

Each of these functions represents a causal process (or mechanism) that determines the value of the left variable (output) from those on the right variables (inputs). The absence of a variable from the right hand side of an equation encodes the assumption that Nature ignores that variable in the process of determining the value of the output variable.

For example, the absence of a variable $Z \in V$ from the arguments of f_Y conveys the empirical claim that variations in Z will leave Y unchanged, as long as variables U_Y , and X remain constant [Pearl, 2010].

Pearl’s structural causal model has its roots to the *structural equation models (SEMs)*, which were first used by Wright [1921, 1934]. Wright revolutionised the standard practice of using regression to predict an outcome by combining information from more than one regression equations using a causal approach to regression modeling rather than merely predictive. Pearl disentangled linearity from causal models in his theory by representing causal models with non-parametric structural equations and introducing a framework for conducting causal inference

graphically without the need of any algebraic analysis [Pearl, 2010].

Pearl proposed the use of causal diagrams to encode all the information that the non-parametric structural equations represent and created the do-calculus, a set of rules applied to the causal diagram of a causal model, to represent interventions. He introduced the $\text{do}(x)$ mathematical operator to simulate physical interventions on x . By this operator certain functions are deleted from the model and replaced by a constant $X = x$ (e.g., f_x in equation 6.1 is replaced by a constant value x), while the rest of the model is unchanged. According to Pearl, the causal effect of a variable $X \in V$ on $Y \in V$ is given by the probability $\text{Pr}(Y = y | \text{do}(X) = x)$ [Pearl, 2010], namely the post-intervention probability of Y .

The central question in the analysis of causal effects is the question of *identification*: Can the controlled (post-intervention) probability distribution, $p(Y | \text{do}(X) = x)$, be estimated from data governed by the pre-intervention joint distribution of the model, $p(Y, X, \cup V_i \notin \{Y, X\})$. In the absence of unobserved confounding variables (when the set of variables U does not contain any unobserved confounding variable), all causal effects are identifiable [Pearl, 1993; Tian and Pearl, 2002]. The problem of identifiability arises when some confounders are not measured. In such cases a precise estimation of a desired causal effect depends critically on the locations of those confounders in the causal diagram.

Sufficient graphical conditions for ensuring the identification of the post-intervention distribution were established by several authors [Pearl, 1993, 1995, 2009; Spirtes et al., 2000; Tian and Pearl, 2002]. For example, all causal effects are identifiable whenever the causal model is Markovian, namely its graph is acyclic and all the variables in U are jointly independent [Pearl, 2010]. For such models the local Markov property is valid and used to factorize the joint distribution of the model based on its original and mutilated causal diagram before and after the intervention (see Section 3.3.1 for an example on factorizing a joint distribution based on the local Markov property). It is proved that the post-intervention joint distributions of the endogenous variables of the model equals the factorization of the joint distribution of the endogenous variables of the model before the intervention [Pearl, 2010]. Thus, the identification problem can be reduced to a graphical procedure and the do operator permits the use of structural equations as a basis for modeling causal effects and counterfactuals [Pearl, 2010].

This work focuses on representing causal models by linear probabilistic regression models in the spirit of the structural equation model; one linear probabilistic regression model is defined per observed variable using as covariates only the other variables of the model that are assumed to affect this variable. Some assumptions are made as explained in the following subsection. The probabilistic regressions are specified in a PPL; appropriate priors are set for the parameters of the model and likelihoods for the observed variables, and the observed data is used to infer the posterior distributions of the regressions' parameters and generate posterior predictive samples for the observed variables.

The intention in this work is to use the inferred posterior distributions to generate simulations

from interventions. When imposing an intervention on a causal model the structure of the causal diagram changes; any incoming arrow in the node we intervene on needs to be removed because the value of the variable corresponding to that node is defined externally and any causes of it do not affect it any more. This implies that the probabilistic model of the causal structure needs to be amended to reflect this change in the causal diagram and generate simulated post-intervention data based on the mutilated causal diagram.

Simulating interventions probabilistically has been encountered in the existing literature again [Huszár, 2019; Lagnado and Sloman, 2004; Witty et al., 2019; Xie et al., 2020]. The existing approaches amend the probabilistic model and usually sample directly from the amended probabilistic model. The proposed approach aims to define a mechanism for amending the probabilistic model and sampling from the pre-estimated inference results in alignment with the amended model. The closest idea to this is presented by Witty et al. [2019]. Witty et al. [2019] had a different ultimate aim; they wanted to infer the causal structure of some observed data from a prior distribution over various possible pre- and post-intervention causal structures. They used Bayesian inference and simulated interventions probabilistically to inform the prior distribution of the possible causal structures of the data. In this work, the inverse of this is suggested; to use the results of Bayesian inference to generate simulated data from interventions.

The following subsection presents a proposed pipeline for simulating interventions probabilistically exploiting the results of Bayesian inference. A concrete example will demonstrate how a PPL-specified causal model can be altered to apply an intervention, and how the inferred distributions can be used to generate simulated posterior predictive samples from the post-intervention mutilated causal model.

6.3.2 Using Bayesian Probabilistic Models to Simulate Interventions

If we have some observed data and prior knowledge for a problem that we cannot conduct actual interventional experiments, can we exploit Bayesian simulating engines to simulate interventions? The intention is to exploit any knowledge learned about the system based on the inferred posterior distributions and posterior predictive sampling distributions to generate data from the same system assuming that an intervention is applied on it. PPLs define a *simulator* via the posterior predictive distribution, whose parameters are “pre-baked” by inference on observed data. There are various advantages in using a Bayesian probabilistic approach and PPLs to model causal structures. The following points summarize them.

- The uncertainty in the causal model (e.g., because of lack of knowledge about the values of exogenous unobserved variables that affect the endogenous variables) can be modelled in a systematic way.
- Any prior knowledge about the problem and any available observed data can be used to generate posterior distributions.

- PPLs are powerful simulating engines that could be exploited to simulate interventions. This is very important when interventions cannot be conducted.

Simulating interventions by using probabilistic models and PPLs is not that straightforward. The purpose of this subsection is to present a pipeline for modeling a causal model using probabilistic models to define the structural equations of the causal model, and exploit the inference of these probabilistic models to generate simulated data of interventions applied on the causal model. This work restricts itself to linear models; probabilistic linear regression models will be used to determine the structural equations of the causal model. The probabilistic models will be expressed in a PPL, and specifically PyMC3. Although PyMC3 is used as a reference PPL in this work, the modeling approaches discussed here could be applicable to other PPLs presenting similar inflexibilities with PyMC3 in causal modeling and interventions simulating. These inflexibilities will be revealed along the discussion in this subsection and the identified ways to overcome them will be discussed.

The proposed pipeline to model a causal model probabilistically and create a Bayesian simulator for interventions is presented in Fig. 6.1. The main component in this pipeline is the *Probabilistic Modeling of Causal Models*. This component exploits the modeling and simulation mechanisms of PPLs to generate simulated posterior predictive data before (D) and after an intervention (G). It is designed to hide the implementation details from the user. The functionalities of this component could be split in four discrete subcomponents (B,C,D,G). The inputs to this component consist of a causal model and the available observed data (A), and the details of an applied intervention (i.e., the variable to intervene on and the value to set it) (F). A concrete implementation of this pipeline in Python using PyMC3 is provided in [Taka, 2023c].

There are two distinct paths in the proposed pipeline; the path $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ that describes the process of modeling probabilistically a causal structure, and the path $F \rightarrow G \rightarrow B \rightarrow D \rightarrow H$ that describes the process of simulating interventions. The following paragraphs discuss the most important aspects of the design in regards with these two paths.

6.3.2.1 Probabilistic Modeling of Causal Structures

Let us start the discussion about the proposed pipeline with a concrete example of a causal model found in Lagnado and Sloman [2004]. Box 6.1 presents the causal model of three observed variables: insomnia, anxiety, and tiredness. To model this causal structure some assumptions are made:

1. there is no other observed variable that affects these three variables;
2. there is no unobserved variable confounding any of these three variables;
3. any unobserved variables influencing these three variables are jointly independent;
4. when any affects another of the three variables, it does this in a linear way;

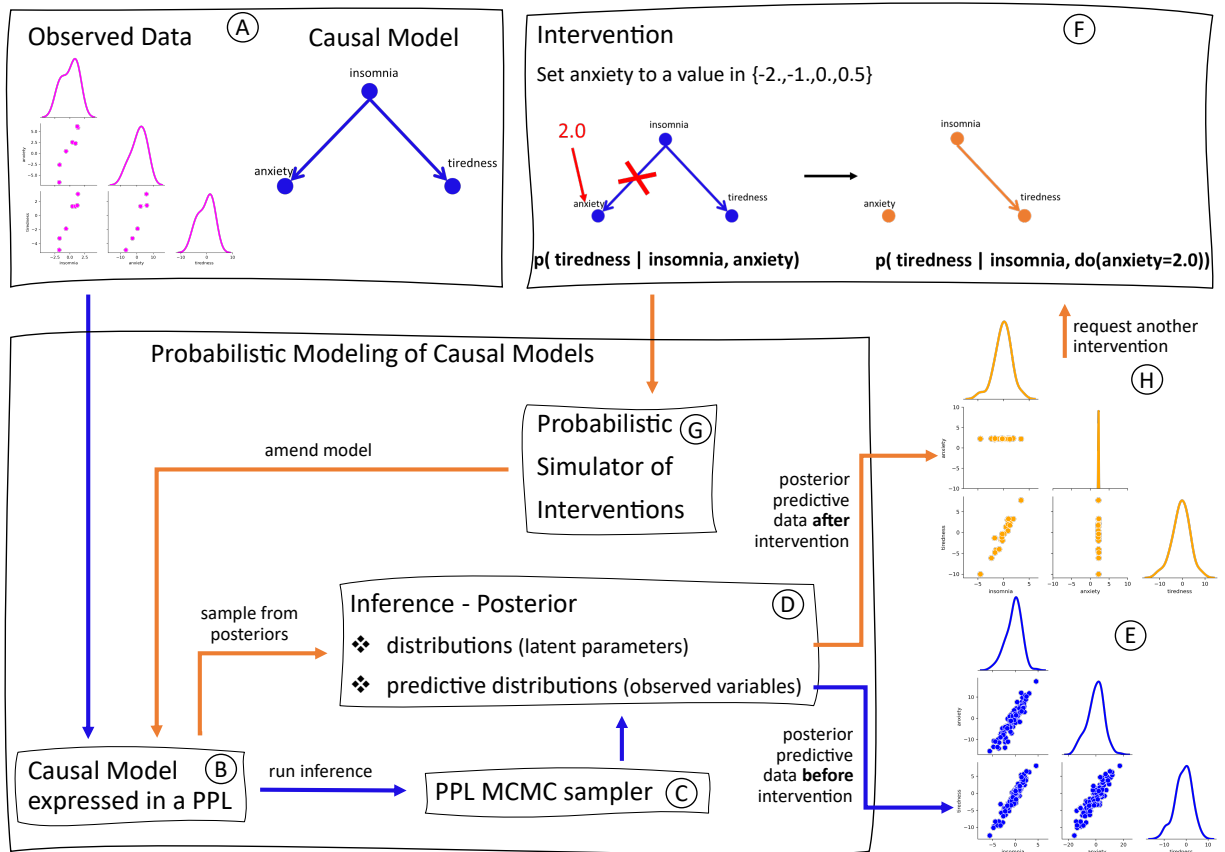


Figure 6.1: Proposed pipeline for modeling a causal model probabilistically and simulating interventions.

- the distributions of the observed variables and unobserved parameters of their distribution are known.

Box 6.1 A Causal Model Expressed By Linear Regression Models

Fig. 6.2(c) presents the causal diagram of three observed variables; insomnia, anxiety, and tiredness. The variables are represented by nodes and the direct causal relations by directed edges starting from a cause and pointing to an effect. Based on this causal diagram, insomnia causes anxiety and tiredness. The following probabilistic statements describe the probabilistic linear regression models of this three variables' causal model.

```

 $\mu_{ins} \sim \text{Normal}(0,1)$ 
 $\sigma_{ins} \sim \text{HalfNormal}(1)$ 
 $\sigma_{anx} \sim \text{HalfNormal}(1)$ 
 $\sigma_{tir} \sim \text{HalfNormal}(1)$ 
 $b_{anx\_ins} \sim \text{Normal}(0,1)$ 
 $a_{anx} \sim \text{Normal}(0,1)$ 
 $b_{tir\_ins} \sim \text{Normal}(0,1)$ 
 $a_{tir} \sim \text{Normal}(0,1)$ 
 $insomnia \sim \text{Normal}(\mu_{ins}, \sigma_{ins})$ 
 $anxiety \sim \text{Normal}(b_{anx\_ins} \cdot insomnia + a_{anx}, \sigma_{anx})$ 
 $tiredness \sim \text{Normal}(b_{tir\_ins} \cdot insomnia + a_{tir}, \sigma_{tir})$ 

```

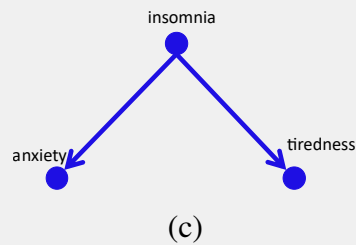
(a)

```

with pm.Model() as model:
    ## priors
    mu_ins = pm.Normal('mu_ins', mu=0, sd=1)
    sigma_ins = HalfNormal('sigma_ins', sd=1)
    sigma_anx = HalfNormal('sigma_anx', sd=1)
    sigma_tir = HalfNormal('sigma_tir', sd=1)
    b_anx_ins = pm.Normal('b_anx_ins', mu=0, sd=1)
    a_anx = pm.Normal('a_anx', mu=0, sd=1)
    b_tir_ins = pm.Normal('b_tir_ins', mu=0, sd=1)
    a_tir = pm.Normal('a_tir', mu=0, sd=1)
    ## likelihoods
    insomnia = pm.Normal('insomnia', mu=mu_ins,
                        sd=sigma_ins)
    anxiety = pm.Normal('anxiety', mu=b_anx_ins*insomnia+a_anx,
                       sd=sigma_anx)
    tiredness = pm.Normal('tiredness', mu=b_tir_ins*insomnia+a_tir,
                          sd=sigma_tir)

```

(b)



(c)

Figure 6.2: Definition of the insomnia, anxiety, and tiredness model in (a) probabilistic statements, and (b) PyMC3 code. (c) Causal diagram of insomnia, anxiety, and tiredness.

The path A->B->C->D->E in the pipeline in Fig. 6.1 describes the process for modeling this insomnia-anxiety-tiredness causal structure probabilistically and generating posterior predictive data from it. Based on the assumptions above a probabilistic linear regression model is used for each one of the observed variables in the notion of the structural equations. These probabilistic regressions are defined in Fig. 6.2(a) and expressed in PyMC3 in Fig. 6.2(b). This process takes place in component B.

The input to component B consists of the available observed data and the causal model (A). Some synthetic observations are generated from the insomnia-anxiety-tiredness causal model (see details about this in Appendix D.1) and used for the inference of the PyMC3 model. The causal structure is encoded in a Python dictionary structure in the way described in Appendix D.1.

Component C consists of the MCMC sampler of the PPL. The posterior distributions of the model's parameters and posterior predictive distributions of the three observed variables are estimated here. Component D holds these distributions along with a sampler to sample from them. The posterior predictive data can be then visualized e.g., using a scatter plot matrix (component

E). Fig. 6.3 presents in a joint scatter plot matrix the synthetic observed and simulated posterior predictive data for the insomnia, anxiety, and tiredness. In both cases all three variables seem to be positively correlated.

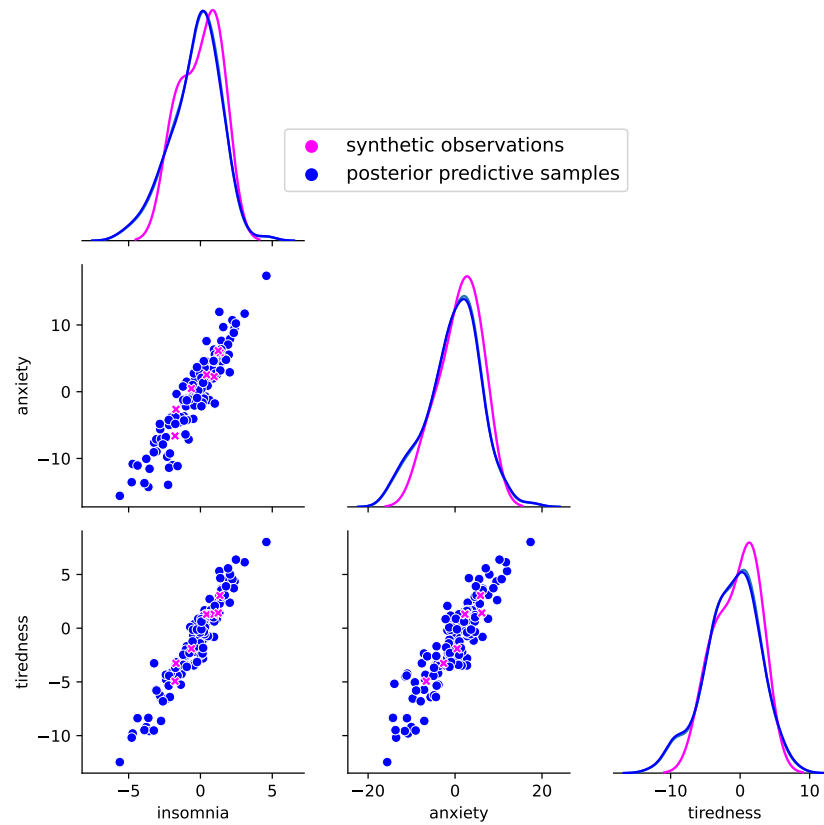


Figure 6.3: Scatter plot matrix of the synthetic observed and simulated posterior predictive data of the insomnia-anxiety-tiredness causal model.

6.3.2.2 Simulating Interventions Probabilistically

So far, the uncertainty in the causal model is modelled using any prior knowledge about the problem and the observed data, and given some assumptions. The intention is now to exploit the generated inference results, namely the inferred posteriors, to generate data after an intervention. The path $F \rightarrow G \rightarrow B \rightarrow D \rightarrow H$ describes the process of achieving this according to the proposed pipeline. Let us assume that we want to apply an intervention on anxiety and set its value equal to 2.0 (F). This intervention causes a change in the structure of the causal diagram as depicted in component F that needs to be reflected in the specification of the probabilistic model in component B.

A problem in achieving this in the context of a PPL is that usually the probabilistic model once specified cannot be amended. The proposed way to overcome this is to use a deterministic transformation of the likelihood of each observed variable that would serve as a “switch” activating and deactivating the likelihood and applying an intervention to the observed variables.

Two factors are needed for this transformation; f_1 and f_2 . Using these factors the likelihoods in the PPL code can be set as following: $f_1 \cdot \text{likelihood} + f_2$. The first factor is used to activate and deactivate the likelihood and the second to set apply the intervention to the variable and set it equal to a specific value. More concretely, the two factors are initialized as following for the path $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$: $f_1 = 1.0$ and $f_2 = 0.0$ meaning that the likelihood is initially activated and no intervention is applied to the observed variable setting it equal to x . The factors are set as following when an intervention needs to be applied; $f_1 = 0.0$ and $f_2 = x$ meaning that the likelihood is deactivated and an intervention is applied on the variable setting it equal to x .

In PyMC3, the `pymc.Data` object is used to define these factors of the likelihood. The factors defined in this way can be set at any point after the specification of the probabilistic model. This is an indirect way of amending the structure of the model on-the-fly. The `pymc.sample_posterior_predictive` method can be used after setting the factors of the likelihood appropriately to apply an intervention to sample from the pre-estimated posterior distributions. Doing so, we would essentially sample from the mutilated causal diagram after the intervention. Component G receives as input the details of the intervention, amends the model by setting appropriately the control factors of the likelihoods, and samples from the pre-estimated posterior predictive distributions. The posterior predictive samples constitute the post-intervention data that can be visualized, e.g., in a scatter plot matrix (H).

Three different types of interventions were considered in the provided implementation of the proposed pipeline in line with the work of Witty et al. [2019]:

- *atomic*;
- *shift*;
- *variance*.

The atomic intervention is when the value of a variable is set to a specific value as discussed so far. In this case, the f_1 and f_2 factors are used for the definition of the likelihood.

The shift intervention is applied to the mean value of a variable by shifting it by a fixed amount. A third factor, f_3 is included in the specification of the probabilistic model of the causal model to allow the application of shift interventions. For a normal likelihood, the specification of the likelihood given all three factors is as following: $f_1 \cdot \text{Normal}(\mu + f_3, \sigma) + f_2$. The factor f_3 is initialized to 0.0 and is set to the specific value the shift intervention requires it to be when a shift intervention needs to be applied.

The variance intervention is applied to the variance of a variable by scaling the standard deviation by a fixed amount. A fourth factor, f_4 is included in the specification of the probabilistic model of the causal model to allow the application of variance interventions. For a normal likelihood, the specification of the likelihood given all four factors is as following: $f_1 \cdot \text{Normal}(\mu + f_3, f_4 \cdot \sigma) + f_2$. The factor f_4 is initialized to 1.0 and is set to the specific value when a variance intervention needs to be applied.

6.4 The Visualizer of Causal Structures' Simulated Interventions

This section deals with how data from interventional experiments could be visualized to convey information about the causal structure of the underlying data visually. This data could either derive from actual interventional experiments or from simulations of such experiments. This section focuses mainly on the visualization of simulated interventional data deriving from probabilistic approaches. The main aim is the design and implementation of a visualization tool that would automatically transform the outputs of the simulation pipeline proposed in the previous section into uncertainty-aware visualizations. Section 6.4.1 presents the objectives that were set for the design of this tool, Section 6.4.2 presents the details of this tool's design and implementation and the limitations of the implementation, and Section 6.4.3 presents a use case of the tool.

6.4.1 Objectives of Design

This work identifies and focuses on two types of user tasks that could be supported by visualizations of simulated data from causal interventions. The *first task type (TT1)* is the identification of the underlying causal structure of some observed data of interest. In this type of tasks the visualizations of the simulated interventional data should present all the useful information entailed by the intervention to help users make judgements about which causal relations exist among variables. The *second task type (TT2)* is more complicated and builds upon the first. In cases when researchers want to conduct interventional experiments to infer the causal structure of some observed data, there might be many possible such experiments. Visualizing simulated interventional data could help users to identify which experiments could be more informative for determining the causal structure of the data and thus, make more informative decisions on which experiments to conduct.

These two types of tasks determined the design objectives of the visualization tool presented in this section and later, the research questions of the evaluation user study that was conducted. In this subsection the design objectives of the visualization tool for the simulated data of interventions are presented.

6.4.1.1 Inclusion of Uncertainty

The first objective is the inclusion of uncertainty in the visualization of simulated data from interventions. The data from interventional experiments usually entails inherent uncertainty even if the experiments are conducted in consistent environments and conditions. Researchers and analysts who conduct such experiments with the aim of identifying causal relations among observed variables have to deal with this uncertainty often and infer relations out of it. It is a reasonable

design objective to include uncertainty in the visualizations of simulated interventions with the aim to create a realistic depiction of the outcomes of an actual interventional experiment.

6.4.1.2 Parallel Presentation of the Simulated Data Before and After the Intervention

The second objective is the parallel presentation of the simulated data before and after the intervention. Some aspects of causal inference could be informed by the direct comparison of these two. For example, the fact that a correlation between two variables of interest breaks after intervening on a third variable implies that this correlation is spurious (a concrete example of how this could be inferred through the visualizations of the simulated data is presented in Section 6.4.3). This is interpreted as absence of causal relation between the two variables of interest.

6.4.1.3 Visualization of the Pairwise Distributions of the Simulated Data

The third objective is the visualization of the pairwise distributions of the simulated data. The purpose of this objective is similar to the objectives IPP had to serve; the communication of the variables' statistical relations. As explained in the provided example for the explanation of the previous objective, communicating how the statistical relation of two variables changes after an intervention can be informative in reasoning about the causal relation of the variables.

6.4.1.4 Exploration of the Outcomes From Applying Different Interventions

The fourth objective is the exploration of the outcomes from applying different interventions. The possible interventions considered in this work might vary in the following aspects:

- in the type of intervention (i.e., atomic, shift, and variance);
- in the observed variable the intervention is applied on. This variable will be called *intervened variable*;
- in the exact value to which the intervention is set. This value represents the value to which the intervened variable is set in an atomic intervention, or the amount of shift that is applied to the mean value of the intervened variable in a shift intervention, or the scaling factor that is applied to the variance of the intervened variable in a variance intervention. This value will be called *interventional value*.

Offering the possibility of exploring different intervention types provides the flexibility to researchers to explore many different scenarios and designs of potential experiments. Exploring the interventional data corresponding to different interventional values could help identify which other variables in the model are affected by or remain correlated with the intervened variable after the intervention. These inferences could inform users' reasoning as to which the causes

and the effects of the intervened variable could be (a concrete example of how this could be inferred through the visualizations of the simulated data is presented in Section 6.4.3). The sensitivity of the affected variables to various changes in the value of the intervened variable could also be observed.

6.4.1.5 Joint Presentation of Causal Models & Simulated Interventional Data

The fifth objective is the graphical representation of multiple hypothesized causal models in conjunction with the simulated data of interventions generated from a single causal model. Presenting causal models graphically is a standard practice in causal reasoning and has been advocated to facilitate effective reasoning [Greenland et al., 1999; Kale et al., 2022; Yen et al., 2019]. This possibility would serve various purposes. First, an intervention might not be informative enough to determine the causal relations between all variables in a model and ultimately, the exact causal structure of the data. If there are some hypothesized causal models that are considered as possible explanations of the data generating mechanism the information provided by an intervention could narrow down these possibilities. Thus, analysts could explore which of their hypothesized causal models could be compatible with beliefs and observations under the interventional data. Second, a user could explore various interventions and determine those that are informative enough to distinguish the hypothesized causal models. This could support decisions on the conduction of actual interventional experiments.

6.4.1.6 Automatic Transformation of Simulated Interventional Data Into Visualizations

The sixth objective is the automatic transformation of the outputs (i.e., posterior distributions, and posterior predictive samples before and after interventions) of the interventions' simulation pipeline presented in the previous section into the visual representations of the visualization tool proposed in this section. The outputs of the simulation pipeline will be standardized in similar way like in the case of IPME and IPP and used as an input to the proposed visualization tool in this section. The aim is to create a seamless representation of the simulated data to support TT1 and TT2 tasks.

6.4.2 Design & Implementation

This subsection presents the *Visualizer of Causal Assumptions and Uncertainty-Aware Simulations of Interventions* (*vicausi*). This is a visualization tool for visualizing the simulated data before and after interventions as generated by the pipeline presented in Section 6.3 and for exploring the probabilistic outputs of various causal interventions. This tool was designed according to the objectives presented in the previous subsection. The main visualization elements used are the scatter plot matrix and the causal diagrams; the first is used for presenting the simulated interventional data of the variables in pairs, and the second for presenting the causal

structure of the hypothesized models. The tool is provided in the form of a Python module in [Taka, 2023d]. The following paragraphs present the main aspects of the tool’s design and implementation and explains why certain design decisions were made where appropriate.

6.4.2.1 Input

The input of the `vicausi` tool consists of the pre-computed simulated data and some user-defined configurations. More concretely, the following elements should be given as arguments to the plotting method:

- a list of `npz` files similar in format to that given as input to IPME and IPP. Each one of these files corresponds to a different causal model (e.g., these models could be the hypothesized causal models of the analyst), contains sets of prior and posterior samples for each parameter in the model, and prior and posterior predictive samples for each observed variable in the model, and the DAG of the probabilistic model from which the assumed causal structure could easily be retrieved through the parent-child relations of the observed variables. These files are enhanced with extra `numpy` matrices and corresponding metadata for the pre-computed posterior predictive samples of a range of possible interventions. There is the possibility to add simulated data of interventions applied on each one of the observed variables in the model, for a range of interventional values for each observed variable, and for all intervention types considered in this work; the atomic, shift, and variance intervention. The code for generating this compact file can be found in [Taka, 2023c];
- an index to the model’s file whose simulated data is to be visualized in the scatter plot matrix. The simulated data of one of the provided causal models is shown in the scatter plot matrix each time. On the contrary, the causal structures of all provided causal models are shown along with the scatter plot matrix. This design option will be explained later in this section;
- the order in which the observed variables will appear on the visualization (i.e., a scatter plot matrix) and the graphical representation (i.e., causal diagrams) of the causal models;
- the preferred mode for visualizing the simulated data. The scatter plot matrix has been implemented in three different modes (i.e., static, interactive, and animated) for reasons that will be explained later in this section;
- the types of interventions that the user will be able to select and explore, and for each type of intervention the variables that could be intervened on. Pre-computed simulated data from these interventions should be included in the `npz` file of the model.

6.4.2.2 The Components of the Design

Fig. 6.4 presents the outlook of the `vicausi` tool for each one of the three visualization modes. The layout is similar in all modes. At the top there is a set of radio buttons for each intervention type with the radio buttons in each set corresponding to model's observed variables. These radio button widgets allow users to select the intervention they wish to explore.

Below the radio buttons widget there is the scatter plot matrix of the simulated data on the left, and the causal diagrams of the causal models contained in the input `npz` files on the right. The intention is that these causal diagrams would represent the various hypothesized causal models that according to an analyst or expert could have possibly generated the observed data. The scatter plot matrix presents the simulated data of one of these causal models, which is defined by the user in the arguments of the plotting method.

Designing the view of the simulated interventions in this way presents many advantages. Users can explore the expected outcomes from a variety of different interventions applied on a causal model of his choice. They can create a different view for each one of the hypothesized causal models. This allows them to explore the expected outcomes of interventional experiments under the various causal hypotheses of how data could have been generated. They can use the provided visual cues to test various causal hypothesis against the outputs generated by a concrete intervention under a specific causal hypothesis. This could help users to realize how informative a specific intervention can be given their assumptions.

6.4.2.3 The Design of the Scatter Plot Matrix

The scatter plot matrix presents only the data corresponding to the observed variables of the probabilistic model, which are the variables of the causal models. The data of the parameters of the probabilistic model are not presented here as they are irrelevant to the tasks of causal reasoning we focus on.

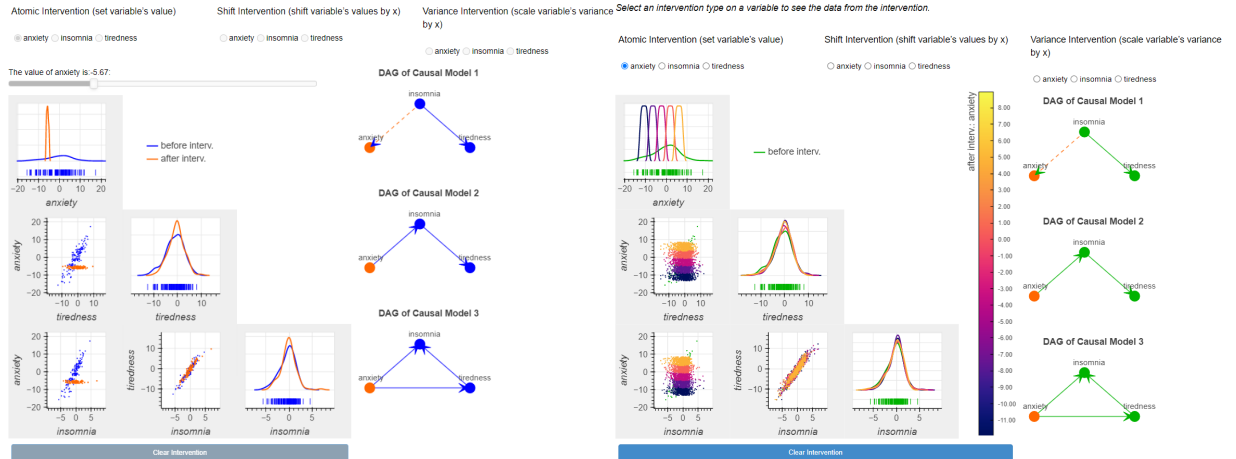
The outlook of the scatter plot matrix is similar to that of the IPP. The lower triangular part of the matrix is kept. The diagonal presents the KDE plots of the posterior predictive samples of the observed variables before (in blue for the interactive and animated mode or green in the static mode) and after (in orange for the interactive and animated mode or various colors determined by the provided colorbar for the static mode) the intervention. The rug plot of the posterior predictive samples before the intervention is presented below each KDE curve. The presence of the rug plot serves only the purpose of giving the user the sense of frequency and discretization for the KDE curve (the same principle was followed in the design of the IPME and IPP). Thus, presenting the simulated data after the intervention in a second rug plot below the KDE curve was omitted for reasons of decluttering. The scatter plots of the pairwise posterior predictive samples before and after the intervention are presented across the rows and columns of the matrix.

Select an intervention type on a variable to see the data from the intervention.



(a)

Select an intervention type on a variable to see the data from the intervention.



(b)

(c)

Figure 6.4: The (a) interactive, (b) animated, and (c) static visualization mode of the `vicausi` tool. The simulated data before the intervention is shown in blue in the interactive and animated visualization mode, and in green in the static. The simulated data after the intervention is shown in orange in the interactive and animated visualization mode, and in colors determined by the colorbar (colors are mapped to slices of data of increasing ranges) in the static.

6.4.2.4 The Presentation Problem of the Interventional Data

The scatter plot matrix in `vicausi` has very similar outlook to that in IPP. There is an important difference in the purpose each serves that led to a different implementation. The scatter plot

matrix in IPP was required to present only the posterior and posterior predictive samples of the probabilistic model's (observed and latent) variables. The scatter plot matrix in `vicausi` is required to visualize the posterior predictive samples both before and after an intervention at the same time. This is required to be in line with the second design objective.

Visualizing the posterior predictive samples before the intervention creates a view similar to IPP. In this case there is the interventional data that should be presented in conjunction with simulated data before the intervention to allow direct comparisons. Various possible design options were considered as to how this could be implemented. One option could be to visualize the whole set of interventional data for all considered interventional values along with the simulated data before the intervention using a different color. Some stratification of the interventional data could be used, e.g., in the form of interactive conditioning like in IPME or IPP or statically by using a different color for the strands. This would enable users to infer causal information like which variables in the model are affected by an intervened variable and the sensitivity of these variables to changes in the intervened variable.

This approach is reasonable and offers a good overview of the interventional data in a single view but can easily lead to cluttered views of the scatter plot matrix (for example see Fig. 6.4(c)). The number of the presented data points in the scatter plots is heavily dependent on the number of the different interventional values considered. More concretely, the size of the trace that holds the posterior predictive samples of an observed variable depends on the following parameters:

- the number N of observations that were used for the inference for this particular observed variable;
- the number S of samples drawn from the posterior predictive distribution (S samples are drawn for each one of the N data points).

The total size of such a trace is defined by $N \times S$. Assuming that there are I different interventional values considered, there would be I traces of size $N \times S$ of interventional data and another trace of same size $N \times S$ for the simulated data before the intervention, all required to be plotted in the same scatter plot matrix. In the case of IPP only a trace of size $N \times S$ would be required to be shown.

As the number of interventional values, I , increases the number of the data points that need to be presented increases linearly with the size of the trace, $N \times S$ leading to cluttered views. On the other hand, keeping the number of interventional values low to avoid cluttering decreases the granularity in the exploration of the resulted effects after a certain kind of intervention. To mitigate this a different approach in presenting the interventional data was considered. Specifically, it was considered to visualize only the interventional data corresponding to a single interventional value each time and provide a mechanism for querying and updating the scatter plot matrix with the interventional data corresponding to a queried interventional value. This approach would reduce the number of data points being plotted at the same time (only two traces

of simulated data of size $N \times S$ would need to be shown, one for before and the other for after the intervention) leading to a clearer view. It could also create more intuitive views of the interventional data by exploiting means like interaction or animation for the update of the scatter plot matrix.

Given these design options for the visualization of the interventional data, three different designs of the scatter plot matrix were considered in this work. Each one of them uses a different visualization mode for presenting the interventional data; statically, interactively, or with animation. A user study was conducted for evaluating the effect of the visualization design on users' causal reasoning in TT1 and TT2 tasks. These three visualization modes were used as the conditions in the user study, which will be presented in detail in Section 6.5. For the moment, the detail of each one of these visualization modes will be explained.

In the static mode (Fig. 6.4(c)), the whole set for all interventional values along with the simulated data before the intervention is presented. The simulated data of the interventions are stratified in 5 strands corresponding to 5 sequential, non-overlapping, and increasing value ranges of the intervened variable so that its whole value range is covered. This number of strands was chosen to provide enough granularity and at the same time avoid cluttering the plots too much by plotting many KDEs or groups of pairwise samples. The data corresponding to each strand is coloured based on a mapping of the strand's index to a color from a colorbar (the smallest index corresponds to the strand of the lowest value range and the highest index to the highest value range). A continuous palette was used from the `colorcet` Python library [Colorcet], and specifically the linear `bmy`. The continuous colormaps in this library were designed to be "perceptually uniform, with each new color equally perceptually distinct from the previous and following colors" [Colorcet].

In the interactive and animated mode (Fig. 6.4(a) and (b)), only a slice of interventional data is presented along with the simulated data before the intervention; the slice that corresponds to a single interventional value (uniform noise is added to interventional values to account for uncertainty present in actual realizations of interventions). A slider at the top of the scatter plot matrix indicates this interventional value. In the interactive mode the user can set the value of the slider and the scatter plot matrix gets updated once the slider is set to present the corresponding slice of the interventional data. In the animated condition, once a user selects a specific intervention from the radio button widgets an animation of the outcomes corresponding to all possible interventional values is triggered. The slider is automatically set to one interventional value after the other sequentially starting from the smallest and following each slider's value setting, the scatter plot matrix gets updated to show the corresponding slice of interventional data. Any interactive element (radio buttons, slider, clear button) is deactivated as long as the animation is played and once it finishes they get activated again. All subplots of the scatter plot matrix are updated simultaneously once the slider is set to a new value in both the interactive and animated mode. Videos demonstrating the visualization modes can be found in [Taka, 2023d].

The “Clear Intervention” button presented below the scatter plot matrix in all conditions was added to clear any selection of intervention, remove any annotation or highlighting from the causal diagrams and any presented interventional data from the scatter plot matrix, and resets the slider and the radio buttons. This could be useful especially in the static mode for creating a separate view of the simulated data before the intervention in case it hides behind the interventional data and is not clearly discernible.

6.4.2.5 The Design of Causal Diagrams

The causal diagrams of the hypothesized causal models are presented on the right of the scatter plot matrix. The variables are represented by nodes and their hypothesized causal relations by arrows starting from the cause and pointing to the effect. The DAGs are initially drawn in the same color as the posterior predictive samples of the variables before the intervention (blue in the interactive and animated mode and green in the static mode). After the intervention the node corresponding to the intervened variable is highlighted (it turns orange). The intention is to create a color consistency for the presentation of the information before and after an intervention.

In the case of an atomic intervention any incoming arrow to the node corresponding to the intervened variable is annotated with a dashed orange line. These arrows normally should be removed in standard causal diagrams. The choice here is to keep them and simply use some annotation to indicate the breaking of the link due to the intervention. In this way it will be ensured that users are aware that these links existed before the intervention at any point of their explorations. The incoming arrows to the node corresponding to the intervened variable should not be removed in the case of a shift or variance intervention. The reason is that in these interventions the values of the intervened variable are not completely set externally, but they are determined by the variable’s causes as usual. The intervention is then applied to the variable’s values to transform them either by shifting their mean value or scaling their standard deviation.

6.4.2.6 Implementation and Limitations

The development of the `vicausi` tool was based on the same technologies used for IPME and IPP (e.g., the Panel and Bokeh visualization libraries in Python), and utilized many code components from IPP. Of course, this code had to be extended to accommodate the extra features entailed by the need of visualizing and exploring the interventional data. The Python package `NetworkX`¹ was used for drawing the causal diagrams.

This concrete implementation of the `vicausi` tool presents some limitations; some of them were unavoidable, others could be avoided by elaborating the tool with extra features. The first major unavoidable limitation of the tool is that the presentation and exploration of the interventional data is restricted to the pre-computed sets of simulated data that are provided as

¹<https://networkx.org/>

input to the tool. The query and generation of interventional data on-the-fly and in real-time is not possible because the PPL-related simulators cannot produce posterior predictive samples for the model's variables in real-time at the moment.

Another limitation of the tool is that interventions can only be applied on a single variable at a time. Allowing combinations of interventions to be applied on the model would require to pre-compute all possible combinations of interventions and provide the outputs as input to the tool. This can be a tedious task most possibly leading to large `npz` files that would require to be loaded and hold in the computer memory every time an instance of `vicausi` is created. The greater the exploration flexibility offered, the larger the file with the pre-computed simulated data. This problem would go away if dynamic computation was available.

Other limitations of the current implementation of the tool are that only continuous variables are supported, the hypothesized models should consist of the same set of variables, and the density of the interventional variables, and hence the level of granularity in the interventional data, cannot be adjusted. Extra features could be added to the tool in the future to extend the range of possibilities to explore.

Finally, there is the possibility that the conclusions the user will reach through the interaction with the exploratory tool are not realistic because the assumptions applied do not reflect the actual data generating mechanism. The less informed (by prior knowledge and experience) the explorations using such tools and the more complex the assumed causal structures are, the greater the risk to reach unrealistic conclusions is. This might constitute a source of confusion or error for the users of the tool. Hence, the users should be conscious about and aware of these risks considering and testing as many causal assumptions as possible. The current design of the tool, which only presents the simulated data from a single hypothesized causal model and dataset in a single instance of the tool, might not encourage extensive explorations of assumptions. Alternative designs like the ones discussed in Section 6.6.2 could be more effective in raising users' attention to such risks and encourage them to conduct more thorough explorations.

6.4.3 Use Case

A sleep specialist wants to investigate the causal relations among insomnia, anxiety, and tiredness. He thinks that there are two possible causal structures for these three variables; first, insomnia is a mediator from anxiety to tiredness, and second, insomnia is a common cause of the other two variables. Also he has some ratings of patients' insomnia, anxiety, and tiredness levels that he maps in a continuous scale. He uses `vicausi` to test his hypotheses and explore the expected outcomes of various interventional experiments to decide which ones would be more informative to conduct.

He uses the implementation of the pipeline presented in Fig 6.1 provided in [Taka, 2023c] to generate the input files for `vicausi`; he inputs his two hypothesized causal structures and the available observed data to the simulation pipeline. The pipeline generates probabilistic models

similar to those in Fig. 6.2(a) for each one of the hypothesized causal structures, and conducts the inference for each one of them. Then it generates simulated interventional data for each one of the models, taking all variables in each model as the intervened variable, one after the other, and simulating data from applying each one of the three available types of intervention (i.e., atomic, shift, variance) on each intervened variable for a range of interventional values, equally spaced one from the other.

One of the explorations the sleep specialist can make with `vicausi` is to assume that the mediator model generates the data. He produces a view of the simulated data from this model using `vicausi` (Fig. 6.5 presents the initial view of `vicausi` that is generated).

Select an intervention type on a variable to see the data from the intervention.

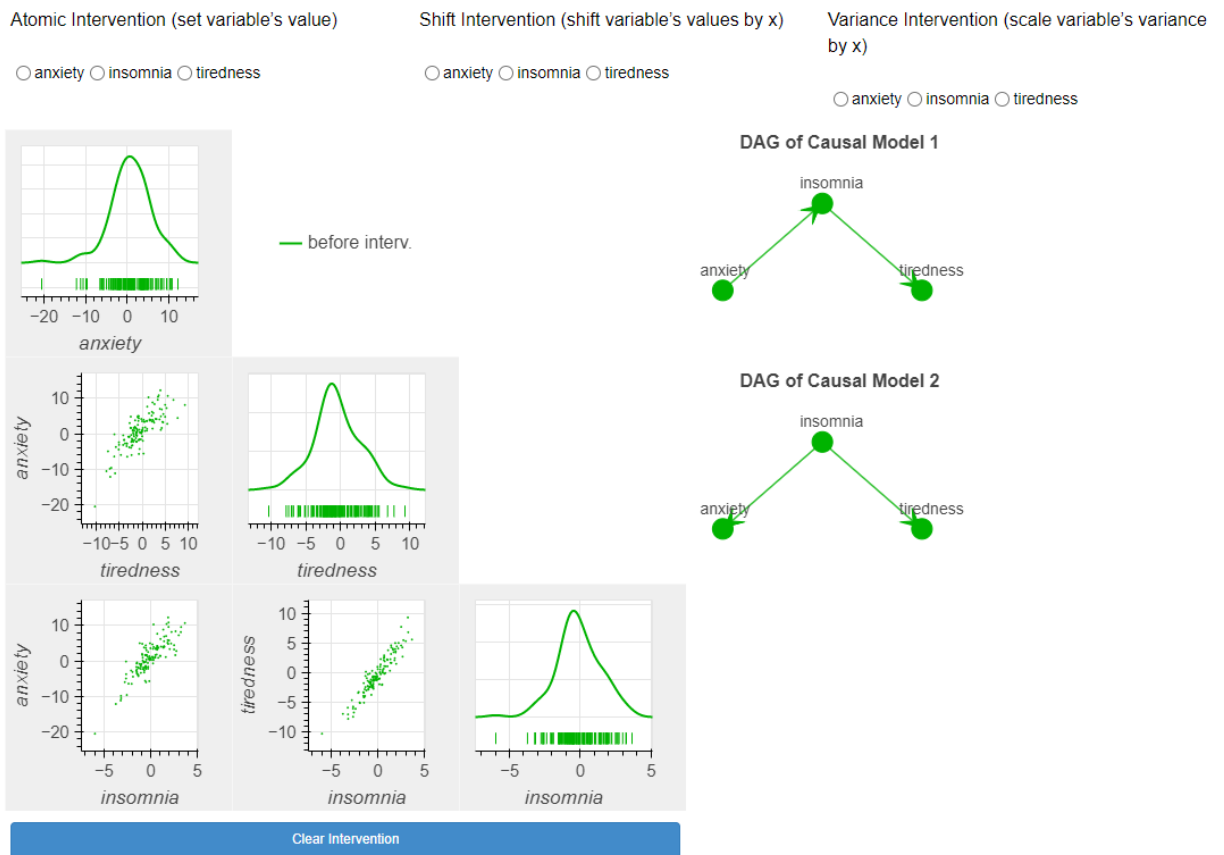


Figure 6.5: The `vicausi` view presenting posterior predictive samples (before any intervention) from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem.

He first observes the posterior predictive data generated by the probabilistic model of the mediator causal model. Any of the two hypothesized causal models could generate observations like the ones presented in the scatter plot matrix where all variables are positively correlated to each other. Causal model 1 dictates that any increase in insomnia is caused by an increase in anxiety, and any increase in tiredness is caused by an increase in insomnia. Although there is no direct link (arrow) between anxiety and tiredness, anxiety affects tiredness indirectly through

insomnia (the pipe effect). So, based on this causal model, anxiety and tiredness would appear correlated. Causal model 2 dictates that any increase in anxiety or tiredness is caused by an increase in insomnia. Although there is no link (arrow) between anxiety and tiredness, they both have a common cause, insomnia, which when increases causes an increase to tiredness and anxiety at the same time. So, plotting the data of tiredness and anxiety in a scatter plot, each will seem to increase when the other increases, although none of them causes the other. Both causal models are plausible explanations of the process that generated this data.

The data before the intervention alone cannot tell if and which of the variables cause other variables in the model. The data of insomnia, anxiety, and tiredness occur naturally through a data generating mechanism described possibly by causal model 1 or 2. The sleep specialist could make an intervention externally on this data generating mechanism, and by collecting the new data that will be generated after the intervention make some inferences about the causal relations of the variables.

Intervention allows the identification of the variables that an intervened variable affects. A change applied on a variable would affect only the variables caused (directly or indirectly) by this variable and leave the remaining variables of the model unaffected. By an intervention, controlled changes are applied to the value of a variable. Other variables that are affected by this variable are identified by spotting changes to them.

The sleep specialist uses `vicausi` to visualize the simulated interventional data to explore various interventions and observe to what extent each intervention would help him to determine the causal relations among the variables of interest. The exploration of three of these interventions is demonstrated in the following paragraphs.

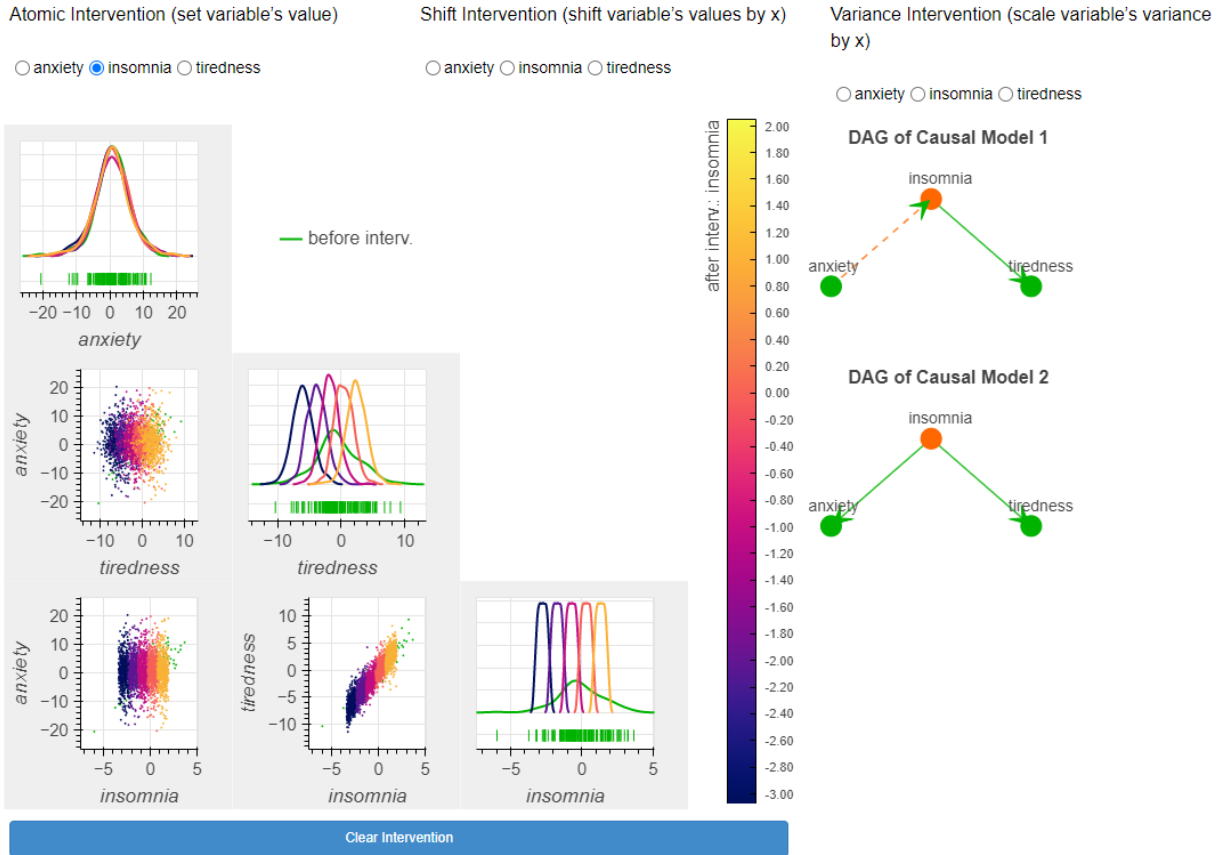
6.4.3.1 Atomic Intervention On Insomnia

The sleep specialist selects the atomic intervention on insomnia from the radio button widgets above the scatter plot matrix. The simulated interventional data of setting insomnia's level to a specific value in the interactive mode (Fig. 6.6(b)-(d)) or ranges of different values in the static mode (Fig. 6.6(a)) is shown. The observations that can be made by setting the level of insomnia to increasing values are the following:

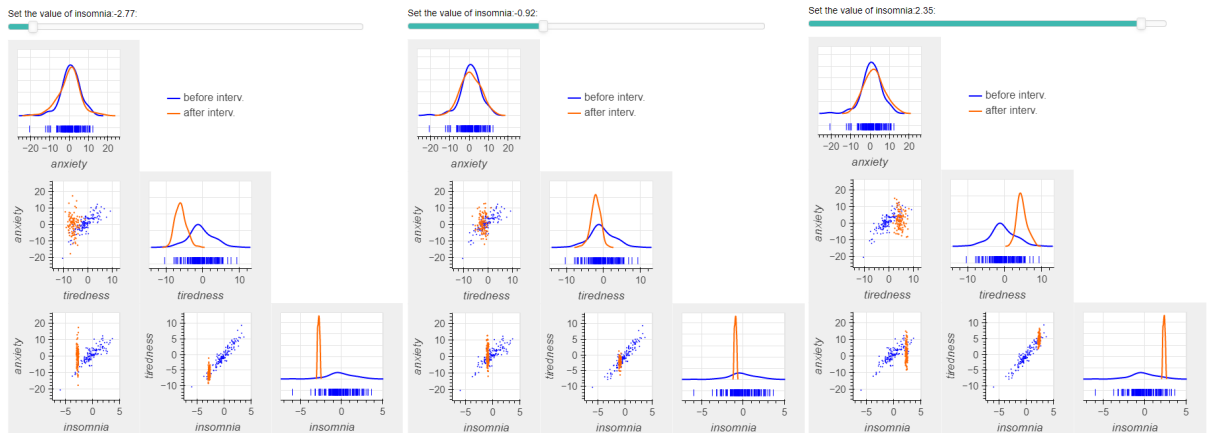
1. tiredness increases. There is a shift of tiredness's range of values to higher values in the scatter plot of tiredness-insomnia and tiredness's KDE plot. This means that *insomnia affects tiredness*;
2. anxiety is not affected. Anxiety's range of values remains essentially unaffected after the intervention in the scatter plot of anxiety-insomnia and anxiety's KDE plot. This means that *anxiety is not affected by insomnia*;
3. tiredness and anxiety appear uncorrelated after the intervention. This can be inferred by the fact that the slices of interventional data in the anxiety-tiredness scatter plot are

uncorrelated as insomnia's value increases, or by that anxiety's KDE does not change as that of tiredness's shifts to higher values.

Select an intervention type on a variable to see the data from the intervention.



(a)



(b)

(c)

(d)

Figure 6.6: The vicausi view presenting posterior predictive samples before and after an atomic intervention on insomnia from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem. (a) The static visualization mode. (b)-(d) Three instances of the interactive visualization mode with each corresponding to a different interventional value.

How could these 3 observations be used to tell which of the two hypothesized causal models generated the data? Both causal model 1 and 2 dictate that *insomnia affects (causes) tiredness* (there is an arrow from insomnia to tiredness) based on the causal diagrams of the two models on the right of the scatter plot matrix (*both causal models confirm observation 1*). Causal model 1 dictates also that *anxiety is not expected to change* after this intervention because anxiety is the cause of (and not caused by) insomnia. In causal model 2 *insomnia causes anxiety* (there is an arrow from insomnia to anxiety) and a change in anxiety would be expected by changing the value of insomnia (*only causal model 1 confirms observation 2*).

Insomnia's value is completely controlled (set) externally in an atomic intervention. This means that any cause of insomnia will no more be able to affect the value of insomnia. In causal model 1, anxiety won't affect insomnia any more after the intervention because insomnia is manipulated externally. For this reason, the correlation of anxiety and insomnia breaks (the link between them breaks and is denoted by a dashed orange arrow in the causal models' DAG). This will make *anxiety unable to affect tiredness* through insomnia and their correlation will also break. In causal model 2 *insomnia would affect tiredness*, so we would expect to see a correlation between them after the intervention (*only causal model 1 confirms observation 3*).

As a result of this exploration, the sleep specialist can conclude that an atomic intervention on insomnia assuming that the data is generated by the mediator causal model provides sufficient information for distinguishing the two causal models and identify which one is that describing the actual data generating mechanism of the variables of interest.

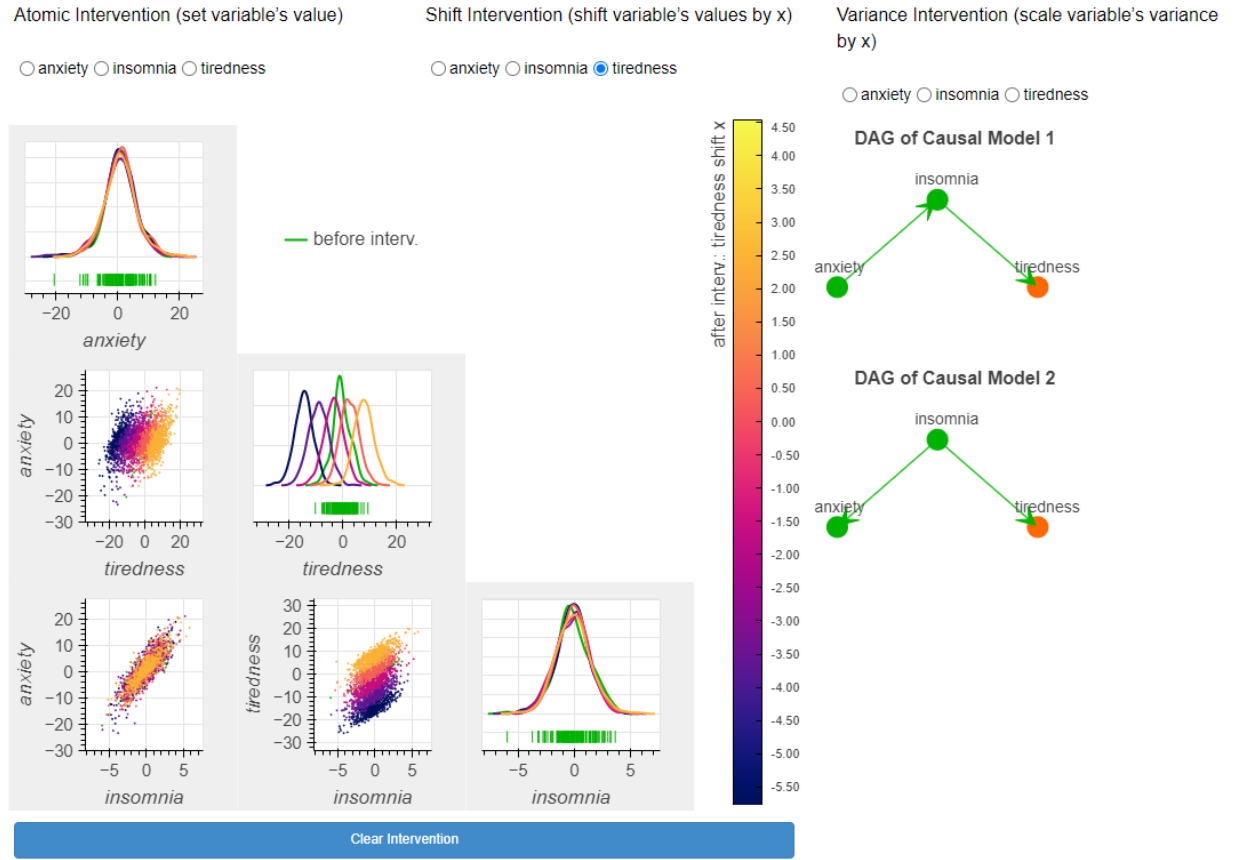
6.4.3.2 Shift Intervention On Tiredness

The sleep specialist selects the shift intervention on tiredness from the radio button widgets above the scatter plot matrix. The simulated interventional data of setting tiredness's level to a specific value in the interactive mode (Fig. 6.7(b)-(d)) or ranges of different values in the static mode (Fig. 6.7(a)) is shown. The observations that can be made by shifting tiredness's mean value by increasing factors are the following:

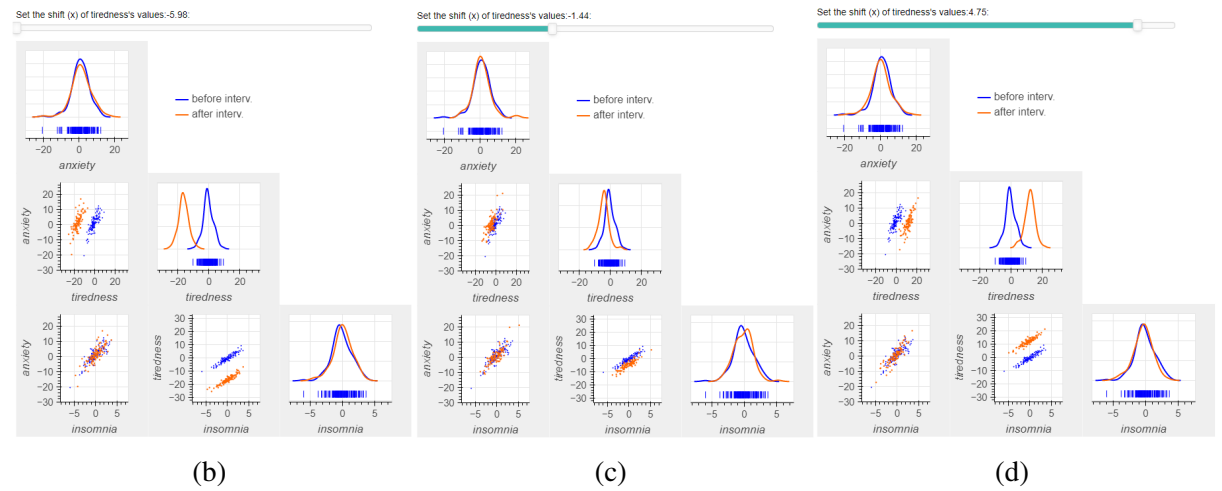
1. insomnia is not affected. Insomnia's range of values remains essentially unaffected after the intervention in the scatter plot of tiredness-insomnia and insomnia's KDE plot. This means that *insomnia is not affected by tiredness*;
2. anxiety is not affected. Anxiety's range of values remains essentially unaffected after the intervention in the scatter plot of anxiety-tiredness and anxiety's KDE plot. This means that *anxiety is not affected by tiredness*.

Causal model 1 dictates that tiredness is caused by anxiety through insomnia. This means that *anxiety and insomnia will not be affected* when intervening on tiredness because tiredness does not cause but is caused by them. Causal model 2 dictates that tiredness and anxiety are

Select an intervention type on a variable to see the data from the intervention.



(a)



(b)

(c)

(d)

Figure 6.7: The *vicausi* view presenting posterior predictive samples before and after a shift intervention on tiredness from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem. (a) The static visualization mode. (b)-(d) Three instances of the interactive visualization mode with each corresponding to a different interventional value indicated by the interactive slider at the top of the scatter plot matrix.

caused by insomnia. Thus, similarly, this means that *anxiety and insomnia will not be affected* when intervening on tiredness because tiredness does not cause but is caused by insomnia and

does not affect anxiety (*both causal models confirm observation 1 and 2*).

As a result of this exploration, the sleep specialist can conclude that a shift intervention on tiredness assuming that the data is generated by the mediator causal model does not provide sufficient information for distinguishing the two causal models and identify which one is that describing the actual data generating mechanism of the variables of interest.

6.4.3.3 Variance Intervention On Anxiety

The sleep specialist selects the variance intervention on anxiety from the radio button widgets above the scatter plot matrix. The simulated interventional data of setting anxiety's level to a specific value in the interactive mode (Fig. 6.8(b)-(d)) or ranges of different values in the static mode (Fig. 6.8(a)) is shown. The observations that can be made by scaling anxiety's variance by increasing factors are the following:

1. insomnia is affected. Insomnia's range of values increases after the intervention in the scatter plot of anxiety-insomnia and insomnia's KDE plot. This means that *insomnia is affected by anxiety*;
2. tiredness is affected. Tiredness's range of values increases after the intervention in the scatter plot of anxiety-tiredness and tiredness's KDE plot. This means that *tiredness is affected by anxiety*.

Causal model 1 dictates that anxiety causes tiredness through insomnia. This means that *insomnia and tiredness will be affected* when intervening on anxiety because anxiety influences the value of insomnia and tiredness. Causal model 2 dictates that anxiety and tiredness are caused by insomnia. This means that *insomnia and tiredness will not be affected* when intervening on anxiety because anxiety does not cause but is caused by insomnia and does not affect tiredness (*only causal model 1 confirms observation 1 and 2*).

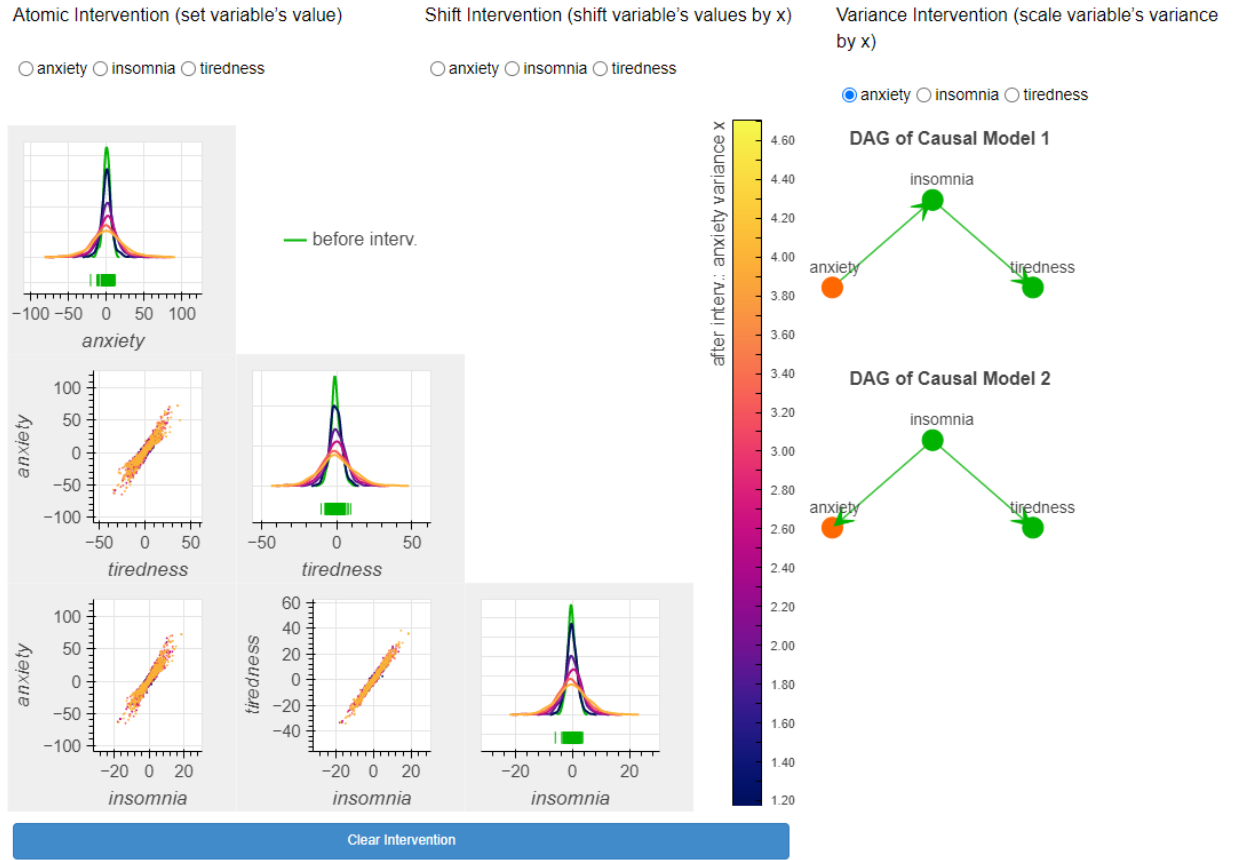
As a result of this exploration, the sleep specialist can conclude that a variance intervention on anxiety assuming that the data is generated by the mediator causal model provides sufficient information for distinguishing the two causal models and identify which one is that describing the actual data generating mechanism of the variables of interest.

6.5 Evaluation User Study

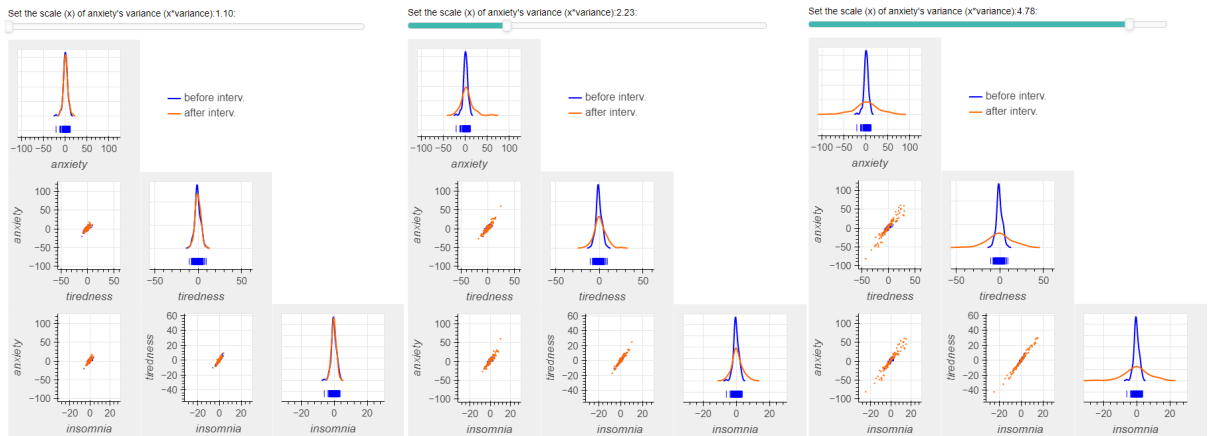
6.5.1 Research Questions, Conditions & Participants

An evaluation user study was designed and conducted with the aim to investigate the effect of the way that the simulated interventional data is presented on users' ability to reason causally. The concrete research questions that were investigated through this user study were the following:

Select an intervention type on a variable to see the data from the intervention.



(a)



(b)

(c)

(d)

Figure 6.8: The *vicausi* view presenting posterior predictive samples before and after a variance intervention on anxiety from the mediator causal model (Causal Model 1) of the insomnia-anxiety-tiredness problem. (a) The static visualization mode. (b)-(d) Three instances of the interactive visualization mode with each corresponding to a different interventional value.

Does interaction or animation when used for the visualization of simulated probabilistic interventional data help users

RQ1 *identify the causal structure of some variables of interest more accurately, faster, and with*

more confidence?

RQ2 *make decisions on interventional experiments more accurately, faster, and with more confidence?*

RQ1 and RQ2 were the two leading reasearch questions in this user study. However, the user study was designed in a way to provide insight into the ways the participants exploited the provided information and visual means to reason causally. Specifically, the user study was designed so that the subsequent analysis of the collected data would allow the investigation of the following aspects:

- ASP1 how people performed in causal reasoning tasks when simulated probabilistic data before and after an intervention was presented to participants in the form of a scatter plot matrix and in three different visualization modes: static, interactive, and animated. RQ1 and RQ2 are part of this investigation;
- ASP2 whether there were any specific settings (e.g., intervention types, causal models) in which participants were able to perform better under these circumstances;
- ASP3 what strategies participants used to identify the causal model when they explored the simulated data from multiple interventions;
- ASP4 whether the statistical or causal inference level of knowledge played a role in participants' performance;
- ASP5 whether there was any strategy followed in how the information presented in the scattter plot matrix was used to reason causally;
- ASP6 whether the visualization mode played any role in the adoption of any such strategy;
- ASP7 to what extent participants' responses were valid;
- ASP8 how participants judged the design of the visualization.

The design of the user study will now be discussed. The investigation aspect that is enabled by the various design choices will be noted in parenthesis using the ASPx codes provided above.

The user study followed a between-subject design with three conditions; the static, interactive, and animated visualization mode of the `vicausi` tool as described in Section 6.4.2 and shown in Fig. 6.4 (ASP1, ASP5, ASP6). 32 participants were randomly assigned in the **static group (SG)**, **interaction group (IG)**, and **animation group (AG)** (11 participants in each of IG and AG, and 10 participants in SG). All participants in the SG, IG, or AG were exposed to only the assigned version of `vicausi`.

Participants were recruited through mailing lists and social media of the institution and personal contacts. They each received a £10 worth online shopping voucher as a compensation.

There was no requirement regarding participants' statistical background to participate in the user study. They were only required to be over 18 years old, have a personal computer (laptop/desktop/tablet etc.) with an at least 11" screen, and be able to read graphs. The intention was to evaluate *vicausi* based on the general audience independently of their statistical or causal inference knowledge. Users who need to use such tools (e.g., researchers, health experts, industry partners) do not often acquire specialized knowledge in these fields but they still need to conduct causal inference and make decisions based on this. An appropriate training was provided to all participants as it will be explained in the following section. In this way it was ensured that all participants were provided with the required knowledge for dealing with the user study's tasks.

6.5.2 Study's Structure

The study was approved by the institution's ethics review board (approval number: 300220041) and conducted online through a web interface developed for this purpose. It consisted of four parts presented to participants in the following order; **demographic questions, training, tasks, and user experience questions.**

In the demographic questions' part, five questions (D1-5) were asked to participants. They were all multiple-choice allowing participants to input any response that was not included in the provided options. The demographic questions asked to participants are presented in the following list. Fig. 6.9 presents participants' demographic statistics.

D1 *What is your age?* 5 options were provided including a "Prefer not to say" option.

D2 *What is the highest level of education you have completed?* 4 options were provided plus the "Other" option allowing another response being input in a text box.

D3 *What former training in statistics have you undertaken?* (ASP4) 5 options were provided plus the "Other" option allowing another response being input in a text box.

D4 *Could you confidently state the difference between correlation and causation?* (ASP4) 4 Likert-scale options were provided.

D5 *Are you aware of having any color vision deficiency?* Yes/no options were provided.

The training part consisted of six pages (find screenshots in Appendix D.2) presented to participants through the web interface of the user study. It consisted mainly of reading and comprehension and demonstration of the *vicausi* tool. The training was similar in all conditions; only the presented visualization mode of the *vicausi* tool varied among the static (SG), interaction (IG), and animated (AG) group. The training's material constituted an introduction to the purpose of the study and concepts like causal diagram, difference of correlation and causation, scatter plot matrix, intervention and the types of intervention considered in the

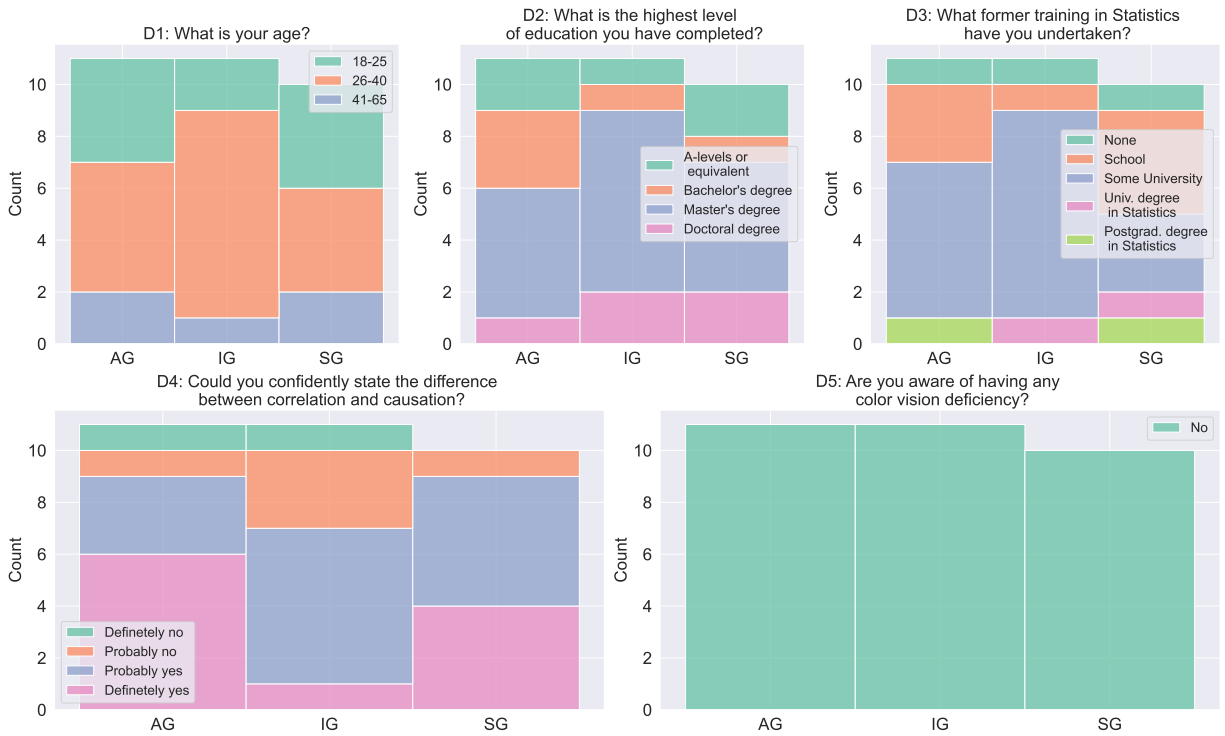


Figure 6.9: **Demographic statistics** of participants in the user study.

user study (atomic, shift, variance). A demonstration of the *vicausi* tool, and a use case of identifying the true causal model by using this tool similar to that presented in Section 6.4.3 were also shown to participants during the training.

The next part after the training was the study tasks. The study tasks were split into two parts each consisting of tasks of a different type; task type TT1 accounting for RQ1, and task type TT2 accounting for RQ2. There were 10 TT1 tasks and 6 TT2 tasks included in the user study. Tables 6.1 and 6.2 summarize the TT1 and TT2 tasks used in the user study, respectively. Screenshots of all tasks in the user study can be found in Appendix D.3. Section 6.5.3 explains how these tasks were designed.

The last part of the user study consisted of 4 user experience questions (UE1-4). The aim of these questions was to understand if participants' responses were informative or not (how often did they respond uninformatively?) (UE1), how they used the scatter plot matrix (did they rely on the scatter plots or KDE plots more often?) (UE2-3), and how they judged the design of the visualizations (UE4). The exact questions were as following:

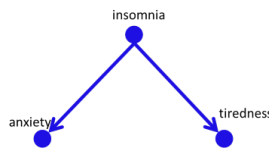
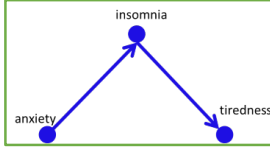
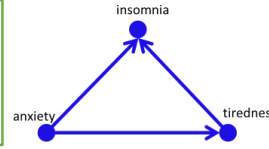
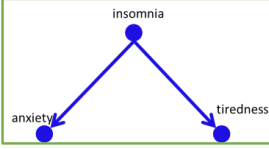
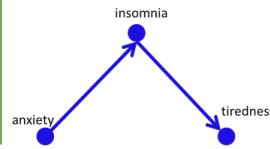
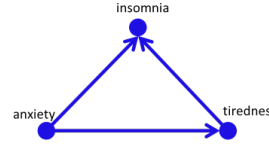
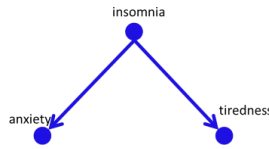
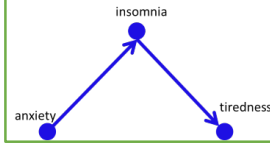
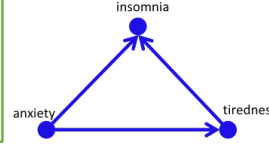
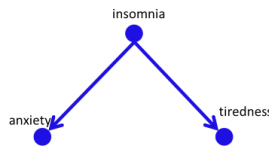
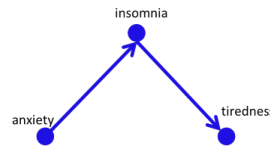
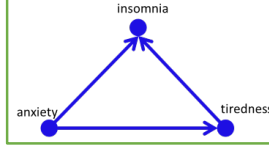
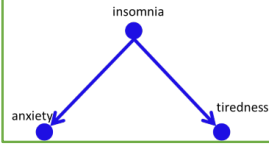
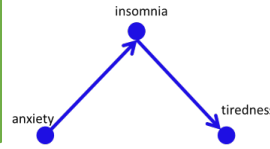
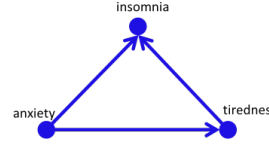
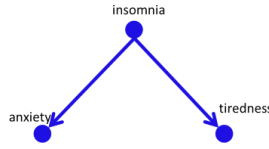
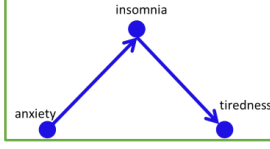
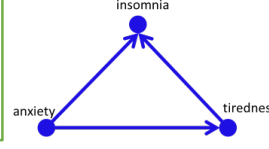
UE1 *In how many tasks in the study do you estimate that you looked at the directed graphs (DAGs) of the causal models?* (ASP7) A slider was provided with options the integers from 0 to 16;

UE2 *In how many tasks in the study do you estimate that you looked at the scatter plots (plots presenting data with dot markers) in the scatter plot matrix?* (ASP5) A slider was provided with options the integers from 0 to 16;

Table 6.1: Summary of TT1 tasks of user study.

Task	Question	Causal Models
		<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid green; padding: 2px;">Correct response.</div> <div style="border: 1px solid green; padding: 2px;">Correct response. The simulated data of this model is shown in the scatter plot matrix.</div> </div> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid red; padding: 2px;">Wrong response.</div> </div>
t1	A sleep doctor conducts an <i>atomic</i> intervention on patients' <i>tiredness</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid green; padding: 5px;"> </div> <div style="border: 1px solid red; padding: 5px;"> </div> </div>
t2	A sleep doctor conducts an <i>atomic</i> intervention on patients' <i>insomnia</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid red; padding: 5px;"> </div> <div style="border: 1px solid green; padding: 5px;"> </div> </div>
t3	A sleep doctor conducts an <i>atomic</i> intervention on patients' <i>anxiety</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid green; padding: 5px;"> </div> <div style="border: 1px solid green; padding: 5px;"> </div> </div>
t4	A sleep doctor conducts an <i>atomic</i> intervention on patients' <i>tiredness</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid green; padding: 5px;"> </div> <div style="border: 1px solid red; padding: 5px;"> </div> </div>
t5	A sleep doctor conducts an <i>atomic</i> intervention on patients' <i>anxiety</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid red; padding: 5px;"> </div> <div style="border: 1px solid green; padding: 5px;"> </div> </div>
t6	A sleep doctor conducts a <i>shift</i> intervention on patients' <i>tiredness</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid red; padding: 5px;"> </div> <div style="border: 1px solid green; padding: 5px;"> </div> </div>
t7	A sleep doctor conducts a <i>shift</i> intervention on patients' <i>anxiety</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid green; padding: 5px;"> </div> <div style="border: 1px solid red; padding: 5px;"> </div> </div>
t8	A sleep doctor conducts a <i>variance</i> intervention on patients' <i>anxiety</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid green; padding: 5px;"> </div> <div style="border: 1px solid red; padding: 5px;"> </div> </div>
t9	A sleep doctor conducts a <i>variance</i> intervention on patients' <i>insomnia</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid red; padding: 5px;"> </div> <div style="border: 1px solid green; padding: 5px;"> </div> </div>
t10	A sleep doctor conducts a <i>variance</i> intervention on patients' <i>anxiety</i> . Which of the hypothesized causal models generated this data?	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid green; padding: 5px;"> </div> <div style="border: 1px solid green; padding: 5px;"> </div> </div>

Table 6.2: Summary of TT2 tasks of user study. The question asked in all tasks was the following: “You want to design and run **one** interventional experiment, which will help you identify the causal model of the data. Which of the provided interventions (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.” The ticked boxes in the interventions column represent the correct response in each task.

Task	Interventions	Causal Models	<input checked="" type="checkbox"/> Correct response. The simulated data of this model is shown in the scatter plot matrix.	
t11	<input checked="" type="checkbox"/> Atomic on insomnia <input type="checkbox"/> Atomic on anxiety <input type="checkbox"/> Atomic on tiredness			
t12	<input checked="" type="checkbox"/> Atomic on insomnia <input checked="" type="checkbox"/> Atomic on anxiety <input type="checkbox"/> Atomic on tiredness			
t13	<input checked="" type="checkbox"/> Shift on insomnia <input type="checkbox"/> Shift on anxiety <input type="checkbox"/> Shift on tiredness			
t14	<input checked="" type="checkbox"/> Shift on insomnia <input type="checkbox"/> Shift on anxiety <input checked="" type="checkbox"/> Shift on tiredness			
t15	<input checked="" type="checkbox"/> Variance on insomnia <input checked="" type="checkbox"/> Variance on anxiety <input type="checkbox"/> Variance on tiredness			
t16	<input checked="" type="checkbox"/> Variance on insomnia <input type="checkbox"/> Variance on anxiety <input type="checkbox"/> Variance on tiredness			

UE3 In how many tasks in the study do you estimate that you looked at the KDE plots (plots presenting the variables’ distribution) on the diagonal of the scatter plot matrix? (ASP5)

A slider was provided with options integers from 0 to 16;

UE4 How did you find the design of the visualization for the simulated data? Was it informative, annoying, distracting? (ASP8) This was an open-ended question. The participants were asked to input their feedback in a text box.

6.5.3 Design of Study's Tasks

Two types of tasks were designed for this user study; task type *TT1* and task type *TT2*, each accounting for the corresponding research question.

In *TT1* tasks participants were asked to identify the causal model of the presented simulated data (before and after an intervention) given a list of 2 different models and a concrete intervention (a specific type of intervention - atomic, shift, or variance- applied on a specific variable of the model) (Fig. 6.10(a)).

In *TT2* tasks participants were required to explore the outcomes of a specific type of intervention applied on any of the models' variables and identify two things: first, the causal model of the presented simulated data from a list of 3 different models, and second, the interventions that each provides sufficient information to identify the causal model (Fig. 6.10(b)). Tasks of type *TT2* were more complex than the tasks of type *TT1*.

The aim of *TT1* tasks was to investigate if users can reason causally when they are provided with simulated probabilistic data (before and after an intervention) presented in a scatter plot matrix. These tasks preceded the more complex *TT2* tasks with the aim to help users to familiarize themselves with the environment and give them the chance to pre-establish a strategy for exploiting the presented information to reason causally before moving on to the more complex tasks of type *TT2*.

The aim of *TT2* tasks was to emulate the task of a specialist or researcher who wants to design an interventional experiment. The intention was to investigate if they can make informative decisions on which interventional experiments to conduct when they rely on simulated probabilistic data of interventions. The complexity was increased in *TT2* tasks in comparison to *TT1* tasks in terms of two aspects; first, an extra causal model was added in the list of provided options, and second, the participants were required to explore information from multiple interventions. The intention was to make the tasks of type *TT2* more realistic by aligning them with the tasks that an actual user would like to conduct.

All questions were multiple-choice. A single selection was allowed for determining the true causal model in *TT1* tasks. Multiple selections were allowed for determining the sufficient interventions in *TT2* tasks, and single selection was allowed for determining the true causal model. At the end of each task participants recorded their confidence about their response. This was input in a five level Likert scale.

A variety of combinations of intervention type - atomic, shift, and variance - and intervened variable was included in *TT1* tasks (ASP2). The provided intervention in most *TT1* tasks was informative enough to identify which one of the two provided causal models is the one having generated the presented data. Only in two tasks (t_3 and t_{10}) the intervention was not providing enough information to distinguish the two causal models. In 5/10 of the *TT1* tasks an atomic intervention was applied, in 2/10 a shift, and in 3/10 a variance intervention.

In *TT2* tasks all types of interventions were considered (ASP2). Two *TT2* tasks for each

Question 1 out of 10:

Time is running...

A sleep doctor conducts an **atomic intervention** on patients' **tiredness** to set its level to a specific value.

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of setting patients' tiredness to a range of different levels can be viewed.

Which is the causal model that generated this data?

Causal Model 1
 Causal Model 2
 Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

1 (not at all)
 2 (slightly)
 3 (somewhat)
 4 (fairly)
 5 (completely)

(a)

Question 1 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

Atomic intervention on insomnia
 Atomic intervention on anxiety
 Atomic intervention on tiredness
 None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

Causal Model 1
 Causal Model 2
 Causal Model 3

How confident are you about your answer?

1 (not at all)
 2 (slightly)
 3 (somewhat)
 4 (fairly)
 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

Atomic Intervention (set variable's value)

insomnia anxiety tiredness

Set the value of S:

(b)

Figure 6.10: (a) Task t1 (TT1) and (b) task t11 (TT2).

one of the atomic, shift, and variance interventions' type were included. There was at least one sufficient intervention in all TT2 tasks. This means that sufficient information was provided in all TT2 tasks to identify the causal model of the presented data (ASP3).

The tasks in each part were presented to participants in a randomized order to account for any learning effect. The order in which the variables were presented in the scatter plot matrix and causal diagrams varied across the tasks but was fixed across participants (see screenshots

in Appendix D.3 to see how the order of variables varied among the tasks). The reason for not keeping the order of variables fixed in all tasks was again to avoid any learning effect.

The insomnia-anxiety-tiredness problem was used as reference in all tasks of the user study. This model was referenced in the existing literature [Lagnado and Sloman, 2004] and was chosen because any direction of the causal link between the variables could be deemed as reasonable. For example, anxiety causes insomnia or insomnia causes anxiety could both be possible facts. In a different model, which for example could include age as a variable, only the fact that age affects insomnia could be possible and not the inverse, namely insomnia affects the age of the patient. It was also deemed that the general audience would not be able to rely on their previous experience or knowledge to infer the causal relations in this model.

The insomnia-anxiety-tiredness model is a three-variable model. The two-variable models would be too trivial, while models with more than three variables would add extra unnecessary complexity given the purposes of the user study and the limited time for its completion by each participant. The three-variable model allows a good variety of possible causal structures to be explored.

Three specific causal models describing the causal relations among insomnia, anxiety, and tiredness were considered (the ones shown in Table 6.2); in the first, insomnia is a common cause of anxiety and tiredness, in the second, insomnia is a mediator from anxiety to tiredness, and in the third, insomnia is a common effect of anxiety and tiredness. These models were not chosen to necessarily reflect the true scientific knowledge about how these three variables might be related causally, but to allow the exploration of a variety of possible causal relations. A mediator, a common cause, and a collider model were the three causal structures considered. These types of causal structures have been included in the tasks of similar user studies [Kale et al., 2022; Lagnado and Sloman, 2004; Yen et al., 2019].

Synthetic observations were created from the common cause model so that all three variables are positively correlated to each other (find exact code in Appendix D.1). These observations along with the three considered causal models were then used in the pipeline presented in Fig 6.1 and provided in [Taka, 2023c] to generate the input files for `vicausi`. The `npz` files that were used in the user study and the code for generating them can be found in the `/examples` folder of the github repository in [Taka, 2023c].

6.5.4 Statistical Analysis & Results

This subsection presents the statistical analysis of the collected data from the user study and discusses the results. The purpose of this analysis is to investigate aspects ASP1-3. Section 6.5.4.1 presents the evaluation measures and the data used for the analysis. Section 6.5.4.2 discusses the analysis's process presenting all the levels of inference considered. Sections 6.5.4.3-6.5.4.5 present the analysis of participants' performance, response time and confidence in TT1 and TT2 tasks. The data and the Python code of the analysis can be found in the github repository in

[Taka, 2023b].

6.5.4.1 Evaluation Measures & Data

Participants' *accuracy*, *response time* and *confidence* are the evaluation measures used for the statistical analysis. Accuracy is measured in a different way depending mainly on the type of the response input used (i.e., single- or multi-selection input).

In the case of a multi-selection response input, accuracy is measured as the number of correctly selected and non-selected options in the multiple-choice input by each participant in every task. Participants' responses in the multiple-choice input are transformed into a binary representation with 0 indicating a wrong and 1 a correct selection or non-selection. The binary representation of each participant's response in these cases consists of as many binary digits as the available options of the multiple-choice input excluding the "none" option, if present. Participants' accuracy in each task is computed as the number of occurrences of digit 1 in their response.

In the case of a single-selection response input, accuracy is measured as 0 or 1 depending on whether the participant selected a wrong or the correct option, respectively.

Participants' response time was measured (in seconds) from the moment the task was displayed until the answer was submitted. Participants rated their confidence in each task on a 1-5 scale with increasing level of confidence (1: not at all, 2: slightly, 3: somewhat, 4: fairly, 5: completely). This was remapped to the $\{-2, -1, 0, 1, 2\}$ set to center the parameterization.

Accuracy data consists of numbers of participants' correct selections or non-selections in the multiple-choice input, or a 0 or 1 in every task depending on the approach followed. Response time data consists of times (in sec). Confidence data consists of ordinal values. No participant or response is excluded from the collected data. Only the multiple blank registrations of some participants, who accidentally clicked the "Register" button multiple times, are omitted.

6.5.4.2 Bayesian Analysis & Levels of Analysis

A Bayesian analysis of the collected data is conducted. Various levels of inference are considered as allowed by the design of the user study and with the aim to investigate the research questions and the other aspects of interest. These levels of inference are summarized in the following list. They will be called *levels of analysis (LA)*.

LA1 All participants are pooled together independently of the condition (ASP1);

LA1.1 inference is conducted by taking participants' responses in the individual tasks separately (ASP2);

LA1.2 inference is conducted by taking participants' responses in all tasks together (ASP1);

LA1.3 inference is conducted by taking participants' responses in the tasks concerning the same type of intervention (i.e., the atomic, shift, and variance) together (ASP2).

LA2 Participants are pooled together based on the visualization condition (IG, SG, and AG) (RQ1 and RQ2);

LA2.1 inference is conducted by taking participants' responses in the individual tasks separately (ASP2);

LA2.2 inference is conducted by taking participants' responses in all tasks together (RQ1 and RQ2).

The LA1 level of analysis allows the investigation of participants' performance overall independently of the visualization mode that the participants used. The intention is to investigate whether and how participants were able to deal with the particular tasks given the probabilistic interventional data.

The LA2 level of analysis allows the investigation of considerable differences in participants' performance between the visualization conditions. The intention is to investigate if the specific way of presenting the interventional data plays any role in participants' performance.

The sublevels allow the investigation of considerable differences in participants' performance for particular tasks (LA1.1, 2.1) or intervention types (LA1.3). The aim is to investigate if there are particular causal conditions or intervention types that are more beneficial.

Sections 6.5.4.3 - 6.5.4.4 present the analysis of participants' accuracy separately for TT1 and TT2 tasks (ASP1-3). This is reasonable because the two types of tasks differ in the questions asked to participants and the response inputs. Thus, the approach followed for the analysis of participants' responses in each type of tasks presents some differences that need to be separately discussed. Section 6.5.4.5 presents the analysis of participants' response time and confidence for all tasks of the user study (ASP1-2).

6.5.4.3 Analysis of Accuracy in TT1 Tasks

Metrics of Accuracy In TT1 tasks (t_{1-10}) participants were able to select only one of three possible options: "Causal Model 1", "Causal Model 2", "Both causal models are possible". Two approaches are considered for measuring accuracy in these tasks.

The first approach considers that only one of the three available options is correct. A participant's response is marked as 0 when he selects a wrong option and as 1 when he selects the correct option. This approach will be called *Single Correct Option (SCO)* approach. This approach is reasonable but following this, there is a risk of missing some existing signal in the data. What if the correct answer is the "Both" (this is the case in tasks t_3 and t_{10}) and the participant identifies only one of the two models? What if the participant selects the "Both" option and only one of the two models is correct? Could a participant's response in these cases

be considered as completely wrong? For this reason, a second approach is considered with the aim to produce non-binary scores for measuring accuracy.

This second approach considers that any of the two models can be correct in each task. A participant's response is transformed into a binary representation consisting of 2 digits, one for each model. Each digit is marked as 0 when the participant thinks the corresponding model is wrong and as 1 when he thinks it is correct. The accuracy of a participant's response is measured as the number of correct selections and non-selections. This approach will be called *Multiple Correct Options (MCO)* approach.

Participants' responses in TT1 tasks are transformed in accuracy scores based on the SCO and MCO methods; that is 0s and 1s for the SCO and numbers in $\{0,1,2\}$ for the MCO. Fig. 6.11 presents the numbers of participants per task that scored each one of the possible values per task for both approaches independently of the visualization condition (LA1) and based on the visualization condition (LA2).

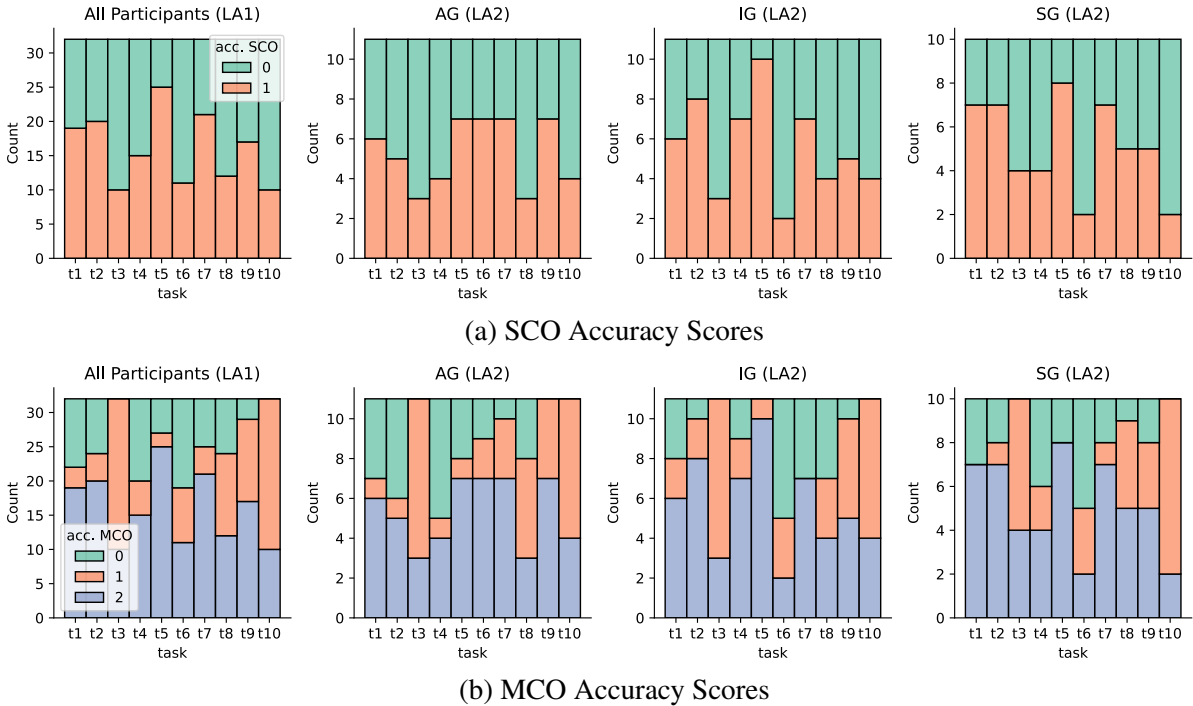


Figure 6.11: Count plots of number of participants scoring a particular level of accuracy per task independently of the visualization condition (LA1) and based on the visualization condition (LA2) for the (a) *Single Correct Option (SCO)*, and (b) *Multiple Correct Options (MCO)* approach.

Tasks t_3 and t_{10} in which the “Both” option is correct present the worst performance in SCO in comparison to the rest of the tasks (see the LA1 plot in the first row of Fig. 6.11). In this case participants who possibly identified only one of the two provided causal models were considered as having given a completely wrong response. Using MCO these participants could be assigned an intermediate accuracy score, namely 1. Following the MCO approach scores of

0 are not possible in such tasks (see that score 0 is missing from tasks t_3 and t_{10} in the plots of the second row in Fig. 6.11) because whichever model the participants select they would get 1 unless they select the “Both” option and get 2.

For the rest of tasks in which only one causal model is correct, the 0-score components (green bars) of the SCO count plots in the first row of Fig. 6.11 are broken into two separate components in the MCO count plots in the second row of Fig. 6.11; one component of completely wrong responses (0-score green bars) and a component of half-correct response (1-score orange bars). The half-correct components in all tasks except t_3 and t_{10} represent the cases when the participants select the “Both” option.

The SCO and MCO approaches for measuring accuracy are both used for the analysis of participants’ accuracy in TT1 tasks. They complement each other and enrich the analysis by providing two different perspectives at looking at participants’ choices.

Analysis Models Two models are created for each one of the SCO and MCO approaches to account for the two main levels of analysis (LA1 and LA2). The first model is used to conduct the inference on the level of all participants independently of the visualization condition (LA1). The second model is used to conduct the inference on the level of visualization conditions (LA2). Fig. 6.12 presents the Kruschke-style diagrams of all the Bayesian probabilistic models used for the analysis of accuracy in TT1.

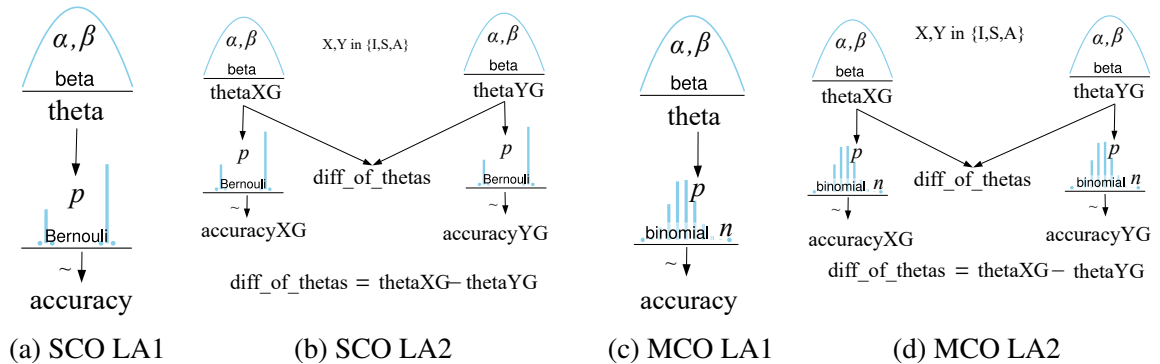


Figure 6.12: Kruschke-style diagrams of the analysis models for participants’ accuracy in TT1 tasks. Two models for each one of the *Single Correct Option (SCO)* and *Multiple Correct Options (MCO)* approaches for measuring accuracy are created. (a) SCO LA1 Model for all participants independently of visualization condition. (b) SCO LA2 Model for participants in each visualization condition. (c) MCO LA1 Model for all participants independently of visualization condition. (d) MCO LA2 Model for participants in each visualization condition.

Both models (LA1 and LA2) in each one of the SCO and MCO approaches are similar; they use the same type of likelihood and prior distributions. The likelihood of the two models in the SCO is a Bernoulli with a beta prior for the θ (represented by p in Fig. 6.12(a), (b)) parameter of the Bernoulli. The likelihood of the two models in the MCO is a binomial with number of trials $n = 2$ (one trial per causal model) and a beta prior for the θ (represented

by p in Fig. 6.12(c), (d)) parameter of the binomial.

Parameter θ expresses the *probability of success* in each trial. In SCO, θ represents the probability of a participant selecting the correct option from the three available: “Causal Model 1”, “Causal Model 2”, “Both causal models are possible”. In MCO, θ represents the probability of a participant to correctly select or not select each one of the available models: “Causal Model 1”, and “Causal Model 2”. In summary θ expresses the propensity of a participant to make a correct selection in a task within the context of these analyses.

The difference between the two models (LA1 and LA2) in each approach is that the second model (LA2) uses a separate likelihood and set of the priors for each visualization conditions. Also the second model estimates the difference between the groups’ (IG, SG, and AG) posterior distributions of θ .

Results Fig. 6.13 presents the inference results (posterior distributions) in the form of forest plots for all 4 models used for the analysis of participants’ accuracy in TT1 tasks. The orange vertical lines are reference lines representing the *chance* in the case of the θ s or the *zero difference* in the case of the differences between the conditions. The chance level for the θ depends on the approach used for measuring accuracy (i.e., the SCO or MCO). In the SCO approach the probability of randomly selecting the correct option from the three available is $1/3$. In the MCO approach the probability of randomly selecting or not selecting correctly an available option is $1/2$. Thus, the reference lines in the estimates of θ from the SCO models are set to $1/3$ and from the MCO models to $1/2$.

The horizontal bars represent the posterior highest density interval (the thinnest represent the 94% and the thickest the 50% HDI). The furthest away the horizontal bars are from the reference lines, the less likely the reference value is under the corresponding posterior distribution.

Based on the analysis conducted on the LA1 level both approaches of measuring accuracy (SCO and MCO) lead to estimations of θ that are better than the chance (Fig. 6.13(a) and (c)).

The LA1.2 level of analysis shows strong evidence that participants were able to perform better than chance (the “all tasks (LA1.2)” HDI bars in Fig. 6.13(a) and (c) are further on the right of the reference line and do not overlap with it).

The estimations of θ are more uncertain in LA1.1 than LA1.2 (the HDI bars in LA1.1 are wider than those in LA1.2 in Fig. 6.13(a) and (c)) because only the subset of participants’ responses corresponding to a specific task are used for the inference of these tasks’ posterior distributions. But still there is strong evidence in LA1.1 that participants performed better than chance in many tasks. Specifically, participants performed well in tasks t_1 , t_2 , t_5 , t_7 , t_9 based on both the SCO and MCO approaches. The performance in tasks t_3 and t_{10} becomes more certainly better than chance based on the MCO approach. The chance (value of reference line) seems to be quite possible under the posterior of θ for participants’ performance in

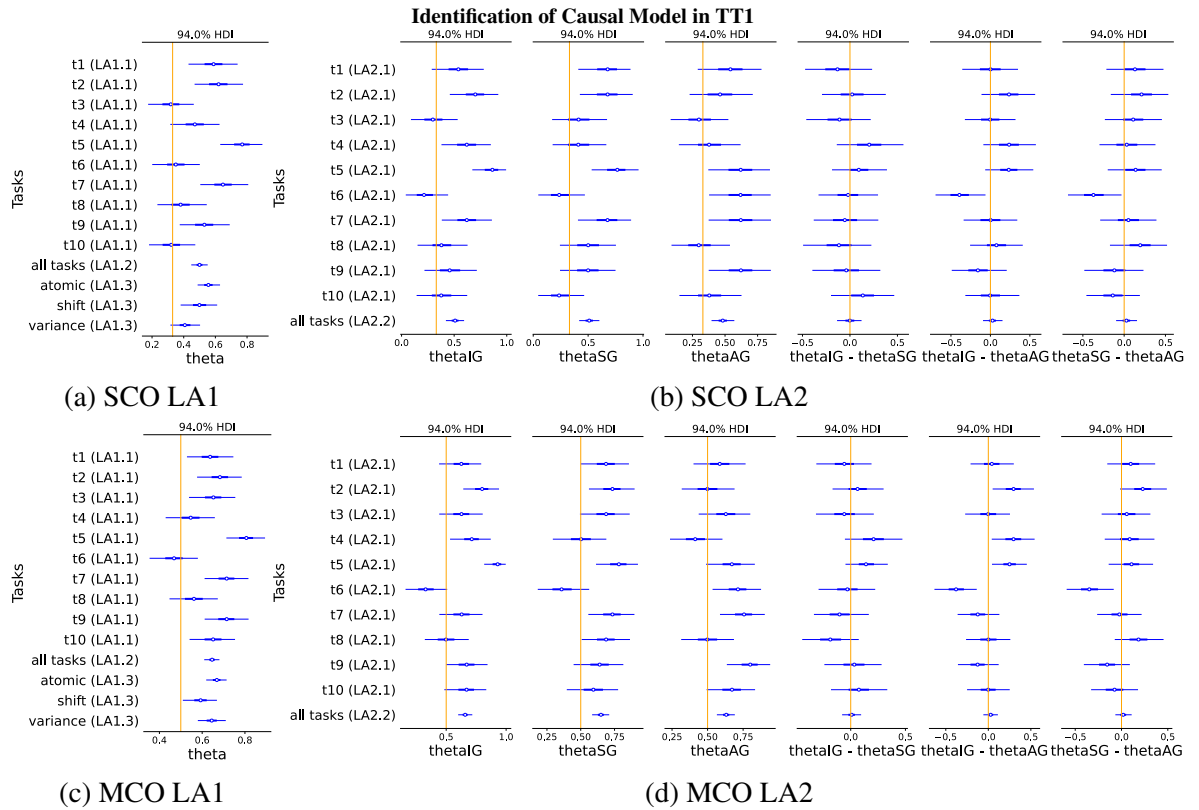


Figure 6.13: Forest plots of the posterior distributions inferred based on participants' accuracy in TT1 tasks. The results of the analysis models for both the *Single Correct Option (SCO)* and *Multiple Correct Options (MCO)* approaches for measuring accuracy are included. Forest plot of the (a) SCO LA1 Model for all participants independently of visualization condition, (b) SCO LA2 Model for participants in each visualization condition, (c) MCO LA1 Model for all participants independently of visualization condition, (d) MCO LA2 Model for participants in each visualization condition.

tasks t_4 , t_6 , and t_8 in LA1.1 for any of the SCO and MCO approaches.

All tasks t_1 , t_4 , and t_6 apply an intervention on tiredness with the first two applying an atomic and the third a shift intervention (see Table 6.1). These three tasks differ in which model's simulated data is presented to participants in the scatter plot matrix: in t_1 participants see the simulated data from the common cause model, and in t_4 and t_8 from the collider. The intervention on tiredness results in similar effects in the alternative causal models that are provided as a second option in these three tasks. Assuming that the order in which the variables are presented in the scatter plot matrices and causal diagrams does not play any important role in participants' causal reasoning, it could be inferred that the causal model whose simulated data is presented to participants plays a role in participants' ability to reason causally in this specific case.

The LA1.3 level of analysis shows strong evidence that participants were able to perform better than chance in the tasks of atomic interventions (the "atomic (LA1.3)" HDI bars in Fig. 6.13(a) and (c) are further on the right of the reference line and do not overlap with it).

The HDI bars of the tasks of shift interventions are the widest of the three in LA1.3 because only 2/10 TT1 tasks fall in this category in contrast to 5/10 in the atomic and 3/10 in the variance. The estimation of θ in the variance intervention tasks seems to move to higher values in the MCO approach. The reason for this is that one of the three tasks that fall into this category is task t_{10} for which participants' accuracy improves based on the MCO approach.

The LA2 level of analysis confirms partly these observations about the estimation of θ but this time within each visualization condition (Fig. 6.13(b) and (d)).

The LA2.2 level of analysis shows strong evidence that participants were able to perform better than chance (the “all tasks (LA2.2)” HDI bars of the θ estimations in each visualization condition in Fig. 6.13(b) and (d) are further on the right of the reference line and do not overlap with it), although θ is not estimated to be high in most of the visualization conditions in the SCO and MCO approach.

The estimations of θ are more uncertain in LA2.1 than LA2.2 (the HDI bars in LA2.1 are wider than those in LA2.2 in Fig. 6.13(b) and (d)) because only the subset of participants' responses corresponding to a specific task are used for the inference of these tasks' posterior distributions. The θ increases in tasks t_3 and t_{10} based on the MCO approach in LA2.1 in comparison to SCO but it does not seem to become clearly better than chance in any of the visualization conditions. The performance in the tasks that were found to be better than chance in the LA1.1 (t_1, t_2, t_5, t_7, t_9) is confirmed to be so in many cases in LA2.1, as well, but not in all visualization conditions.

The differences of the θ s between two visualization conditions in LA2 suggest an effect of one condition over the other when the value 0.0 (reference lines in columns presenting the differences of θ s and indicating no difference in Fig. 6.13(b) and (d)) becomes less likely under the difference of the posterior distributions. The effect of one condition becomes less likely when the highest density intervals of the posteriors of differences are pulled towards the reference value.

The inference on LA2.2 does not suggest any difference between any of the visualization conditions; the “all tasks (LA2.2)” HDI bars of the differences of θ s estimations in each visualization condition in Fig. 6.13(b) and (d) overlap with the reference line.

Similarly, the inference results in the LA2.1 level of the individual tasks do not seem to suggest any considerable difference between the visualization conditions. The HDI bars are consistently close and overlap with the reference line across the tasks.

Summary of Findings Overall participants' performance in TT1 tasks was better than chance independently of the visualization condition. The performance of participants in tasks in which the intervention was not informative enough to identify the causal model (tasks t_3 and t_{10}) is estimated to be better than chance when the MCO approach is used for measuring accuracy. Atomic interventions seemed to be well-interpreted by participants as their performance in TT1

tasks of atomic interventions was better than chance. Finally, in the case of distinguishing the collider from the common cause or mediator causal model through applying an intervention on tiredness, the model whose simulated data is presented to participants seems to play a role in their ability to identify the true causal model.

6.5.4.4 Analysis of Accuracy in TT2 Tasks

Metrics of Accuracy In TT2 tasks (t11-16) participants were asked to identify two things:

- all the interventions that were sufficient on their own to reveal the causal model of the presented data. Participants were given 3 interventions and had to select the ones that were sufficient (and not to select the ones that were not sufficient). A multiple-choice response input that allowed multiple selections was used for recording participants' responses.
- the true causal model of the presented simulated data. Participants had to select one of three possible options: "Causal Model 1", "Causal Model 2", "Causal Model3". A multiple-choice response input that allowed a single selection was used for recording participants' responses.

For the analysis of participants' responses in TT2 tasks, the MCO approach is used for measuring participants' accuracy in identifying the sufficient interventions and the SCO approach is used for measuring participants' accuracy in identifying the true model.

Analysis Models Similarly to the analysis of accuracy in TT1 tasks, a LA1 and a LA2 MCO Bayesian probabilistic model is created for the analysis of participants' responses in regards with the sufficient interventions, and a LA1 and a LA2 SCO Bayesian probabilistic model is created for the analysis of participants' responses in regards with causal model to account for the two main levels of analysis. These models are similar to those presented in Fig. 6.12.

Initially, the inference is conducted separately for participants' responses regarding the sufficient interventions and the causal model. Subsequently, only the sufficient interventions responses of participants who got the causal model correct are analysed.

There might be different ways in which participants could conclude which of the provided causal models is true for the presented simulated data. They might be able to combine information from the different available interventions to exclude one after the other the possible causal models until they reach to the one that is the true. They might also be able to determine an intervention that provides sufficient information on its own to identify the true model. Thus, the selections of models and interventions are initially analysed separately to investigate participants' performance independently of how they determined the causal model. The subsequent analysis of only the responses of participants who got the model correct aims at investigating to what extent their performance could be attributed to the identification of sufficient interventions (ASP3). All TT2 tasks included at least one sufficient intervention.

produces estimates of the θ taking participants' responses together independently of the visualization condition (Fig. 6.14(a)).

The LA1.2 level of analysis shows somewhat strong evidence that participants were able to perform better than chance (the “all tasks (LA1.2)” HDI bar in Fig. 6.14(a) is further on the right of the reference line and do not overlap with it). In LA1.1 there is strong evidence in LA1.1 that participants performed better than chance in tasks t_{11} , t_{12} , t_{14} (the “ t_{11} (LA1.2)”, “ t_{12} (LA1.2)”, and “ t_{14} (LA1.2)” HDI bars in Fig. 6.14(a) are further on the right of the reference line and do not overlap with it). The chance (value of reference line) seems to be quite possible under the posterior of θ for participants' performance in tasks t_{13} , t_{15} , and t_{16} based on LA1.1.

All tasks t_{11} , t_{13} , and t_{16} presented the simulated data of the mediator model (see Table 6.2) to participants. These tasks differ in the type of interventions that were provided as options to participants: the first allows atomic, the second shift, and the third variance interventions. Assuming that the order in which the variables are presented in the scatter plot matrices and causal diagrams does not play any important role in participants' causal reasoning, it could be inferred that the type of interventions that participants are able to explore plays a role in participants' ability to identify a mediator model.

All tasks t_{12} , and t_{15} presented the simulated data of the common cause model (see Table 6.2) to participants. These tasks differ in the type of interventions that were provided as options to participants: the first allows atomic interventions, and the second variance interventions. Assuming that the order in which the variables are presented in the scatter plot matrices and causal diagrams does not play any important role in participants' causal reasoning, it could be inferred that the type of interventions that participants are able to explore plays a role in participants' ability to identify a common cause model.

The LA1.3 level of analysis shows strong evidence that participants were able to perform better than chance in the tasks of atomic interventions (the “atomic (LA1.3)” HDI bar in Fig. 6.14(a) is further on the right of the reference line and do not overlap with it). On the other hand, chance seems to be possible under the posterior of θ in the tasks of variance interventions.

The LA2 level of analysis confirms partly these observations about the estimation of θ but this time within each visualization condition (Fig. 6.13(b)).

The LA2.2 level of analysis shows a rather fragile improvement in participants' performance than the chance (the “all tasks (LA2.2)” HDI bars in columns 1, 2, and 3 of Fig. 6.14(b) are on the right of the reference line and they nearly overlap with it). In LA2.1 level of analysis the HDI bars of the tasks that were found to be better than chance in the LA1.1 (t_{11} , t_{12} , t_{14}) nearly overlap with the reference line in all visualization conditions.

The inference on LA2.2 does not suggest any difference between any of the visualization conditions; the “all tasks (LA2.2)” HDI bars in columns 4, 5, and 6 of Fig. 6.14(b) overlap with the reference line. Similarly, the inference results in the LA2.1 level of the individual tasks do

not seem to suggest any considerable difference between the visualization conditions. The HDI bars are consistently close and overlap with the reference line across the tasks.

The *second row* of the figure (Fig. 6.14(c) and (d)) concerns the analysis of participants' selections of sufficient interventions conducted on the LA1 and LA2 levels. The LA1 level of analysis produces estimates of the θ taking participants' responses together independently of the visualization condition (Fig. 6.14(c)).

The LA1.2 level of analysis shows that chance is possible under the posterior of θ (the "all tasks (LA1.2)" HDI bar in Fig. 6.14(b) nearly overlaps with the reference line) and θ is estimated to be quite low. In LA1.1 only task t_{11} seems to indicate some evidence of better performance than the chance. Chance seems to be possible under the posterior of θ for participants' performance in the rest of tasks based on LA1.1 with the exception of task t_{14} that indicates a very poor performance. The LA1.3 level of analysis shows strong evidence that participants were able to perform better than chance in the tasks of atomic interventions (the "atomic (LA1.3)" HDI bar in Fig. 6.14(c) is further on the right of the reference line and do not overlap with it).

The LA2 level of analysis confirms partly these observations about the estimation of θ but this time within each visualization condition (Fig. 6.14(d)). In the LA2.2 level of analysis the "all tasks (LA2.2)" HDI bars of the θ estimations in columns 1, 2, and 3 of Fig. 6.14(d) overlap with the reference line in all visualization conditions. In LA2.1 the HDI bars of the task t_{11} (in columns 1, 2, and 3 in Fig. 6.14(d)) that was found to be better than chance in the LA1.1 overlaps with the reference line in all visualization conditions except for the static condition. The inference on LA2.2 does not suggest any difference between any of the visualization conditions; the "all tasks (LA2.2)" HDI bars in columns 4, 5, and 6 of Fig. 6.14(d) overlap with the reference line. Similarly, the inference results in the LA2.1 level of the individual tasks do not seem to suggest any considerable difference between the visualization conditions. The HDI bars are consistently close and overlap with the reference line across the tasks.

The *third row* of the figure (Fig. 6.14(e) and (f)) concerns the analysis of participants' selections of sufficient interventions for only those participants who got the causal model correct, conducted on the LA1 and LA2 levels.

The LA1.2 level of analysis shows that the "all tasks (LA1.2)" HDI bar in Fig. 6.14(e) is slightly shifted away from the reference line and towards higher values but still it is close to the reference line. In LA1.1 only task t_{16} joins task t_{11} in those that seem to present some stronger evidence of better performance than the chance. In the LA1.3 level of analysis the variance tasks join the atomic ones in those that seem to present some evidence of better performance than chance (the "atomic (LA1.3)" and "variance (LA1.3)" HDI bars in Fig. 6.14(e) are further on the right of the reference line and do not overlap with it).

The LA2 level of analysis confirms partly these observations about the estimation of θ but this time within each visualization condition (Fig. 6.14(f)). In the LA2.2 level of analysis the

“all tasks (LA2.2)” HDI bars in columns 1, 2, and 3 of Fig. 6.14(f) overlap with the reference line in all visualization conditions except for the animated. In LA2.1 the HDI bars of tasks t_{11} and t_{16} (in columns 1, 2, and 3 in Fig. 6.14(d)) that were found to be better than chance in the LA1.1 overlap with the reference line in all visualization conditions except for the static condition. The inference in LA2.2 does not suggest any difference between any of the visualization conditions; the “all tasks (LA2.2)” HDI bars in columns 4, 5, and 6 of Fig. 6.14(f) overlap with the reference line. Similarly, the inference results in the LA2.1 level of the individual tasks do not seem to suggest any considerable difference between the visualization conditions. The HDI bars are consistently close and overlap with the reference line across the tasks.

A final observation is that although there is strong evidence that participants were able to identify the true model in task t_{14} (the “ t_{14} (LA1.1)” HDI bar in Fig. 6.14(a) is further on the right of the reference line and do not overlap with it), their performance in identifying the sufficient interventions in this task was very poor in both the analysis considering all participants and only those that found the correct model (the “ t_{14} (LA1.1)” HDI bars in Fig. 6.14(c) and (e) are further on the left of the reference line and do not overlap with it).

Summary of Findings Overall participants’ performance in identifying the causal model through the exploration of multiple interventions in TT2 tasks was better than chance. This is more certain in the SG than in IG and AG. Participants’ performance was better than chance when the type of the interventions in the TT2 tasks was atomic or shift. The exploration of simulated data from atomic interventions seemed to have helped participants to identify common cause and mediator causal models.

The analysis of participants’ selections of sufficient interventions suggests that the task of identifying interventions that provide sufficient information for identifying the causal model of the presented data was a hard task for participants independently of the visualization condition. The analysis of participants’ selections of sufficient interventions for only those participants who got the causal model correct does not seem to change the picture drawn by the similar analysis conducted based on all participants.

The poor performance of participants in identifying sufficient interventions seems to suggest that they did not rely on the identification of sufficient interventions at a great extent for identifying the causal model. An assumption that could be made is that participants combined information from the different interventions to reach a conclusion as to which the true causal model is. This assumption about the existence of such a strategy is enhanced by the case of task t_{14} .

6.5.4.5 Analysis of Response Times and Confidence

Analysis Model The same Bayesian probabilistic model is used for the analysis of the response times and confidence adjusting only the values of the fixed-value parameters. Two such

models are created to account for two main levels of analysis, LA1 and LA2. Fig. 6.15(a) and (b) present the Kruschke-style diagrams of the Bayesian probabilistic models used for the analysis of the response time and confidence on LA1 and LA2 levels, respectively. A normal likelihood has been used in both cases. The simplifying assumption that the ordinal values of confidence could be treated as if they lay on a common continuous scale was made.

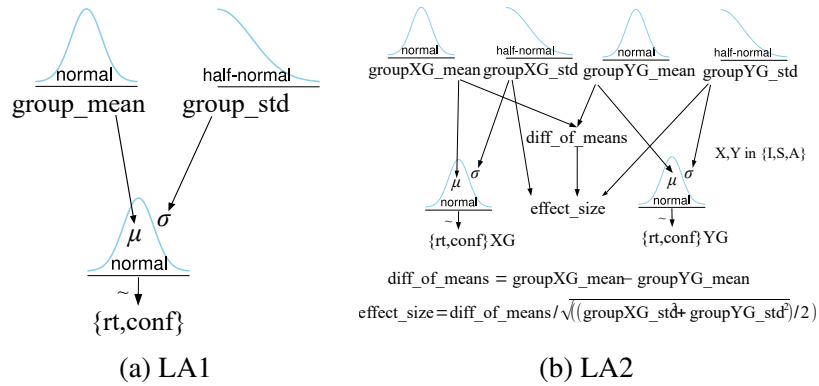


Figure 6.15: Kruschke-style diagrams of the analysis models for participants’ response time and confidence in the user study’s tasks and analysis on the (a) LA1 level for all participants independently of visualization condition, and (b) LA2 level for participants in each visualization condition.

The posterior distribution of *effect size* (*Cohen’s d*) is estimated for the comparison of the three groups (IG, SG, and AG) in terms of the response time to normalise for the varying duration (and thus typical variances) of the tasks. The posterior *mean confidence level* is estimated for each group as confidence takes ordinal values and there was no need to normalise. The difference of the mean confidence posterior distribution of each group for every task are estimated to compare the groups.

Results Fig. 6.16 presents the inference results (posterior distributions) in the form of forest plots for the analysis of participants’ response time (Fig. 6.16(a) and (b)) and confidence (Fig. 6.16(c) and (d)). No reference lines were drawn in the forest plots presenting the posterior distributions of the mean response time and confidence as there was no reference value for these two measures.

The analysis of participants’ response times and confidence in the LA1 level shows strong evidence that participants were faster and more confident in TT1 tasks than TT2 tasks (the “all tasks TT1 (LA1.2)” HDI bars are further apart from the “all tasks TT2 (LA1.2)” HDI bars with no overlap in Fig. 6.16(a) and (c)).

This effect seems to replicate in the forest plots of mean response time and confidence in the LA2 level of analysis for the AG in the case of response times and the IG and SG in the case of confidence (in these cases the “all tasks TT1 (LA2.2)” HDI bars are further apart from the “all tasks TT2 (LA2.2)” HDI bars with no overlap in Fig. 6.16(b) and (d)). The posterior

distributions of the mean response time in TT1 and TT2 tasks overlap well with each other in SG (see the “all tasks TT1 (LA2.2)” and “all tasks TT2 (LA2.2)” HDI bars in the second column of Fig. 6.16(b)) meaning that participants were able to deal with two types of tasks equally efficient. Nevertheless, their confidence levels in TT2 tasks were fairly lower than in TT1 tasks (see the “all tasks TT1 (LA2.2)” and “all tasks TT2 (LA2.2)” HDI bars in the second column of Fig. 6.16(d)).

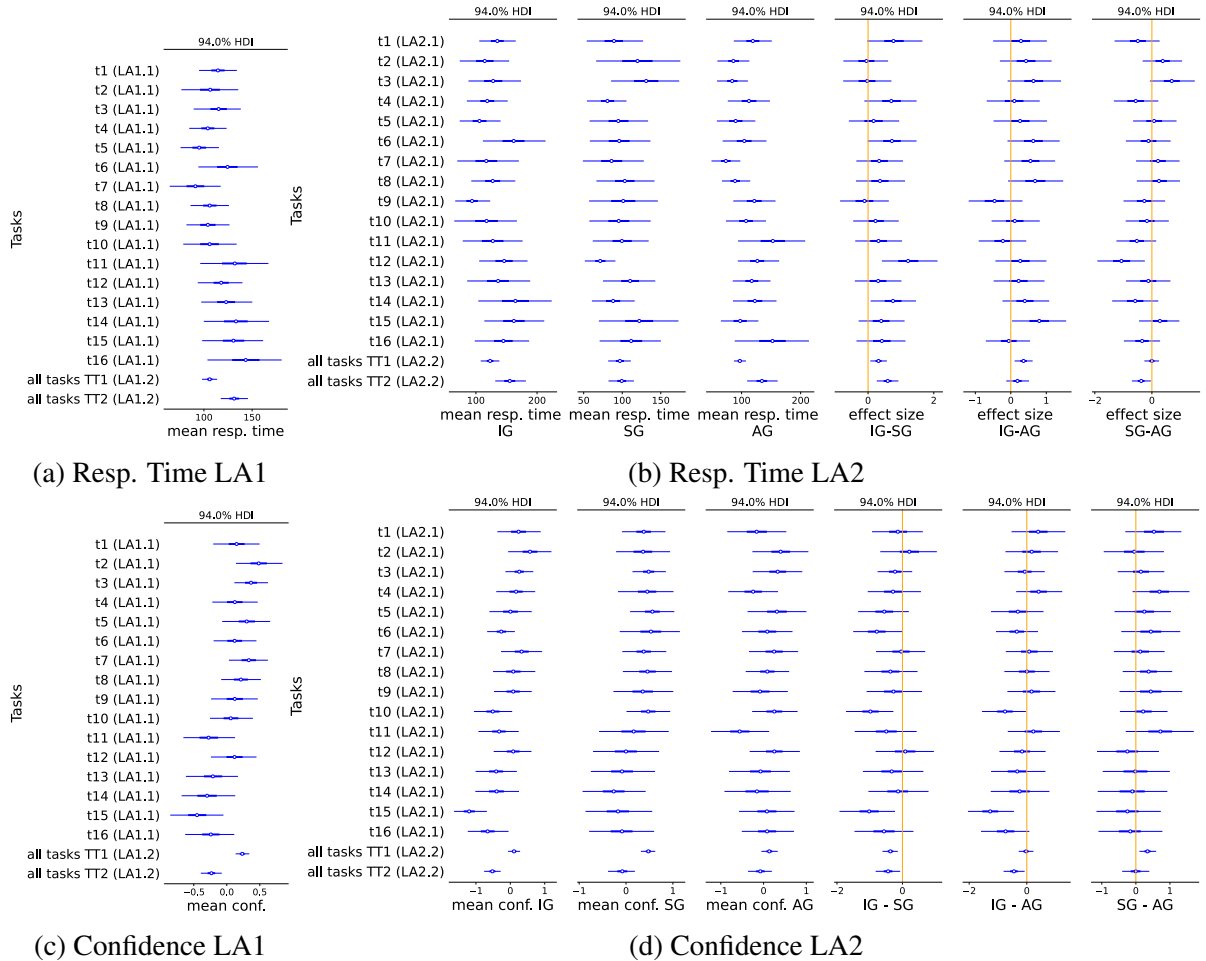


Figure 6.16: Forest plots of the posterior distributions inferred based on the analysis model for the response time and confidence of participants in the user study’s tasks. Forest plot of response time analysis on level (a) LA1 and (b) LA2. Forest plot of confidence analysis on level (a) LA1 and (b) LA2.

Another observation is that the posterior distribution of participants’ mean response time in task t_6 is shift to higher values than the rest of the TT1 tasks (see “ t_{16} (LA1.1)” in Fig. 6.16(a)). That was one of the TT1 tasks that participants did not seem to have performed well (see “ t_{16} (LA1.1)” in Fig. 6.13(a)). A similar effect in the response time of the rest of tasks that indicated poor performance (t_4 and t_8) is not observed.

Participants in the IG needed more time than participants in SG in all tasks (the “all tasks TT1 (LA2.2)” and “all tasks TT2 (LA2.2)” HDI bars in the fourth column of Fig. 6.16(b) are on

the right of the reference line and do not overlap with it), while they were less confident (the “all tasks TT1 (LA2.2)” and “all tasks TT2 (LA2.2)” HDI bars in the fourth column of Fig. 6.16(d) are on the left of the reference line and do not overlap with it).

Summary of Findings Overall the analysis of participants’ response times and confidence showed that participants needed more time and their confidence levels were lower in TT2 tasks than in TT1 tasks. Also participants in IG were less efficient and confident in their responses than participants in SG.

6.5.5 Other Analyses

This subsection presents other data analyses of the collected data from the user study and discusses the findings. The purpose of these analyses are to investigate aspects ASP4-8. Each one of the following paragraphs investigate one of these aspects. The data and the Python code of these data analyses can be found in the github repository in [Taka, 2023b].

6.5.5.1 Did participants’ statistical or causal inference literacy play a role in their performance?

Participants’ statistical or causal inference literacy could have played a role in their performance (ASP4). Question D3 and D4 in the demographic questions’ part of the user study captured the level of participants’ former statistical training and their awareness level of the difference between correlation and causation. Participants’ responses to these questions are plotted along with their overall performance in the TT1 tasks. The overall performance of a participant in TT1 tasks is calculated as the number of TT1 tasks that the participants got correct following the SCO approach.

Fig. 6.17 presents the strip and box plots of participants’ performances plotted against their level of statistical (D3 question) and causal inference (D4) knowledge based on the recorded responses to the demographic questions.

Most of the participants stated that they had either a school or some university training in statistics. The performances of these participants vary from very low values (2 or 3 correct TT1 tasks) to high values (9 correct TT1 tasks) and are distributed quite evenly along the axis between these values. Similarly, most of the participants stated that they could state the difference between correlation and causation quite confidently (definitely yes or probably yes). The performances of these participants vary from very low values (1 or 2 correct TT1 tasks) to high values (9 correct TT1 tasks) and are distributed quite evenly along the axis between these values.

Based on these observations it does not seem that participants’ statistical and causal inference background plays a role in their ability in reasoning causally in the TT1 tasks.

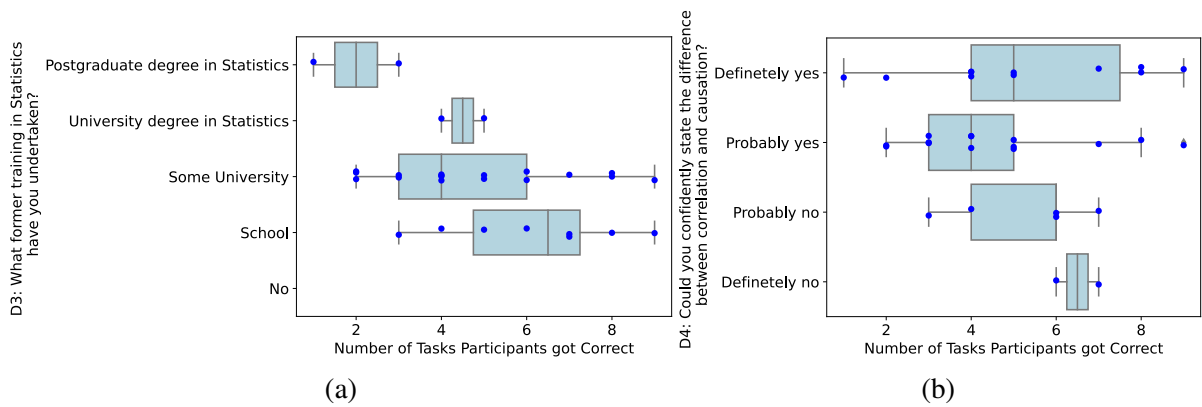


Figure 6.17: Strip and box plots of participants' performance measured as the number of TT1 tasks that each participant got correct plotted against their statistical (D3 question) and causal inference (D4 question) level of knowledge as these were recorded in the demographic questions. Blue dot markers are used in the strip plots to represent the observations (each dot represents a participant in the user study).

6.5.5.2 Did the visual exploration strategy followed by participants play a role in their performance?

There are various ways that the information presented in the scatter plot matrix (i.e., the simulated data before and after the intervention) could be read and perceived by the participants (ASP5). The question is how participants exploited this information and whether the visual exploration strategy that they used to do this played any role in their performance.

In the user study part with the user experience questions participants were asked in how many tasks they looked at the scatter plots (UE2 question) and at the KDE plots (UE3 question). The intention was to use the data from these questions to verify if participants relied mainly on one of them to respond and if that might have affected their performance. Fig. 6.18 presents a graph showing the responses of each participant in these two questions along with their performance. Participants' performance is calculated as the number of TT1 tasks that the participants got correct following the SCO approach.

Observing Fig. 6.18 one could say that there were roughly three different strategies followed for reading the scatter plot matrix. In the central part of the figure participants who looked at both the scatter and KDE plots in all tasks (the circles and squares overlap for these participants) are shown. On the right there are the participants who relied more on the KDE plots (squares are higher than circles for these participants). On the left there are the participants who relied more on the scatter plots than the KDE plots (circles are higher than squares for these participants).

Participants seem to be almost equally split among these three categories of visual exploration strategy. Participants who relied more on the KDE than the scatter plots seem to have scored higher than the rest of participants (the shapes are darker for these participants).

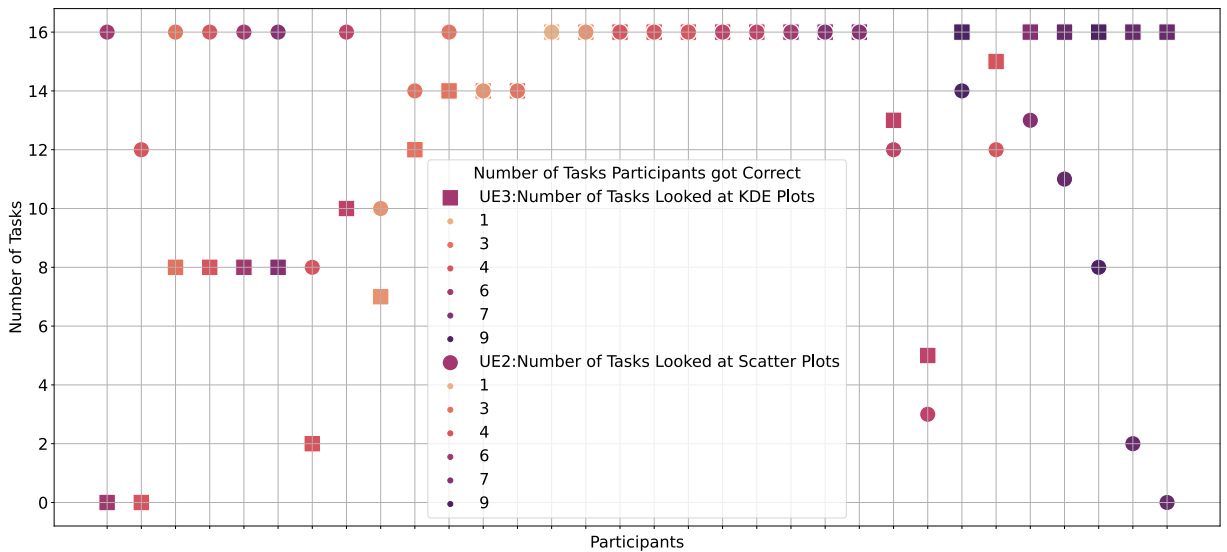


Figure 6.18: Participants responses in UE2 question (*In how many tasks in the study do you estimate that you looked at the scatter plots (plots presenting data with dot markers) in the scatter plot matrix?*) represented by circles and UE3 question (*In how many tasks in the study do you estimate that you looked at the KDE plots (plots presenting the variables’ distribution) on the diagonal of the scatter plot matrix?*) represented by squares. The darker a shape is, the higher the performance of the participant is.

6.5.5.3 Did the visualization condition favor a specific visual exploration strategy?

Participants seem to have followed different strategies in exploring the presented data in the scatter plot matrices. The question is if this was a personal choice or the strategy followed could have been favored by the visualization condition (ASP6). Fig. 6.19 presents a similar graph to Fig. 6.18 showing the responses of each participant in the UE2 and UE3 user experience questions. This time the color of the shapes represents the visualization condition that each participant was assigned to.

Based on Fig. 6.19 participants that followed a specific exploration strategy seem to be equally distributed to visualization conditions. Thus, the visual exploration strategy seems to be driven more by a personal choice of the participant than the visualization condition.

6.5.5.4 Are participants’ responses valid?

A validity check of participants responses is conducted based on their responses to the UE1 user experience question: *In how many tasks in the study do you estimate that you looked at the directed graphs (DAGs) of the causal models?* (ASP7). If many participants report small numbers of tasks in this question, it would mean that they responded uninformatively and most possibly at random. The questions in the user study’s tasks could not be answered without looking at the causal diagrams, at all.

Fig. 6.20 presents a graph showing the responses of each participant in the UE1 question along with their visualization condition. Participants’ performance is calculated as the number

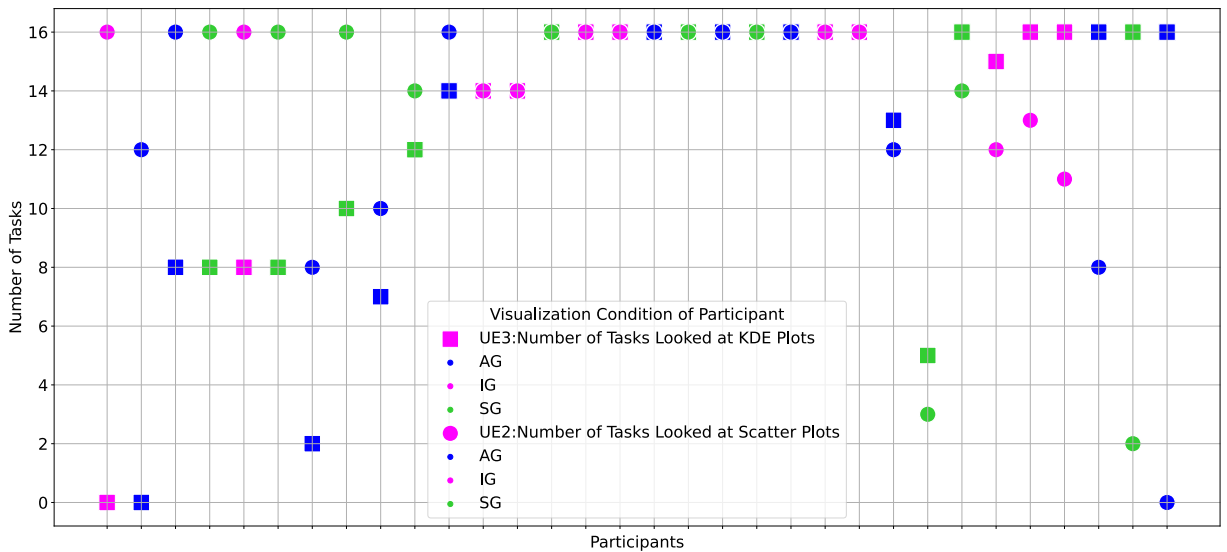


Figure 6.19: Participants responses in UE2 question (*In how many tasks in the study do you estimate that you looked at the scatter plots (plots presenting data with dot markers) in the scatter plot matrix?*) represented by circles and UE3 question (*In how many tasks in the study do you estimate that you looked at the KDE plots (plots presenting the variables' distribution) on the diagonal of the scatter plot matrix?*) represented by squares. The color of the shape represents the visualization condition to which each participant was assigned.

of TT1 tasks that the participants got correct following the SCO approach.

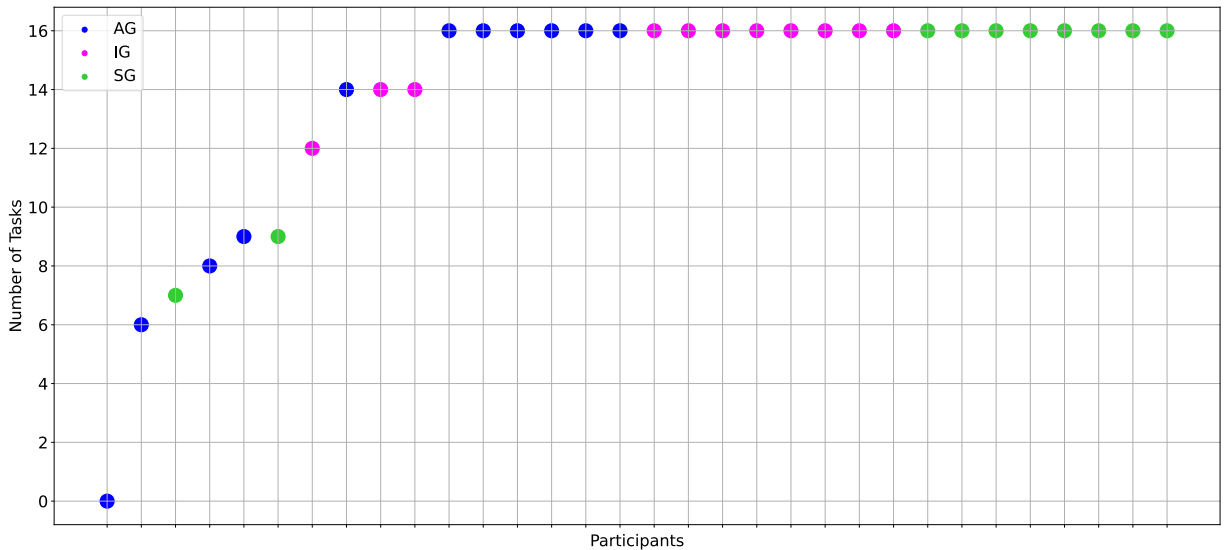


Figure 6.20: Participants responses in UE1 question (*In how many tasks in the study do you estimate that you looked at the directed graphs (DAGs) of the causal models?*) represented by circles. The color of the circles represents the visualization condition of the participant.

Based on Fig. 6.20 very few participants reported that they looked at the causal diagrams in less than 8 tasks. These participants were not excluded from the analysis because we could not be sure in which tasks they might have answered randomly and in which they answered based on some thinking. The first participant on the left of the figure reported that they did not look at

the causal diagrams in any of the tasks. Nevertheless, this participant was not excluded because in the UE2 question in Fig. 6.18 they reported that they looked at the scatter plots in 12/16 of the tasks. Based on this information it is not certain how engaged they were and to what extent they exploited the provided information to answer the questions.

Regarding the distribution across conditions, the responses in the AG group seem to have more noise than these in the IG and SG. There are more participants in AG who stated that they did not look at the causal diagrams in all of the tasks.

6.5.5.5 How did participants judge the design of the visualization?

In the part of the user experience questions participants were asked to state how they found the design of the visualization for the simulated data in UE4 question (ASP8). 27 out of 32 participants left a comment in the provided text box. Table 6.3 presents the responses of participants in the first column and the condition they were assigned to in the second. The comments were tagged with one of three codes to characterize the impression of participants about the visualization: *positive*, *negative*, *ambivalent*.

From the 27 participants who left a comment 19 reported a positive impression, 3 reported a negative impression, and the rest 5 were ambivalent. In IG 6 comments were positive, 3 negative, and 1 ambivalent. In AG 7 comments were positive and 2 ambivalent. In SG 6 comments were positive and 2 ambivalent.

Considering the content of the comments, 17/27 of the comments mentioned that the visualization was informative. The negative cues in the negative or ambivalent comments regarding the visualization were the following: “distractive” (2/8), “annoying” (1/8), or “complex” (1/8). A few of the negative cues in participants’ comments seem to refer to the design of the study rather than of the visualization: “annoying, lengthy” (1/8), “descriptive” (1/8), “repetitive” (1/8).

6.5.6 Summary of Findings

The analysis of the collected data from the user study led to various interesting results and findings. The analysis investigated all the different aspects mentioned in Section 6.5.1:

- ASP1 how people performed in causal reasoning tasks when simulated probabilistic data before and after an intervention was presented to participants in the form of a scatter plot matrix and in three different visualization modes: static, interactive, and animated. RQ1 and RQ2 are part of this investigation;
- ASP2 whether there were any specific settings (e.g., intervention types, causal models) in which participants were able to perform better under these circumstances;

Table 6.3: Participants' responses to the UE4 (*How did you find the design of the visualization for the simulated data? Was it informative, annoying, distracting?*) shown in the first column, the condition each was assigned to shown in the second column, and the code with which each comment was tagged shown in the third column.

Comment	Condition	Code positive/negative/ambivalent
It gave information in a simple, clear and organized way.	IG	positive
Informative	IG	positive
Good	IG	positive
Very Informative	IG	positive
I found it distracting	IG	negative
Very good and easy to understand the changes.	IG	positive
Informative, the visualization was actually very good	IG	positive
annoying, lengthy	IG	negative
Could have been spaced out a little more.	IG	negative
Informative but too descriptive at times	IG	ambivalent
The visualisations were informative and fairly easy to follow	AG	positive
found it useful; I like that it was animated and you could see how changing one variable "moves" the others	AG	positive
the colours were a good contrast - the scatter graph i found easier to view	AG	positive
Informative but complex	AG	ambivalent
Informative	AG	positive
Informative	AG	positive
Informative	AG	positive
The design of visualization for simulated data was informative but repetitive after a point of time. Overall it gave good information	AG	ambivalent
Clear enough	AG	positive
A bit distracting but on the whole comprehensive	SG	ambivalent
Informative	SG	positive
Informative	SG	positive
Informative	SG	positive
Informative	SG	positive
Quite informative	SG	positive
Somewhat informative	SG	positive
it was what it needed to be	SG	ambivalent

ASP3 what strategies participants used to identify the causal model when they explored the simulated data from multiple interventions;

ASP4 whether the statistical or causal inference level of knowledge played a role in participants' performance;

ASP5 whether there was any strategy followed in how the information presented in the scatter plot matrix was used to reason causally;

ASP6 whether the visualization mode played any role in the adoption of any such strategy;

ASP7 to what extent participants' responses were valid;

ASP8 how participants judged the design of the visualization.

For ASP1 participants performed well overall in tasks of identifying the causal model given a specific intervention (TT1 tasks) or given multiple interventions of the same type (TT2 tasks). In the latter a better performance is more certain when the simulated interventional data is presented in a static scatter plot matrix, while in the first it is independent of how the data is presented. The tasks of identifying sufficient interventions for the determination of the causal model were hard for all participants independently of the visualization condition. Participants needed more time and their confidence levels were lower in the tasks they were required to explore multiple interventions (TT2 tasks). Participants in IG were less efficient and confident in their responses than participants in SG independently of whether they had to explore multiple interventions.

For ASP2 the performance of participants in single-intervention tasks (TT1 tasks) in which the intervention was not informative enough to identify the causal model (tasks t_3 and t_{10}) is estimated to be better than chance when the MCO instead of the SCO approach is used for measuring accuracy. Participants seemed to be able to identify the causal model in the case of atomic interventions independently of whether they had to explore multiple interventions. The exploration of simulated data from multiple atomic interventions seemed to have helped participants to identify common cause and mediator causal models. Finally, in the case of distinguishing the collider from the common cause or mediator causal model through applying an intervention on tiredness the model whose simulated data is presented to participants seems to play a role in their ability to identify the true causal model.

For ASP3 participants who were able to explore multiple interventions and got the causal model correct were not able to determine the sufficient interventions. The poor performance of participants in identifying sufficient interventions seems to suggest that they did not rely on the identification of sufficient interventions at a great extent for identifying the causal model. An assumption that could be made is that participants combined information from the different interventions to reach a conclusion as to which the true causal model is. This assumption about the existence of such a strategy is enhanced by the case of task t_{14} .

For ASP4 it does not seem that participants' statistical and causal inference background played a role in their ability in identifying the causal model in single-intervention tasks.

For ASP5 three different strategies followed by participants for reading the scatter plot matrix; looking at both the scatter and KDE plots, looking more on the KDE than the scatter plots, and looking more on the scatter than the KDE plots. Participants seem to be almost equally split

among these three categories of visual exploration strategy. Participants who relied more on the KDE than the scatter plots seem to have scored higher than the rest of participants.

For ASP6 participants who followed a specific exploration strategy seemed to be equally distributed to visualization conditions. Thus, the visual exploration strategy seems to be more a personal choice than driven by the visualization mode.

For ASP7 few participants reported that they did not look at the causal diagrams in all tasks, and very few of them in less than 8 tasks. The selections of participants in the user study's tasks seem to have been informative according to their responses in the user experience questions.

For ASP8 the majority of participants reported a positive impression about `vicausi` and that was consistent in all visualization conditions. Most of the participants mentioned that they found the provided visualization informative.

6.5.7 Limitations of the User Study

The number of questions had to be limited to ensure the completion of study by participants in roughly an hour. The user study was designed to include the three main three-variable causal primitives (common cause, mediator, and collider), three different types of interventions, interventions on any of the three variables, and two different types of tasks: one for identifying the causal model given a specific intervention and two possible causal models, and another for identifying the causal model by exploring multiple interventions of a single type and given three possible causal models. These were enough to produce a multitude of possible settings to be explored. Only a subset of these possible combinations was finally included in the user study. This limited the extent of insight into the causal contexts in which people might perform better or worse (ASP2).

The design of this user study provided some insight into how participants identified the causal model when they explored multiple interventions (ASP3), but still the exact strategy is not known. How did they combine the information from multiple interventions to reason about the causal relations found in the simulated data? Other design protocols (e.g., think-aloud protocol) could help with this.

A single causal reasoning problem (the insomnia-anxiety-tiredness problem) was used both in the training and all tasks of the user study. It was not investigated to what extent the visual causal reasoning skills and strategies participants used in the context of this problem would be transferable to the context of another problem.

The tasks in the user study were cognitively demanding, especially the TT2 ones (see the increased variability in the inferred mean response time in Fig. 6.16(a) and 3 first columns of Fig. 6.16(b) for tasks t_{11} - t_{16}). A few of the participants' comments in their feedback in question UE4 (see Table 6.3) could also imply the complexity of the tasks and a possible tiredness of the participants; some of these comments mentioned the phrases "complex", "annoying, lengthy", "repetitive", "descriptive". Participants' possible tiredness could constitute a form of

bias especially for the second part of the tasks, the TT2 ones, that came last in the study and might be the reason why participants failed to identify the sufficient interventions. To exclude the factor of tiredness from biasing the findings of the user study, the design of the user study could have further simplified the participant's session, possibly by investigating each set of tasks in separate sessions or user studies.

6.6 Discussion

6.6.1 Contributions

A common task in many fields that are concerned with defining the cause of observed effects (e.g., medicine, epidemiology, biology, marketing) is the design and conduction of interventional experiments. By applying interventions on the variables of a system of interest the underlying causal relations of the variables can be learned reliably. Interventional experiments are not always possible; they might be unethical or too expensive to conduct. On the other hand, when they are possible, the decisions on which variables to intervene, especially when there are many possibilities, might not be very informative. Is any prior knowledge or expectation (hypotheses) of the causal relations among variables taken into consideration? Are researchers' hypotheses about the expected outcomes of an interventional experiment tested before the conduction of an experiment to verify whether this is the right experiment? Is uncertainty accounted for?

Simulating interventions could provide valuable information for designing interventional experiments or in cases when these experiments are not possible. For example, a researcher or analyst might want to simulate interventions to test (validate) their expectations about a planned experiment and possibly decide whether this is the right experiment to conduct. Or there might be multiple possible interventional experiments and the analyst needs to decide which ones would be more informative and helpful in identifying the causal structure of the data.

Producing such simulations involves specialized knowledge (e.g., in statistics or causal inference) or skills (e.g., in programming) that researchers or analysts often do not possess. There are very few existing tools that support the simulation and exploration of interventions [Xie et al., 2020] or the incorporation of assumptions and prior knowledge [Wang and Mueller, 2015; Wang and Mueller, 2017]. There is need for tools that are specifically designed to simulate interventional experiments by accounting for uncertainty and researchers' prior knowledge and causal hypotheses. There is also need for tools that would allow the exploration of the outcomes of these simulated experiments with the aim to support researchers in causal reasoning and designing of such experiments. These tools could also provide insight into the outcomes of interventional experiments that are not realizable.

This work introduced a simulation pipeline for exploiting existing simulation tools of probabilistic models and Bayesian inference (e.g., PPLs and sample-based inference using MCMC)

to produce uncertainty-aware simulations of interventions. These interventions are applied on hypothesized causal models, which possibly describe the data generating mechanisms of data variables of interest according to an analyst. A concrete implementation of this pipeline is provided in Python [Taka, 2023c]. This implementation enables the user to input their hypothesized causal models into a Python method in a simple way. This method automatically retrieves a batch of simulated data before and after a variety of different interventions applied on the hypothesized causal models along with metadata, all combined into a standardized output structure. In this way, users can have access to simulated data of interventions through the call of a single method.

The output of this pipeline is quite complex consisting of multitudes of samples from multidimensional distributions. To facilitate the exploration of such outputs, an automatic visual exploration tool was designed, *vicausi* [Taka, 2023d]. This is a browser-based application that takes as an input the pre-computed simulated data (in the standardized form produced by the simulation pipeline) and some preferred configurations. It renders a page on the browser showing the simulated data of a chosen hypothesized causal model in a scatter plot matrix, the causal diagrams of all the hypothesized causal models, and controls that enable the exploration of various interventions.

We could not rely on existing work too much to inform the design of this tool because this is limited. Existing visualization tools supporting causal reasoning were not primarily designed to present probabilistic data of interventions with the aim to support causal reasoning and design of interventional experiments. Most importantly there is very little work on understanding how people exploit visual information to reason causally [Lagnado and Sloman, 2004; Yen et al., 2019].

For this reason, two target tasks were determined and the design of this tool was informed by a set of objectives set to facilitate these two target tasks. These target tasks were the identification of the causal model of the presented data given a specific intervention, and the determination of interventions that provide sufficient information to identify the causal model of the presented data. Three different ways of showing the simulated data after an intervention were included in the design of the tool. The first way was statically showing the simulated data after the intervention as slices of continuous ranges of the intervened variable. The second way was to use interaction for querying the slice of simulated data to be shown. The third way was to show an animation of all sequential slices that cover the range of the intervened variable.

A user study was designed and conducted to evaluate the design of *vicausi* on the basis of the two target causal reasoning tasks. The effect of interactive and animated presentations of the simulated interventional data in *vicausi* on participants' ability to reason causally was investigated. The visual exploration strategies used by participants to exploit the information provided by the scatter plot matrix were analysed. Participants' user experience was also captured by this user study.

The findings from the analysis of the collected data (the code of the analysis and the data can be found in [Taka, 2023b]) suggest that participants using `vicausi` were able to identify the causal model of the presented data either given a single intervention or by exploring various interventions. On the other hand, their performance in identifying sufficient interventions was poor. It was found that participants did not rely on the sufficient interventions to identify the causal model in the case of multi-interventional tasks. They might have relied more on combining information from multiple interventions to draw their conclusions.

In terms of `vicausi`'s design, no strong effect of any of the visualization modes used for presenting the interventional data was found. It was found that there were three different visual exploration strategies of the information in the scatter plot matrices which participants followed; roughly 1/3 of them relied on both the scatter and KDE plots, another 1/3 of them relied more on the scatter plots, and the last 1/3 of them relied more on the KDE plots. Those who followed the last strategy had a better performance in identifying the causal model given a specific intervention. Finally, most participants judged the design of the visualization positively (this was consistent in all visualization conditions) and mentioned that they found it informative.

6.6.2 Future Work

This research presents some original ideas in the field of supporting causal reasoning by using simulation and visualization, creates a working framework for demonstrating their value, and provides research evidence about this value. This research could pave the way for further future research towards this direction.

Probabilistic modeling and Bayesian inference become more and more popular lately because of the emergence of powerful inference tools that rely on simulation engines, the PPLs. PPLs' simulation engines exploit efficient and well-tested implementations of MCMC sampling algorithms and are continuously improving. Interventions on causal structures could be simulated by exploiting these simulation engines to account for uncertainty and the complexity the conduction of such simulations could involve.

Existing implementations of PPLs do not offer great flexibility in intervening on a defined model and amend it to apply interventions. The simulation pipeline for interventions proposed in this work showed that this could be achieved through the manual definition of linear transformations of the observed variables' likelihood when expressing the model in PPL code. In the future, PPLs' designers and developers could include these options in the syntax of the language to make the process of modeling interventions more intuitive and flexible.

Automatic tools for designing and defining in PPL code the probabilistic models into which the causal models are transcribed are also necessary. This would offer more flexibility in the types of variables (continuous or discrete), likelihoods and priors used for the modeling.

From the visualization perspective, it is very little known about how visualization could support the complex cognitive task of causal reasoning. This research provides some context

towards this direction but there are still many questions that are unanswered.

For example, it is not clear why participants in the conducted user study could not identify sufficient interventions, although they were able to identify the causal model of the presented data. Participants seemed to be able to use the information provided by `vicausi` to reason causally and identify the causal model given a concrete intervention. This process involves many cognitive steps of validating or rejecting causal links between the variables of the hypothesized causal models and it seemed that the use of a scatter plot matrix and causal diagrams supported them.

The question is why participants failed to replicate this primary task in the case of multi-interventional tasks for as many as the interventions to identify the sufficient ones. Was it a matter of the visualizations' design, ability of reasoning in complex causal reasoning tasks (is there a threshold in the complexity of the causal reasoning task that people can deal with?), comprehension of the task, or tiredness (these tasks were shown last to participants in the user study)? Future research could investigate more thoroughly how people reason causally in more complex causal reasoning tasks, how the presented visual information could support them in this, and whether the design of the visualization has an impact on this.

Regarding the design of the visualization to support such tasks, there are various extensions of the tool that could possibly offer more flexibility in the exploration of causal assumptions. For example, `vicausi` shows the simulated data from a single causal model in the scatter plot matrix. This was decided to emulate the condition of possessing and exploring actual interventional data. An alternative design could include the comparative presentation of the simulated data from multiple hypothesized causal models (e.g., the parallel presentation of a separate scatter plot matrix showing the simulated data from each one of the hypothesized causal models) or datasets in a single instance of the visualization. Such designs could possibly encourage users to explore multiple causal assumptions or datasets at once and thus, become implicitly aware of the fact that any conclusions drawn are at stake as they are based on assumptions, which might not reflect the reality. Conducting interviews with real users of such visualizations like researchers who decide, design, and conduct interventional experiments would inform the design of such a tool and provide useful feedback as to which views and functionalities might be more helpful in their work routine.

An alternative more compact design of the tool would be a presentation of the causal diagrams of the hypothesized causal models, which would incorporate the KDE plots into the nodes of the causal diagram. Such a design would remind IPME. Users would be able to explicitly compare the simulated data before and after an intervention for each variable across the different hypothesized models and infer which interventions could be more informative. Also as discussed in Section 6.5.5.2, participants who focused more on KDE plots and less on the scatter plots seemed to have performed better than those who relied more on the scatter plots. Further experimentation is required to investigate whether such alternative designs would be

more effective in supporting complex causal reasoning tasks.

6.6.3 Conclusions

Simulating interventions could provide valuable information for designing interventional experiments or in cases when these experiments are not possible. Producing such simulations involves specialized knowledge or skills that analysts often do not possess. There is need for automatic tools for simulation and visual exploration of interventions to support causal reasoning and designing of interventional experiments. This work focused on proposing a pipeline to generate simulated probabilistic data from interventions applied on causal structures that are expressed in PPLs using probabilistic modeling and Bayesian inference. The implementation of an automatic visualization tool for visualizing the simulated probabilistic data generated by this pipeline was presented. An evaluation user study of the proposed tool was conducted. The evaluation focused on how people reason causally and make decisions on interventional experiments when the uncertainty in the simulated data of interventions was presented using static, animated, or interactive visualizations.

The findings suggested that participants were able to identify the causal model of the presented data either given a single intervention or by exploring various interventions. Their performance in identifying sufficient interventions was poor. Participants did not rely on the sufficient interventions to identify the causal model in the case of multi-interventional tasks. They might have relied more on combining information from multiple interventions to draw their conclusions. There were three different visual exploration strategies of the information in the scatter plot matrices which participants followed; roughly 1/3 of them relied on both the scatter and KDE plots, another 1/3 of them relied more on the scatter plots, and the last 1/3 of them relied more on the KDE plots. Those who followed the last strategy had a better performance in identifying the causal model given a specific intervention. Most participants judged the design of the visualization positively (this was consistent in all visualization conditions) and mentioned that they found it informative. These initial findings evoke the need for more research to understand how users can benefit from visual representations of simulated interventional data and could pave the way for future investigation into the role of visual means to support more complex causal reasoning tasks like the design of interventional experiments.

Chapter 7

Conclusions

7.1 Summary

This chapter summarizes the purpose of the research presented in this thesis, its contributions and the conclusions drawn from it. The future direction of this work is discussed.

7.2 Purpose of Research

The aim of this research was the proposal and creation of novel and automated tools for visualizing Bayesian probabilistic models and their outputs using interaction and animation. These tools would make Bayesian analysis and its outputs more intuitive and interpretable for a general audience. Bayesian models' structure and inference results are complex and difficult to grasp and interpret without specialized knowledge. Thus, the proposed tools were intended to support users in understanding and validating the structure of probabilistic models, and exploring and understanding their outputs through visual representations.

Uncertainty was a main aspect of the proposed tools. It is a cornerstone of probabilistic models and is accounted for in them. When included in the representation of a probabilistic model and its outputs, uncertainty can offer rich valuable information about the model and its variables. This could convey, for example, the amount of risk involved in a decision. Uncertainty could also reveal the causal- and non-causal relations of variables in a probabilistic model. For example, representing the pairwise distribution of variables in a model could convey correlations between variables, which, within the context of a causal model, could be interpreted as causal relations or not.

The evaluation of the proposed visualizations was also part of the purpose of this research. The aim of the evaluation was two-fold. First, to investigate the effectiveness of the proposed tools in tasks where participants would have to identify the (causal and non-causal) relations of variables in probabilistic models. Second, to gain insight into how participants exploit the visual information to draw conclusions about variables' relations.

7.3 Summary of Findings

The research presented in this thesis consisted of three separate parts which all focused on the visualization of probabilistic models and their outputs but from a different perspective.

The *first part* (Chapter 4) presented a generic pipeline to transform a probabilistic programming model and associated sample-based MCMC inference results into a standardized format. This standardized output could then be automatically translated into an *interactive probabilistic models explorer (IPME)*. IPME is the first novel representation of Bayesian probabilistic models proposed in this research. It fuses the structural and distributional display of a probabilistic model in an interactive graphical representation. An initial tool was developed to render these representations from standard PPL definitions and offered in the form of a Python package [Taka, 2020a].

IPME aimed at a threefold representation effort; the representation of a collapsible tree-like structure of the model’s variables and parameters to reveal internal levels of statistical or mathematical dependencies among them; the representation of each inferred parameter as a node that presents graphically the marginal prior or posterior distribution of the inferred parameter’s MCMC samples; the representation of the observed variables through prior or posterior predictive distributions of the model’s predictive samples. Appropriate uncertainty visualization techniques were used to graphically represent the marginal distributions. Interactive conditioning was used to enable a form of “sensitivity analysis” of the variables by querying and presenting the conditional marginal distributions.

IPME was designed with the intention to enhance informativeness, transparency and explainability and ultimately, the potential of increasing trust in models. It provides at-a-glance communication of a probabilistic program’s structure and uncertainty of latent parameters, and allows interactive exploration of the multi-dimensional prior or posterior MCMC sample space. It makes the sample-based (MCMC) inference results easily explorable.

The *second part* (Chapter 5) explored the *interactive pair plot (IPP)* (the code can be found in [Taka, 2020b]) for visualizing a model’s distribution; that is a scatter plot matrix of a probabilistic model allowing interactive conditioning on the model’s variables. IPP was built on IPME’s framework inheriting its design elements and most importantly, its interactive conditioning’s mechanism and design. Interactive conditioning worked as a brushing-and-linking effect in IPP highlighting queried information and presenting the conditional marginal distributions of variables.

A user study investigated whether adding interactive conditioning to classical scatter plot matrices helped users better understand the relations of variables in a probabilistic model and ultimately, its structure. It was also investigated whether there were levels of structural detail and model designs for which interactive conditioning was beneficial. The analysis of the collected data (the code of analysis and the data are provided in [Taka et al., 2022]) showed that using interactive conditioning was beneficial in cases of sophisticated model designs in comparison to

a static scatter plot matrix. The difference in response time between the interaction and static group of participants became less important in higher levels of structural detail. Participants using interactive conditioning were more confident about their responses overall with the effect being stronger in tasks of lower level of structural detail.

The *third part* (Chapter 6) proposed a pipeline to generate simulated probabilistic data from interventions applied on causal structures that are expressed in PPLs. A concrete implementation of this pipeline in Python using PyMC3 is provided in [Taka, 2023c]. This pipeline outputs the simulated data before and after interventions in a standardized format which is used as an input to the *Visualizer of Causal Assumptions and Uncertainty-Aware Simulations of Interventions* (`vicausi`).

`vicausi` is a proposed visualization tool for visualizing and exploring simulated probabilistic data before and after interventions applied on causal structures. The main visualization elements used in this tool were the scatter plot matrix and the causal diagrams; the first is used for presenting the simulated interventional data of the variables in pairs, and the second for presenting hypothesized causal models that could have generated the presented data according to an analyst or researcher. The tool is provided in the form of a Python module in [Taka, 2023d].

Two target tasks were determined and the design of this tool was informed by a set of objectives set to facilitate these two target tasks. These target tasks were the identification of the causal model of the presented data given a specific intervention, and the determination of interventions that provide sufficient information to identify the causal model of the presented data. Three different ways of showing the simulated data after an intervention were included in the design of the tool. In the first the simulated data after the intervention were shown as slices of continuous ranges of the intervened variable statically. IN the second interaction was used to query the slice of simulated data to be shown. In the third way an animation of all sequential slices that cover the range of the intervened variable was shown.

A user study was designed and conducted to evaluate the design of `vicausi` on the basis of the two target causal reasoning tasks. The effect of interactive and animated presentations of the simulated interventional data in `vicausi` on participants' ability to reason causally was investigated. The visual exploration strategies used by participants to exploit the information provided by the scatter plot matrix were analysed. Participants' experience with using the tool was also captured by this user study.

The findings from the analysis of the collected data (the code of the analysis and the data can be found in [Taka, 2023b]) suggest that participants using `vicausi` were able to identify the causal model of the presented data either given a single intervention or by exploring various interventions. On the other hand, their performance in identifying sufficient interventions was poor. It was found that participants did not rely on the sufficient interventions to identify the causal model in the case of multi-interventional tasks. They might have relied more on combining information from multiple interventions to draw their conclusions.

In terms of `vicausi`'s design, no strong effect of any of the visualization modes used for presenting the interventional data was found. It was found that there were mainly three different visual exploration strategies of the information in the scatter plot matrices which participants followed; roughly 1/3 of them relied on both the scatter and KDE plots, another 1/3 of them relied more on the scatter plots, and the last 1/3 of them relied more on the KDE plots. Those who followed the last strategy had a better performance in identifying the causal model given a specific intervention. Finally, most participants judged the design of the visualization positively with many mentioning that they found it informative.

In summary, the research presented in this thesis focused on KDE plots combined with rug plots for the univariate marginal distributions of models' variables and on scatter plots for the pairwise distributions of the variables. Interactive conditioning implemented as a brushing-and-linking effect and other common interactive graphical elements (i.e., drop-down menus, sliders, radio buttons, buttons) were included in the designs of the proposed visualizations. Regarding animation, direct animation without animated transitions was investigated in the last part of the thesis.

7.4 Conclusions & Future Directions

The research presented in this thesis presented visualization tools (`IPME`, `IPP`, and `vicausi`) designed to integrate interaction or animation to make probabilistic models and Bayesian inference more explainable and support tasks like model validation and refinement, decision-making, and causal reasoning. The intention was to communicate the structure of a probabilistic model through visual means in a way to illustrate the "essence" of the models' mathematical definitions intuitively; what variable in a model is linked to what other variable, what the effect of a change in the value of one variable might be on other variables in the model.

The "essence" of the models' mathematical definitions is often difficult to extract from the typical textual or mathematical descriptions of probabilistic models. Complex mathematical transformations or equations, deep hierarchical structures, or other sophisticated formulations might be involved in these definitions, which are difficult to grasp without a strong statistical background. By simply looking at definitions of probabilistic models like probabilistic statements the effects of parameters on each other or on the predictions of the model cannot be easily inferred in cases of such complex models even if users acquire a good statistical background. Chapter 4 and 5 focused on the proposal of ways to communicate this aspect of model's structure visually and interactively. To achieve this, visualizations of the distributional information of the model were employed in combination with a mechanism to specify and report visually conditional queries in the sample space of the model (i.e., the interactive conditioning).

A user study was conducted in Chapter 5 to collect evidence in regard to what extent (level of detail of relations or model designs) visual means like graphs and interactive conditioning of

a probabilistic model's distribution could facilitate people's comprehension of model's structure without conveying details about the mathematical definition of the model. The results seem to suggest that visualizing model's distributional information and interactive conditioning is a way to achieve a comprehensive communication of variables' relations in a probabilistic model.

Up to this point the research in this thesis focused on the communication of variables' relation within the context of a model. Nevertheless, as it was discussed in Chapter 1, a model does not usually describe the exact actual data generating mechanism, but it is used to approximately capture interesting aspects of it. Thus, all efforts to this point concern how users can conceive the structure of these *proxy* data generating mechanisms through visualization. Chapter 6 takes a step further to investigate how visualization could be used to conceive and explore possible *actual* data generating mechanisms described by probabilistic models. Visualization of model's distribution is employed to communicate the results of simulations from interventional experiments to help people realize what the possible implications of certain interventions on a system might be (e.g., what variables will be affected by the intervention and how) under certain assumptions. These explorations could be really useful to make more informed decisions as to which interventional experiments might be more informative to conduct (i.e., help retrieve the causal relations of the variables in the model).

In a user study conducted in Chapter 6, the results seem to suggest that visualizations of probabilistic models' distribution when these reflect causal assumptions were generally helpful for participants to identify variables' causal relations after a simulation of interventional experiments based on specific assumed causal structures of data. This is an interesting finding and aligns with those from the previous user study; visualizations of models' distributional information seems to be able to support people in inferring the relations of variables' relations within the context of probabilistic models. On the other hand, participants in the user study in Chapter 6 did not generally perform well in identifying interventional experiments that could sufficiently reveal the causal relations of the variables based on the proposed visualizations. There is need for further investigation towards this direction as to how visualization could support users in making more informed decisions when they need to design interventional experiments.

New directions are also opened for future research as to how simulation and visualization could be combined and used to support causal reasoning and more complex causal reasoning tasks like the design of interventional experiments. As the computational power of computers improves, and simulation algorithms become more efficient, simulation approaches will become more and more popular as they would offer "inexpensive" alternatives. Then, visualization could play an important role in the presentation and exploration of the simulated data.

There is enormous potential in interactive exploratory tools for probabilistic models that support the elicitation, validation and presentation of Bayesian probabilistic models. As the tools for Bayesian modeling and efficient Bayesian inference improve and become more accessible to a broader audience, there will be increased need for such tools. Users' engagement with the

tools could play an important role in the design of tools to make Bayesian probabilistic models more *explainable*. Interaction and animation could be employed to exploit user's engagement and evaluated for its effectiveness more systematically.

Appendix A

Examples of Models

A.1 Average Minimum Temperature in Scotland

A very simple model that could predict the average minimum temperature in Scotland for the month November could be described as:

$$\mu \sim \text{Normal}(\mu = 2, \sigma = 10) \quad (\text{A.1})$$

$$\sigma \sim \text{Half-Normal}(\mu = 0, \sigma = 10) \quad (\text{A.2})$$

$$y \sim \text{Normal}(\mu = \mu, \sigma = \sigma) \quad (\text{A.3})$$

This model was implemented in PyMC3 using the NUTS [Homan and Gelman, 2014] MCMC sampler for the computation of the posterior. The model definition in PyMC3 is given by the following lines of code:

```
import pymc3 as pm

#DATA: mean min temperature in Scotland for November 1884-2020
t = [1.0, 1.1, 2.5, 0.8, 2.7, 2.9, 1.4, 1.1, 2.2, 0.3, 3.6, 1.7,
2.1, 3.5, 1.5, 4.6, 2.3, 1.2, 3.5, 1.8, 1.3, 1.2, 3.7, 1.4, 2.9,
-0.3, -1.1, 1.1, 1.6, 3.2, 1.9, -1.3, 3.2, 3.1, 0.8, -1.6, 3.6,
0.7, 2.1, -0.5, 3.5, -0.0, 1.1, 1.8, 2.7, 2.0, 0.6, 3.6, 1.8,
1.7, 1.5, 1.7, 1.3, 1.5, 3.8, 3.0, 1.9, 1.7, 1.0, 2.2, 0.4, 3.4,
2.7, 1.6, 3.0, 2.4, 0.6, 3.4, -0.3, 4.0, 1.6, 3.5, 2.6, 2.8, 2.2,
2.9, 1.7, 1.1, 1.5, 2.2, 2.3, -0.2, 0.5, 1.5, 1.7, -0.8, 1.9,
1.7, 0.9, 0.9, 1.7, 1.8, 1.5, 0.9, 3.1, 1.2, 1.9, 2.5, 2.2, 3.0,
3.3, -1.1, 2.5, 2.6, 1.7, 2.0, 2.0, 2.0, 1.5, 0.4, 5.2, 3.0, 0.2,
4.6, 1.3, 2.9, 1.8, 3.0, 3.8, 3.4, 3.3, 1.5, 3.4, 3.4, 2.0, 2.8,
0.0, 4.8, 1.7, 1.1, 3.8, 3.6, 0.3, 1.5, 3.6, 1.1]
```

```

samples = 1000
chains = 2
tune = 1000
model = pm.Model()
with model:
    #PRIORS
    mu = pm.Normal('mu', mu = 2, sd = 10)
    sigma = pm.HalfNormal('sigma', sd = 10)
    #OBSERVED VAR
    y = pm.Normal('y', mu = mu, sd = sigma, observed = t)
    #INFERENCE
    trace = pm.sample(draws = samples, chains = chains, tune = tune)

```

The data provides the average minimum temperature in Scotland in month November for the years 1884 – 2020 and it was retrieved by [metoffice.gov.uk](https://www.metoffice.gov.uk)¹.

A.2 The Eight Schools Model

The aim of the eight schools model is to predict the effect of coaching programs on the scholastic aptitude test (SAT) for the admission to college in the US. The SAT was designed to reflect the knowledge that was acquired over a long-term effort rather than short-term efforts like short-term coaching before the test. The study conducted by Rubin [1981] recorded the change in SAT scores in eight different schools after providing to students short-term coaching before taking the test.

A hierarchical model for the prediction of the change in the SAT scores of the eight schools is specified in probabilistic statements as following:

$$\mu \sim \text{Normal}(\mu = 0, \sigma = 5) \quad (\text{A.4})$$

$$\tau \sim \text{Half-Cauchy}(x_0 = 0, \gamma = 5) \quad (\text{A.5})$$

$$\theta_i \sim \text{Normal}(\mu = \mu, \sigma = \tau) \quad (\text{A.6})$$

$$y_i \sim \text{Normal}(\mu = \theta_i, \sigma = \sigma_i) \quad (\text{A.7})$$

where $i \in \{1, \dots, 8\}$, y_i are the observed changes in the SAT scores, and σ_i are the standard errors of the observed scores' changes.

The model's definition in PyMC3 is given by the following lines of code. In this example, the

¹<https://www.metoffice.gov.uk/pub/data/weather/uk/climate/datasets/Tmin/date/Scotland.txt>

`arviz_json` module is used to export the PyMC3 model into the standardized structure described in Appendix B. For brevity, the steps to do this are omitted from the presentation of the rest models. The only thing that changes in the application of these steps to the rest of the models is the definition of the indexing dimensions and coordinates. This will be included in the code of the rest models.

```

import pymc3 as pm
import numpy as np
import arviz as az
from arviz_json import get_dag, arviz_to_json

#DATA
J = 8
obs = np.array([28., 8., -3., 7., -1., 1., 18., 12.])
sigma = np.array([15., 10., 16., 11., 9., 11., 10., 18.])
#DIMENSIONS
coords = {"school": ["A", "B", "C", "D", "E", "F", "G", "H"]}

samples = 4000
chains = 2
tune = 1000
fileName = "inference_8_schools_centered"
centered_eight = pm.Model(coords=coords)
with centered_eight:
    #PRIORS
    mu = pm.Normal('mu', mu = 0, sigma = 5)
    tau = pm.HalfCauchy('tau', beta = 5)
    theta = pm.Normal('theta', mu = mu, sigma = tau, dims = 'school')

    #OBSERVED VARS
    y = pm.Normal('y', mu = theta, sigma = sigma, observed = obs,
                  dims = 'school')

    #INFERENCE
    trace = pm.sample(samples, chains = chains, tune = tune)
    prior = pm.sample_prior_predictive(samples = samples)
    posterior_predictive = pm.sample_posterior_predictive(trace,
                                                           samples = samples)

#STEP 1: inference results into ArviZ InferenceData object
data_c = az.from_pymc3(trace = trace_c, prior = prior_c,

```

```

posterior_predictive = posterior_predictive_c)
#STEP 2: extract model graph & attach it to InferenceData object
dag_c = get_dag(centered_eight)
data_c.sample_stats.attrs["graph"] = str(dag_c)
#STEP 3: save data
arviz_to_json(data_c, fileName+'.npz')

```

The PyMC3 model specification presented here for the eight schools problem refer to the corresponding PyMC3 example available on-line² in the PyMC3 on-line documentation.

A.3 Drivers' Reaction Time Models

The observed data used in the models of the drivers' reaction time problem are retrieved from a study carried out by Belenky et al. [2003]. The purpose of this study was to measure the effect of sleep deprivation on cognitive performance. Eighteen lorry drivers were chosen and were restricted to 3 hours of sleep per day during the trial. Their reaction time to a visual stimulus was measured on each day of the experiment for 10 days. A simple model that could model the variation in reaction times of the lorry drivers across the days of the sleep-deprivation is a linear regression model of the following form: the reaction time for a driver $i, i \in [0, 17]$ on day $t, t \in [0, 9]$ would follow a normal distribution with a mean value floating on a straight line of the form $a_i + t \cdot b_i$. A greater reaction time on every next day of the sleep-deprivation trial is anticipated. Thus, a straight line with positive slope would be an appropriate model for modelling the reaction time of the drivers across the days of sleep-deprivation.

The *homogenous* model is described by the following probabilistic statements:

$$a \sim \text{Normal}(\mu = 100, \sigma = 250)$$

$$b \sim \text{Normal}(\mu = 10, \sigma = 250)$$

$$\text{sigma} \sim \text{Half-Normal}(\mu = 0, \sigma = 250)$$

$$y_{\text{pred}_i} \sim \text{Normal}(\mu = a + t \cdot b, \sigma = \text{sigma})$$

The homogenous model's definition in PyMC3 is given by the following lines of code:

```

import pymc3 as pm
import pandas as p

#DATA
DATAPATH = './evaluation_sleepstudy.csv'

```

²https://www.pymc.io/projects/docs/en/v3/pymc-examples/examples/diagnostics_and_criticism/Diagnosing_biased_Inference_with_Divergences.html

```

reactions = pd.read_csv(DATAPATH, usecols=['Reaction', 'Days', 'Subject'])
driver_idx, drivers = pd.factorize(reactions["Subject"], sort=True)
day_idx, days = pd.factorize(reactions["Days"], sort=True)
#DIMENSIONS
coords = {"driver": drivers, "driver_idx_day": reactions.Subject}

samples = 4000
chains = 2
tune = 2000
with pm.Model(coords = coords) as fullyPooled_model:
    #PRIORS
    a = pm.Normal("a", mu = 100, sd = 250)
    b = pm.Normal("b", mu = 10, sd = 250)
    sigma = pm.HalfNormal("sigma", sd = 200)
    #OBSERVED VARS
    y_pred = pm.Normal('y_pred', mu = a + b*day_idx, sd = sigma,
                        observed = reactions.Reaction,
                        dims = "driver_idx_day")

    #INFERENCE
    trace_p = pm.sample(samples, chains = chains, tune = tune)
    prior_p = pm.sample_prior_predictive(samples = samples)
    posterior_predictive_p =
    pm.sample_posterior_predictive(trace_p, samples = samples)

```

Each lorry driver has his individual characteristics that define their driving practices and performance like skills, experience, and endurance. On the other hand, all lorry drivers have some common driving-related characteristics. For example, they are all professional drivers that underwent similar training or they are all used to driving under sleep-deprivation conditions. Thus, the model that will model the data generation mechanisms of lorry drivers' reaction times should be able to reflect both the individual and common characteristics of the lorry drivers. A hierarchical probabilistic model would be able to capture the uncertainty of the overall reaction time estimation of the lorry drivers taking into consideration the uncertainty of the individual driver's reaction behavior, at the same time. Inferences at the individual driver level and estimations at the whole population of lorry drivers level will be enabled through the hierarchical model.

The *hierarchical* model is described by the following probabilistic statements:

$$\mu_a \sim \text{Normal}(\mu = 100, \sigma = 250)$$

$$\sigma_a \sim \text{Half-Normal}(\mu = 0, \sigma = 250)$$


```

y_pred = pm.Normal('y_pred',
                   mu = a[driver_idx]+b[driver_idx]*day_idx,
                   sd = sigma[driver_idx],
                   observed = reactions.Reaction,
                   dims = "driver_idx_day")

#INFERENCE
trace_hi = pm.sample(draws = samples, chains = chains,
                    tune = tune)

prior_hi = pm.sample_prior_predictive(samples = samples)
posterior_predictive_hi =
pm.sample_posterior_predictive(trace_hi, samples = samples)

```

The hierarchical and homogenous models of the lorry drivers reaction times refer to problems included in the book of Lambert [2018a]. The problem set questions and answers and the data could be found on-line [Lambert, 2018b].

A.4 Stochastic Volatility Model

The aim of the stochastic volatility model is to model the day returns of assets. Assets prices are variable and they have a time-varying volatility. There are periods, when returns are highly variable, while in other periods they are more stable. This variability is modeled by using a latent volatility variable. The following probabilistic statements describe the model:

$$\text{step_size} \sim \text{Exp}(\lambda = 50) \quad (\text{A.8})$$

$$v \sim \text{Exp}(\lambda = 0.1) \quad (\text{A.9})$$

$$\text{volatility}_i \sim \text{Normal}(\mu = \text{volatility}_{i-1}, \sigma = \text{step_size}^{-2}) \quad (\text{A.10})$$

$$\log(\text{return}_i) \sim \text{StudentT}(v = v, \mu = 0, \lambda = \exp(-2 \cdot \text{volatility}_i)) \quad (\text{A.11})$$

where $\log(\text{return}_i)$ is the log of the return on date i , and volatility_i is the latent volatility variable on date i . The actual observations of the model are the daily returns of the S&P 500 from May 2008 to November 2019.

The code for the PyMC3 definition of this model is the following:

```

import pandas as pd
import pymc3 as pm

#DATA
returns = pd.read_csv(pm.get_data('SP500.csv'), parse_dates=True,
                      index_col = 0)

```

```

dates = returns.index.strftime("%Y/%m/%d").tolist()

#DIMENSIONS
coords = {"date": dates}

samples = 2000
tune = 2000
chains = 2
with pm.Model(coords = coords) as model:
    #PRIORS
    step_size = pm.Exponential('step_size', 10)
    volatility = pm.GaussianRandomWalk('volatility',
    sigma = step_size, dims = 'date')
    nu = pm.Exponential('nu', 0.1)
    #OBSERVED VARS
    returns = pm.StudentT('returns', nu = nu,
                          lam = np.exp(-2*volatility),
                          observed = data["change"], dims = 'date')
    #INFERENCE
    trace = pm.sample(draws = samples, chains = chains, tune = tune)
    prior = pm.sample_prior_predictive(samples = samples)
    posterior_predictive =
    pm.sample_posterior_predictive(trace, samples = samples)

```

The PyMC3 model specification presented here for the stochastic volatility problem refers to the corresponding PyMC3 example available on-line³ in the PyMC3 on-line documentation.

A.5 The Coal Mining Disasters' Model

The aim of the coal mining disasters model is to predict the time point in the past when the number of the recorded coal mining-related disasters in the UK started to decline. It is assumed that the decline in the number of the coal mining-related disasters was linked to changes in the safety regulations. The dataset used provides the time series of recorded coal mining disasters in the UK from 1851 to 1962 [Jarrett, 1979]. The following probabilistic statements describe

³https://www.pymc.io/projects/docs/en/v3/pymc-examples/examples/case_studies/stochastic_volatility.html

the model:

$$\text{disasters}_t \sim \text{Poisson}(\lambda = \text{rate}_t), \text{rate}_t = \begin{cases} \text{early_rate} & \text{if } t \leq \text{switchpoint} \\ \text{late_rate} & \text{if } t > \text{switchpoint} \end{cases} \quad (\text{A.12})$$

$$\text{switchpoint} \sim \text{Uniform}(a = t_l, b = t_h) \quad (\text{A.13})$$

$$\text{early_rate} \sim \text{Exp}(\lambda = 1) \quad (\text{A.14})$$

$$\text{late_rate} \sim \text{Exp}(\lambda = 1) \quad (\text{A.15})$$

where disasters_t is the number of disasters in year t , rate_t the rate parameter of the Poisson distribution in year t , switchpoint is the switch-point, namely the year, when the change in the safety regulations occurred and there was a switch in the rate rate, early_rate is the rate parameter before the switch-point, and late_rate is the rate parameter after the switch-point.

The code for the PyMC3 definition of this model is the following:

```
import pandas as pd
import pymc3 as pm

#DATA
disaster_data = pd.Series([4, 5, 4, 0, 1, 4, 3, 4, 0, 6, 3, 3, 4, 0,
                          2, 6,
                          3, 3, 5, 4, 5, 3, 1, 4, 4, 1, 5, 5, 3, 4,
                          2, 5,
                          2, 2, 3, 4, 2, 1, 3, np.nan, 2, 1, 1, 1,
                          1, 3, 0, 0,
                          1, 0, 1, 1, 0, 0, 3, 1, 0, 3, 2, 2, 0, 1,
                          1, 1,
                          0, 1, 0, 1, 0, 0, 0, 2, 1, 0, 0, 0, 1, 1,
                          0, 2,
                          3, 3, 1, np.nan, 2, 1, 1, 1, 1, 2, 4, 2,
                          0, 0, 1, 4,
                          0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
                          1])

years = np.arange(1851, 1962)
years_missing = [1890, 1935]

#DIMENSIONS
coords = {"year": years}
```

```
samples = 10000
tune = 10000
chains = 2
with pm.Model(coords = coords) as disaster_model:
    #PRIORS
    switchpoint = pm.DiscreteUniform('switchpoint',
    lower = years.min(), upper = years.max(), testval = 1900)
    early_rate = pm.Exponential('early_rate', 1)
    late_rate = pm.Exponential('late_rate', 1)
    rate = pm.math.switch(switchpoint >= years, early_rate,
    late_rate)
    #OBSERVED VAR
    disasters = pm.Poisson('disasters', rate,
    observed = disaster_data, dims = 'year')
    #INFERENCE
    trace = pm.sample(samples, chains = chains, tune = tune)
    prior = pm.sample_prior_predictive(samples = samples)
    pm.sample_posterior_predictive(trace, samples = samples)
```

The PyMC3 model specification presented here for the coal mining disasters problem refers to the corresponding PyMC3 example available on-line⁴ in the PyMC3 on-line documentation.

⁴<https://pymcmc.readthedocs.io/en/jss-gp/tutorial.html>

Appendix B

arviz_json Package

The *arviz_json* [Williamson, 2019] Python package exports *ArviZ InferenceData* objects into a zip file containing JSON metadata and a collection of *numpy*¹ format arrays of data from the *InferenceData* object. There are two JSON structures that comprise the JSON metadata. The first one captures the graph-related information of the PyMC3 models and the second one links the model's variables to the data arrays and provides information about the type of the samples (prior, posterior), the dimensions and coordinates. In this section, the JSON format of these structures is described.

B.1 Graph JSON Structure

A prototype JSON structure captures the model-related information that is necessary for the construction of an informative DAG of the model. The JSON structure is according to the following scheme:

```
{<var_name>: {"name"           : <var_name>,
              "type"          : <var_type>,
              "parents"       : [],
              "size"          : 0,
              "dims"          : [],
              "coords"        : {},
              "distribution"  : {"dist"     :<dist_name>,
                                "type"     :<dist_type>,
                                "shape"    : []
                               }
            }
}
```

¹<https://numpy.org/devdocs/reference/generated/numpy.lib.format.html>

where `<var_name>` is a string of the variable's name based on the PyMC3 model specification, `<var_type>` is a string in the set {"free", "observed", "deterministic"}, `<dist_name>` is a string of the prototype distribution name of the variable according to PyMC3 distributions class names², `<dist_type>` is a string of the type of the distribution in the set {"continuous", "discrete", ""}. The field "parents" takes a list of strings with the variables' names of the parent nodes. The field "size" has to do with the *free* variables of the model and captures the overall size of the variable, i.e, product of the "shape" array. This value defines the number of draws from the distribution for each MCMC sample. The field "shape" captures the shape of the MCMC samples array for the corresponding parameter. The field "dims" takes a list of strings indicating the names of the variable's dimensions. Finally, the field "coords" takes a dictionary, where each entry corresponds to an indexing dimension and is of the form: {`<idx_dim>`: []}, `<idx_dim>` is an element in "dims" and its value is a list of the coordinates of each `<idx_dim>`.

At this point, it's worth raising two points that are very important for the creation of visualization tools based on PPLs' outputs. There is some confusion about the definition of variables' shape by the various PPLs. The shape of a variable's MCMC samples arrays cannot easily be interpreted. For example, a shape (2,) could be interpreted as either getting 2 draws from a uni-variate distribution, or 2 draws from 2 separate uni-variate distribution, or finally, one draw from a bi-variate distribution.

The second point that should be highlighted here is the definition of the indexing dimensions of the variables. The semantic definition of model's indexing dimensions and coordinates is necessary for the creation of interactive widgets by the IPME tool. Current implementations of PPLs do not always provide this possibility. In those cases, we should manually provide a semantic definition of the indexing dimensions to ArviZ and *arviz_json*, which is an inevitable extra step that we hope will be taken in the probabilistic programming-level during the model specification process.

B.2 ArviZ InferenceData Export Into *numpy* Arrays and Metadata File

The ArviZ library provides an API for exporting inference data of different inference back-ends and programming languages into ArviZ *InferenceData*³ data structures. Fig. B.1(a) presents the groups that the InferenceData object creates to organize the inference data of the PyMC3 hierarchical model for the eight schools' problem. Fig. B.1(b) presents the `xarray.Dataset` for the posterior group of the InferenceData object of the eight schools' model.

The *arviz_json* package transforms the ArviZ InferenceData object into a collection of `numpy`

²<https://docs.pymc.io/api/distributions.html>

³<https://arviz-devs.github.io/arviz/schema/schema.html>

```

Inference data with groups:
  > posterior
  > sample_stats
  > log_likelihood
  > posterior_predictive
  > observed_data
  > prior
  > prior_predictive
      (a)

<xarray.Dataset>
Dimensions: (chain: 2, draw: 4000, school: 8)
Coordinates:
  * chain      (chain) int32 0 1
  * draw      (draw) int32 0 1 2 3 4 5 6 7 ... 3993 3994 3995 3996 3997 3998 3999
  * school    (school) <U1 'A' 'B' 'C' 'D' 'E' 'F' 'G' 'H'
Data variables:
  mu          (chain, draw) float64 6.026 7.623 7.596 6.857 ... 5.822 4.066 4.066
  theta      (chain, draw, school) float64 7.491 7.137 6.837 ... 5.916 6.214
  tau        (chain, draw) float64 1.9 1.694 1.748 1.699 ... 1.822 1.84 1.84
Attributes:
  created_at:                2020-05-28T11:35:36.495866
  arviz_version:             0.7.0
  inference_library:         pymc3
  inference_library_version: 3.8
      (b)

```

Figure B.1: The *ArviZ InferenceData* data structure of the PyMC3 hierarchical model of the eight schools' problem. (a) The groups in which inference data is organized in the *InferenceData* data structure. (b) The *xarray.Dataset* that corresponds to the posterior group of the *InferenceData* object.

arrays and a JSON metadata structure to provide links between the model's variables and the data arrays and to provide information about the type of the samples (prior, posterior), the dimensions and the coordinates. This JSON structure is according to the following scheme:

```

{"inference_data": {
    "observed_data"           :<sub_structure_1>,
    "posterior"               :<sub_structure_1>,
    "posterior_predictive"    :<sub_structure_1>,
    "prior"                   :<sub_structure_1>,
    "prior_predictive"        :<sub_structure_1>,
    "sample_stats"            :<sub_structure_1>,
    "log_likelihood"          :<sub_structure_1>
  }
}

```

<sub_structure_1> is defined as:

```
{ "array_names": {<var_name>: <array_name>},
  "attrs"       : <attrs_structure>,
  "dims"        : <dims_structure>,
  "coords"      : <coords_structure>,
  "vars"        : <vars_structure>
},
```

where <var_name> is a string of the variable's name according to the PyMC3 model specification, <array_name> is a string of the npy data array name that holds the corresponding MCMC samples of the variable. All the variables of the model that are encountered within the corresponding group, e.g. posterior, are included here and linked to their data array. The field "attrs" captures metadata that has to do with version of the ArviZ used and the inference library, the field "dims" captures all the different indexing dimensions that are encountered in this group's variables, the field "coords" captures the coordinates of each encountered dimension, and the field "vars" captures information related to dimensions, coordinates, shape, and name of data array for each variable.

<attrs_structure> is defined as:

```
{ "arviz_version"      : "0.7.0",
  "created_at"         : <timestamp>,
  "inference_library" : "pymc3",
  "inference_library_version": "3.8"
}
```

where the field "arviz_version" takes a string of the ArviZ version that was used to export the inference data into an ArviZ InferenceData object, e.g. version 0.7.0, <timestamp> is a string of timestamp when the inference data was exported to ArviZ InferenceData, the field "inference_library" takes a string of the PPL that was used for the inference, e.g. PyMC3, and the field "inference_library_version" takes a string of the inference library version, e.g. version 3.8.

<coords_structure> is defined as:

```
{ "chain"       : [1, 2],
  <dim_name>    : ["A", "B", "C"],
  "draw"        : [1, 2, 3, ..., 100]
}
```

where each field corresponds to each dimension that is encountered within this group (e.g. posterior), and takes a list of the corresponding dimension's coordinates. The fields "chain" and "draw" are fixed and common for all variables. The field "chain" takes a list of integers indicating the incremental indexes of the MCMC chains of samples that the sampler

retrieved, e.g. [1,2] for two chains, the field `<dim_name>` is a string of the name of the indexing dimension – extra fields of this type are added for any extra indexing dimension present in this group – and takes a list of the actual semantic coordinates of this dimension. The field "draw" takes a list of incremental indexes for each MCMC sample. The length of this list equals the number of the samples drawn from the (prior) posterior (predictive) distribution and is defined in the `pm.sample()`, `pm.sample_prior_predictive()`, `pm.sample_posterior_predictive()` PyMC3 methods, for example 100 samples.

`<dims_structure>` is defined as:

```
{ "chain"          : 2,
  <dim_name>      : 3,
  "draw"          : 100
}
```

where there is one field per dimension and each field takes an integer indicating the number of coordinates each dimension has.

`<vars_structure>` is defined as:

```
{ <var_name>: { "array_name"  : <array_name>
               "attrs"       : {},
               "dims"        : ["chain", "draw", <dim_name>],
               "dtype"       : "<f8",
               "shape"       : [2, 100, 3]
             }
}
```

where the field "dtype" captures the data type (floating point, integer).

Appendix C

User Study 1

C.1 Participants' Training

The training part of the user study comprised of four training videos. The training videos were same for both groups of participants; the static (SG) and interaction (IG) group, with some slight differences based on the visualization condition. Participants could ask questions once they had watched the videos.

The *first training video* (IG: <https://youtu.be/6yrBrL6amiQ>, and SG: <https://youtu.be/zeonqIgHspk>) was introducing the researcher and provided her contact details, explaining the freedom to withdraw and the purpose of the user study, and provided a description of the tasks and structure of the user study.

The *second training video* (IG: <https://youtu.be/iPf8bwdxKy8>, and SG: <https://youtu.be/q0ZCM5KOxbI>) was making an introduction to basic probabilistic concepts such as random variables, probability, (joint) probability mass function/density, sampling, density/scatter/rug plots.

The *third training video* (IG: <https://youtu.be/zQhy-LYJGQ4>, and SG: <https://youtu.be/ow86A6cvHjE>) was presenting and demonstrating the use of the IPP tool.

The *fourth training video* (IG: <https://youtu.be/9mfhepxCeRU>, and SG: <https://youtu.be/uDcqwLqQFDA>) was presenting example tasks (one per Task Type) and how the presented visualizations could be used to answer the questions. For example, participants in the SG were shown how they can interpret the shape of a pair plot in terms of relations between variables when they think of it on the basis of conditioning.

C.2 Task Models

C.2.1 Model 1

The first model was designed to predict the mean November temperature (°C) in Scotland. The model consists of an observed random variable for the predicted temperature and a set of unidentified parameters a , b , and c . The mean value of the prior distribution for the mean value of temperature's distribution was set to 2 because out of prior experience of living in Scotland, the temperatures at this time of the year are usually nearly above 0. The standard deviations of the prior distributions were set to 10 to make them weakly informative.

$$a \sim \text{Uniform}(\text{lower} = 80, \text{upper} = 100)$$

$$b \sim \text{Normal}(\mu = 2, \sigma = 10)$$

$$c \sim \text{Half-Normal}(\sigma = 10)$$

$$\text{temperature} \sim \text{Normal}(\mu = b, \sigma = c)$$

The PyMC3 code of the model can be found in https://github.com/evdoxiataka/ipme/tree/master/examples/user_study/min_temperature. The prior samples from the model used in the study could be found in the file `study_analysis\data\min_temperature.npz` in [Taka et al., 2022]. The data used for the definition of the likelihood was the average minimum temperature in Scotland in month November for the years 1884-2020 (retrieved from <https://www.metoffice.gov.uk/pub/data/weather/uk/climate/datasets/Tmin/date/Scotland.txt>).

C.2.2 Model 2

The second model was designed to predict the output of an engine that generates random real numbers. The model consists of an observed random variable for the predicted `random_number` and a set of unidentified parameters a , b , and c . For the parameterization of the uniform likelihood's bounds, we subtract a positive number c (sampled from a half-normal distribution) from a number a (sampled from a normal distribution centered around 0) to set the lower bound and we add it to a to set the upper bound.

$$a \sim \text{Normal}(\mu = 0, \sigma = 10)$$

$$b \sim \text{Half-Normal}(\sigma = 10)$$

$$c \sim \text{Half-Normal}(\sigma = 20)$$

$$\text{random_number} \sim \text{Uniform}(\text{lower} = a - c, \text{upper} = a + c)$$

The PyMC3 code of the model can be found in https://github.com/evdoxiataka/ipme/tree/master/examples/user_study/random_number_generator. The prior samples from the model used in the study could be found in the file `study_analysis\data\transformation.npz` in [Taka et al., 2022]. The data used for the definition of the likelihood was synthetically created.

C.2.3 Model 3

The third model was designed to predict the reaction time (in msec) of lorry drivers under sleep deprivation conditions. The model consists of observed random variables for the predicted `reaction_time` of each lorry driver ($i \in 1, 2, \dots, 18$), a set of priors a, b, σ_i and d , and a set of hyper-priors c, e, f, g and h . The day variable takes values in the $day \in 1, 2, \dots, 10$. The visualizations of the tasks in the user study regarding this problem included only the parameters a, b, c, d , and the `reaction_time` observed variable.

For setting the prior for parameter a , namely the intercept of the `reaction_time`'s mean value, we set a hyper-prior for the mean value of its prior distribution with mean value equal to 100 msec (0.1 sec) and standard deviation 150 msec (0.15 sec). Crudely, this would represent the mean value and standard deviation of the drivers' reaction time on day 0. We were expecting the drivers to have some small reaction time above 0 on day 0 of driving, because they were well-rested, and this reaction time to increase as the days pass by and the drivers become sleep-deprived. For setting the prior for parameter b , namely the slope of the `reaction_time`'s mean value that represents the amount of time in msec that the reaction time of the driver increases in each day, we set a hyper-prior for the mean value of its prior distribution with mean value equal to 10 msec (0.01 sec) and standard deviation 100 msec (0.1 sec). We were expecting that the drivers' reaction time would increase with day, but we had no previous knowledge of how much this increase could be.

We set the standard deviation of the a parameter to a higher value (150 msec) than the standard deviation of the b parameter (100 msec), as we were expecting more variation to the drivers' reaction times at rest as this could reflect their individual traits, than to the effect of sleep deprivation on drivers (we thought that tiredness more or less affects drivers in the same ways). Finally, for the prior distribution for the standard deviation of the `reaction_time`'s likelihood a hyper-prior was set with standard deviation equal to 200 msec to account for bigger variations among drivers and days.

$$c \sim \text{Normal}(\mu = 100, \sigma = 150)$$

$$e \sim \text{Half-Normal}(\sigma = 150)$$

$$f \sim \text{Normal}(\mu = 10, \sigma = 100)$$

$$g \sim \text{Half-Normal}(\sigma = 100)$$

$$h \sim \text{Half-Normal}(\sigma = 200)$$

$$a_i \sim \text{Normal}(\mu = c, \sigma = e)$$

$$b_i \sim \text{Normal}(\mu = f, \sigma = g)$$

$$\text{sigma}_i \sim \text{Half-Normal}(\sigma = h)$$

$$d \sim \text{Normal}(\mu = 0, \sigma = 10)$$

$$\text{reaction_time}_i \sim \text{Normal}(\mu = a_i + \text{day} \cdot b_i, \sigma = \text{sigma}_i)$$

The PyMC3 code of the model can be found in https://github.com/evdoxiataka/ipme/tree/master/examples/user_study/reaction_times. The prior samples from the model used in the study could be found in the file `study_analysis\data\reaction_times_hierarchical.npz` in [Taka et al., 2022]. The data used for the definition of the likelihood was taken from the study presented in [Belenky et al., 2003].

C.3 Analysis

The Bayesian models used for the analysis of the user study's collected data were designed and interpreted in PyMC3. The code for the models' specification is presented here and could be found as Python Jupyter Notebooks along with the collected data from the user study in [Taka et al., 2022]. Please note that the variables' names are slightly different in the presented code below to be in alignment with Kruschke-style diagrams of the models presented in Fig. 5 of the paper.

C.3.1 Accuracy's Model

Two separate models were used for the analysis of accuracy; one for T1 tasks and the other for the T2-T3 tasks. The models were different in the likelihood used for the accuracy. A binomial likelihood was used for T1 tasks because multiple selections were allowed. A Bernoulli likelihood was used for the rest of tasks because only a single selection was allowed.

Beta priors with $\alpha = 1.0$ and $\beta = 1.0$ were set in both models for the probabilities of success (thetaIG and thetaSG). These priors correspond to a uniform distribution with bounds between 0 and 1 and is a reasonable uninformative option in this case.

C.3.1.1 Model for T1

```
import pymc3 as pm
import numpy as np
```

```

coords = {"task": t_ids}
with pm.Model(coords=coords) as model:
    #priors
    thetaIG = pm.Beta("thetaIG", alpha = 1.0, beta = 1.0,
                      dims = 'task')
    thetaSG = pm.Beta("thetaSG", alpha = 1.0, beta = 1.0,
                      dims = 'task')

    #likelihood
    accuracyIG = pm.Binomial("accuracyIG", n = n_i,
                             p = thetaIG[t_indices_i],
                             observed = answers_i)
    accuracySG = pm.Binomial("accuracySG", n = n_s,
                             p = thetaSG[t_indices_s],
                             observed = answers_s)

    #comparisons
    diff_of_thetas = pm.Deterministic("difference of thetas",
                                       thetaIG - thetaSG,
                                       dims='task')

    #inference
    trace = pm.sample(2000)

```

C.3.1.2 Model for T2-T3

```

import pymc3 as pm
import numpy as np

coords = {"task": t_ids}
with pm.Model(coords=coords) as model:
    #priors
    thetaIG = pm.Beta("thetaIG", alpha = 1.0, beta = 1.0,
                      dims = 'task')
    thetaSG = pm.Beta("thetaSG", alpha = 1.0, beta = 1.0,
                      dims = 'task')

```

```

#likelihood
accuracyIG = pm.Bernoulli("accuracyIG", p = thetaIG[t_indices_i],
                          observed = answers_i)
accuracySG = pm.Bernoulli("accuracySG", p = thetaSG[t_indices_s],
                          observed = answers_s)

#comparisons
diff_of_thetas = pm.Deterministic("difference of thetas",
                                   thetaIG - thetaSG,
                                   dims='task')

#inference
trace = pm.sample(2000)

```

C.3.2 Response Times' Model

The response times of the participants were continuous values and we assumed a normal likelihood to model them. A normal prior distribution was set for the μ and a half-normal prior distribution for the σ parameter of the response times' likelihood. The user study was designed so that each participant spends 2-3 min on average on each task. So, we set $\mu = 120$ sec for the priors and allowed for a variance of 60 sec to account for the fact that some tasks could be completed in less or more time depending on the complexity of the presented structure.

```

import pymc3 as pm
import numpy as np

coords = {"task": t_ids}
with pm.Model(coords=coords) as model:
    #priors
    groupIG_mean = pm.Normal("groupIG_mean", mu = 120, sd = 60,
                             dims = 'task')
    groupIG_std = pm.HalfNormal("groupIG_std", sd = 90,
                                 dims = 'task')

    groupSG_mean = pm.Normal("groupSG_mean", mu = 120, sd = 60,
                              dims = 'task')
    groupSG_std = pm.HalfNormal("groupSG_std", sd = 90,
                                 dims = 'task')

```

```

#likelihood
rtIG = pm.Normal("rtIG", mu = groupIG_mean[t_indices_i],
                 sd = groupIG_std[t_indices_i],
                 observed = times_i) # sec
rtSG = pm.Normal("rtSG", mu = groupSG_mean[t_indices_s],
                 sd = groupSG_std[t_indices_s],
                 observed = times_s) # sec

#comparisons
diff_of_means = pm.Deterministic("difference of means",
                                 groupIG_mean - groupSG_mean,
                                 dims = 'task')
effect_size = pm.Deterministic("effect size",
                                diff_of_means / np.sqrt((groupIG_std ** 2 +
                                                         groupSG_std ** 2) / 2),
                                dims = 'task')

#inference
trace = pm.sample(2000)

```

C.3.3 Confidence's Model

Although the structure of the model used for the analysis of confidence is the same as that used for the response times, the values of the parameters of priors were different. The recorded confidence levels of the participants were mapped to a $[-2, 2]$ scale. Thus, we centered the prior for the μ of the likelihood around 0, as we had no previous experience or knowledge about how high users' confidence would be, and allowed for a variance of 1 to create uninformative enough priors.

```

import pymc3 as pm
import numpy as np

coords = {"task": t_ids}
with pm.Model(coords=coords) as model:
    #priors
    groupIG_mean = pm.Normal("groupIG_mean", mu = 0, sd = 1,
                             dims = 'task')
    groupIG_std = pm.HalfNormal("groupIG_std", sd = 1,
                                dims = 'task')

```



```

groupSG_mean = pm.Normal("groupSG_mean", mu = 0, sd = 1,
                          dims = 'task')
groupSG_std = pm.HalfNormal("groupSG_std", sd = 1,
                             dims = 'task')

#likelihood
config = pm.Normal("config", mu = groupIG_mean[t_indices_i],
                  sd = groupIG_std[t_indices_i],
                  observed = conf_i)
confSG = pm.Normal("confSG", mu = groupSG_mean[t_indices_s],
                  sd = groupSG_std[t_indices_s],
                  observed = conf_s)

#comparisons
diff_of_means = pm.Deterministic("difference of means",
                                  groupIG_mean - groupSG_mean,
                                  dims = 'task')
effect_size = pm.Deterministic("effect size",
                                diff_of_means / np.sqrt((groupIG_std ** 2 +
                                                         groupSG_std ** 2) / 2),
                                dims = 'task')

#inference
trace = pm.sample(2000)

```

C.4 Tasks

Fig. SC.1-SC.19 present the study questions as they were presented to participants during the study in exactly the same order.

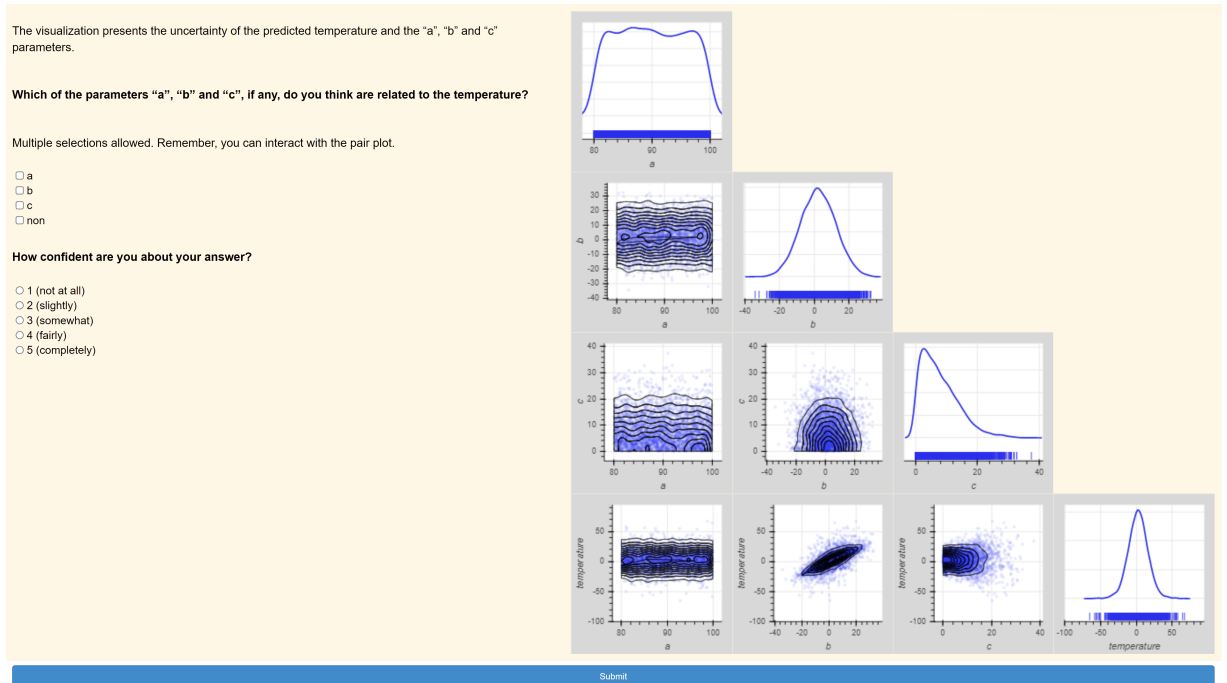


Figure C.1: Model 1 - Task t1 (T1).

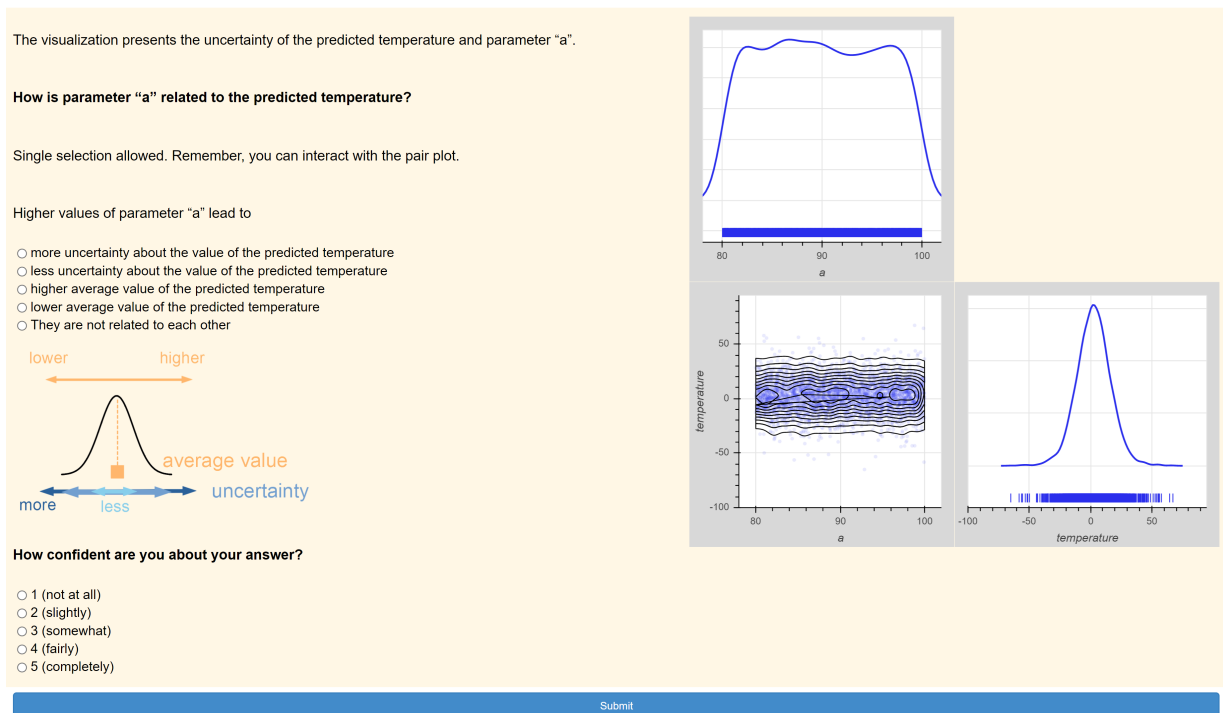


Figure C.2: Model 1 - Task t2 (T2).

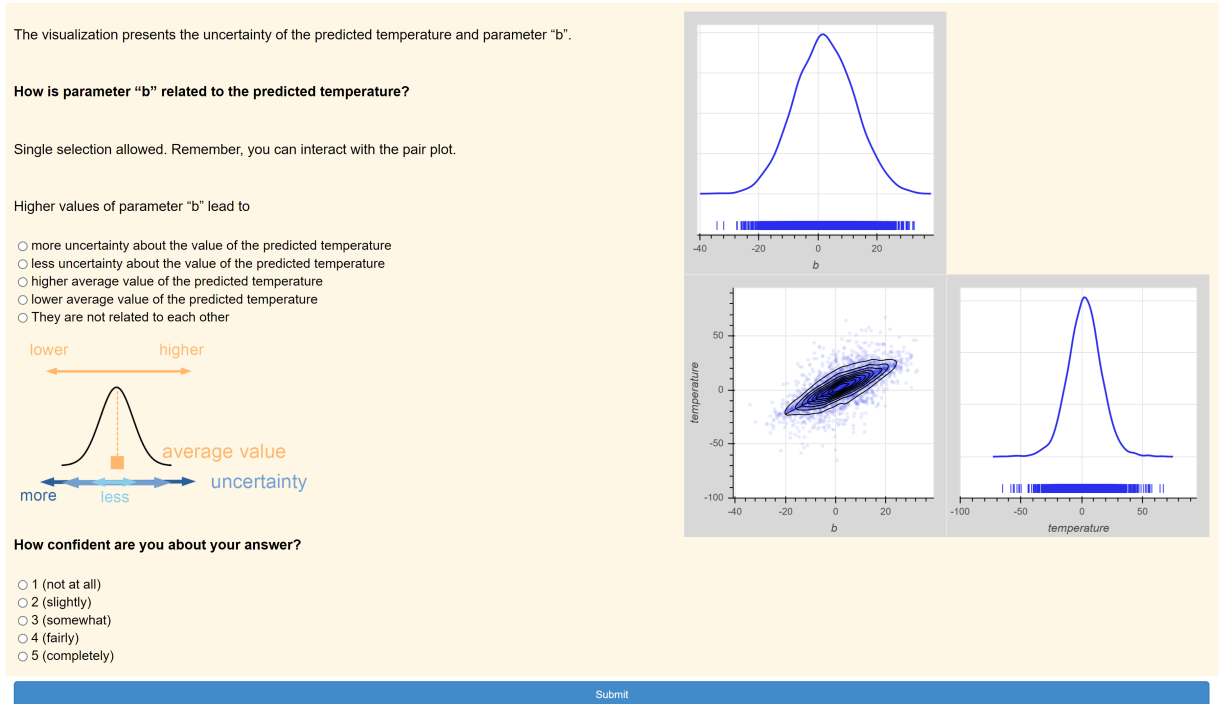


Figure C.3: Model 1 - Task t3 (T2).

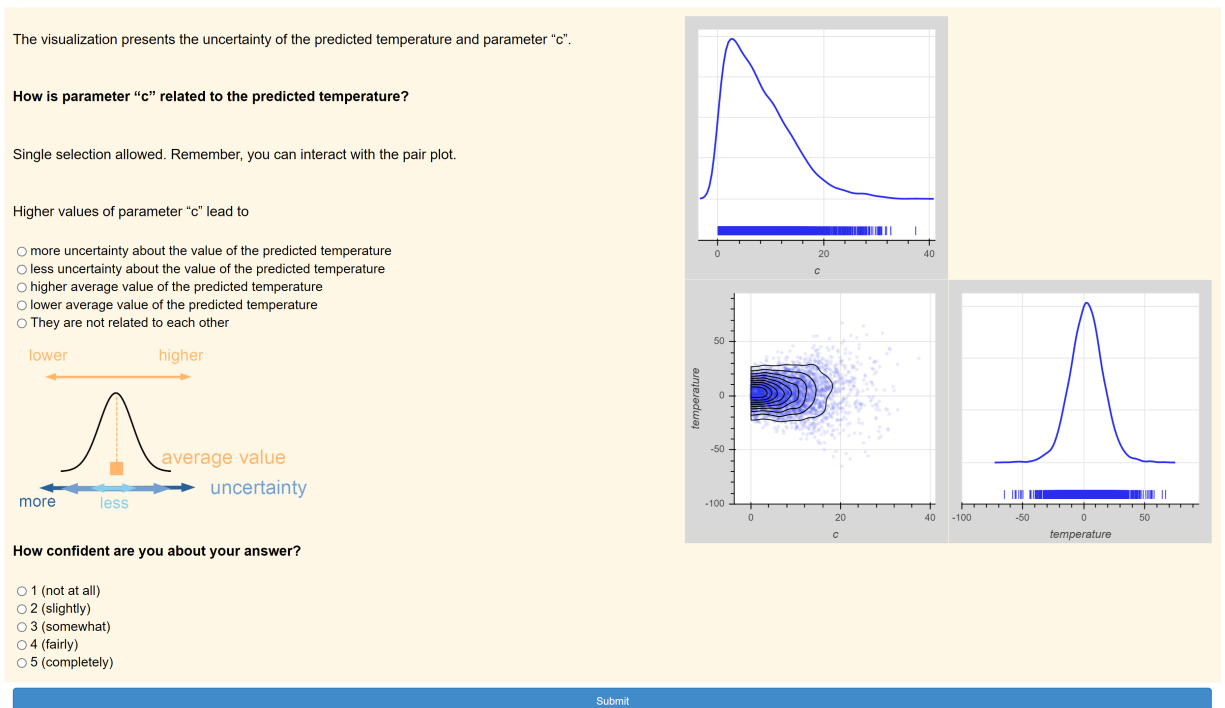


Figure C.4: Model 1 - Task t4 (T2).

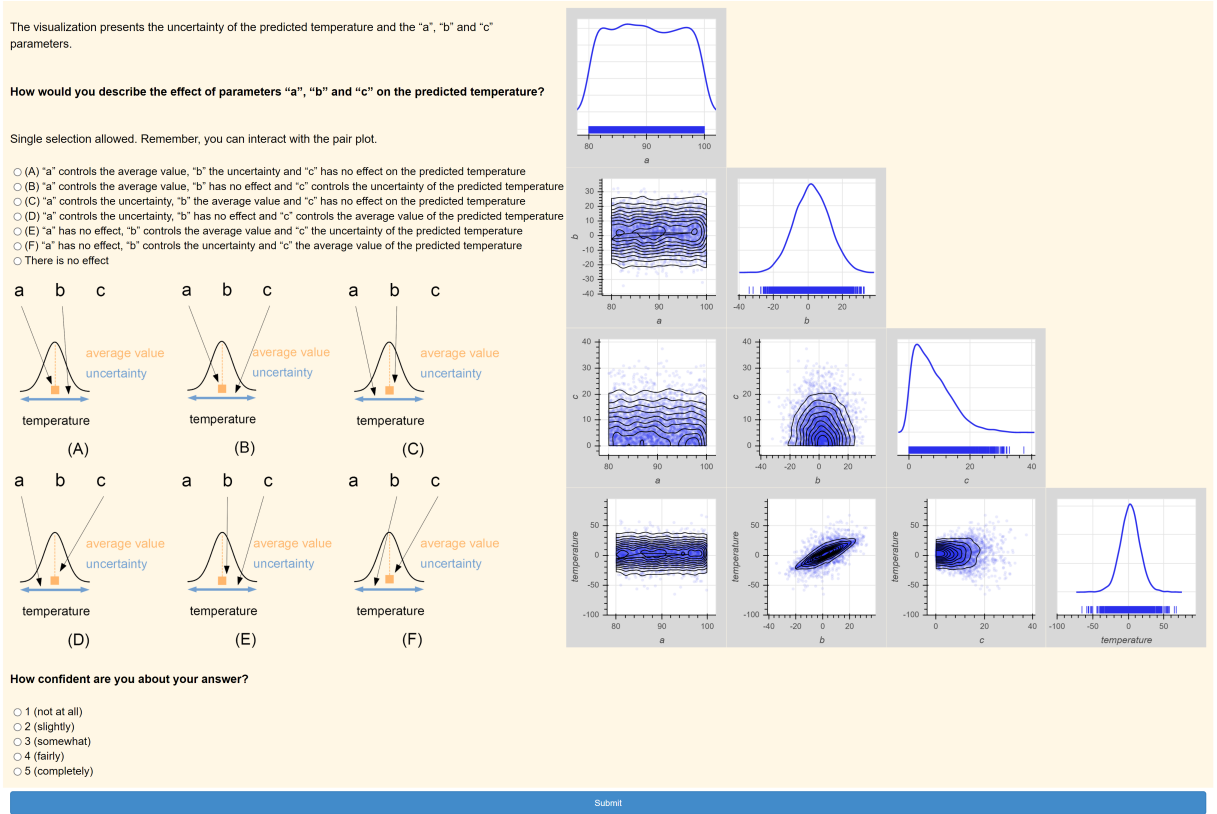


Figure C.5: Model 1 - Task t5 (T3).

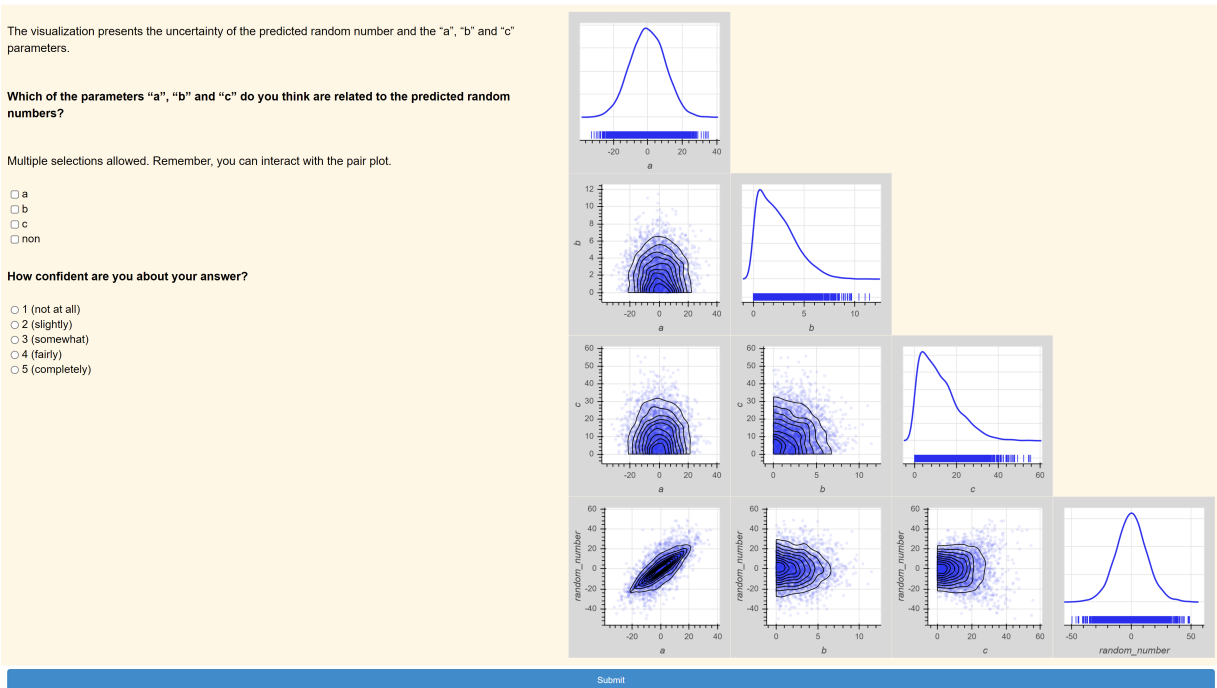


Figure C.6: Model 2 - Task t6 (T1).

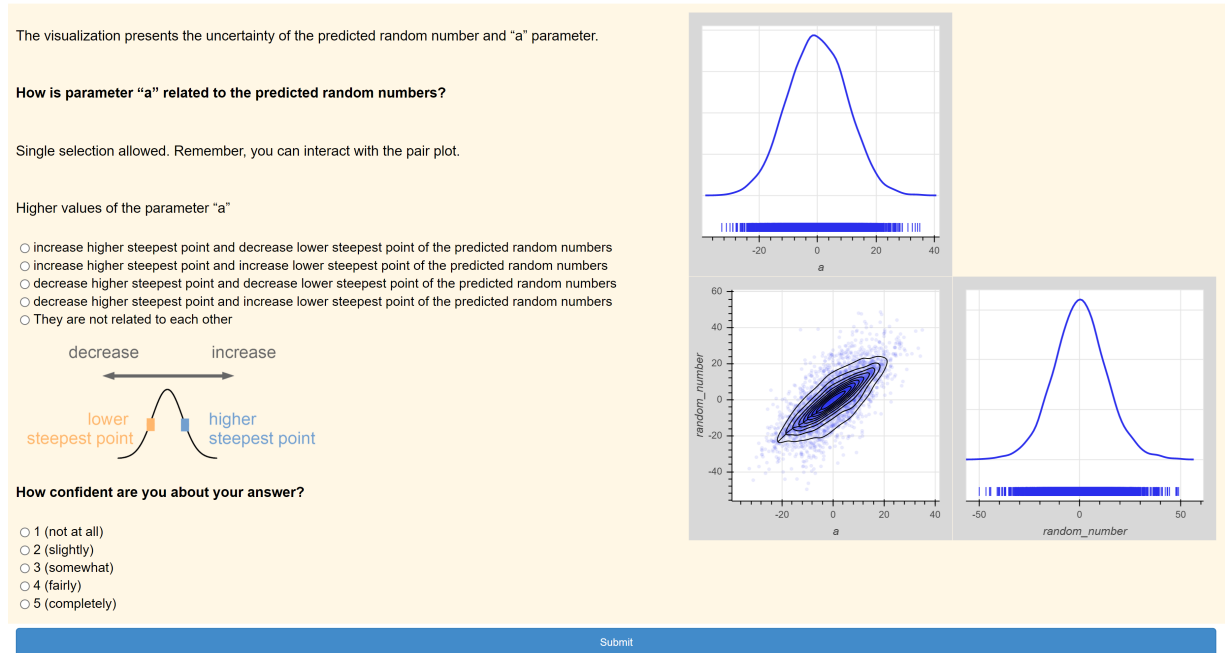


Figure C.7: **Model 2** - Task t7 (T2).

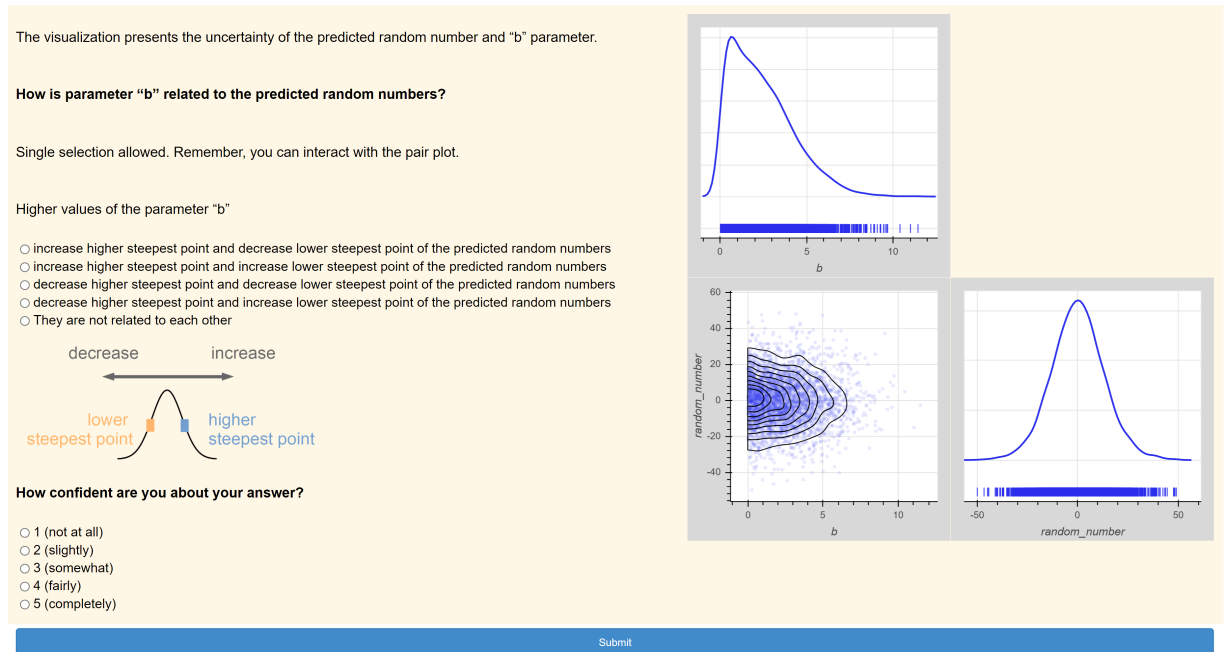


Figure C.8: **Model 2** - Task t8 (T2).

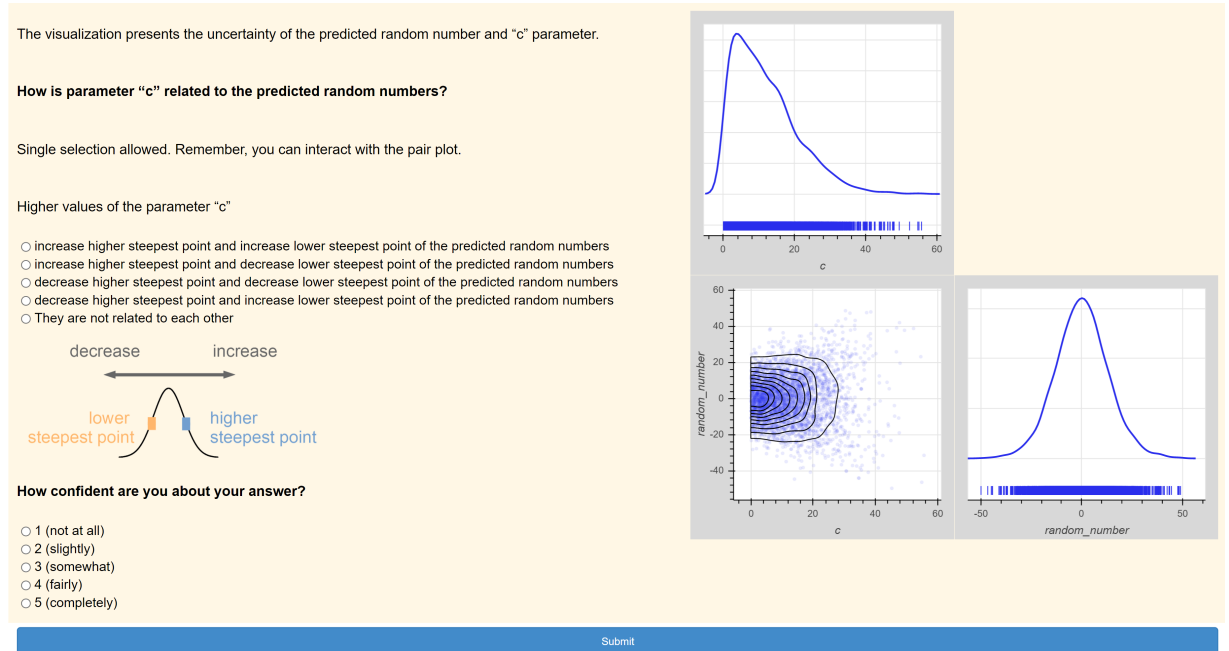


Figure C.9: Model 2 - Task t9 (T2).

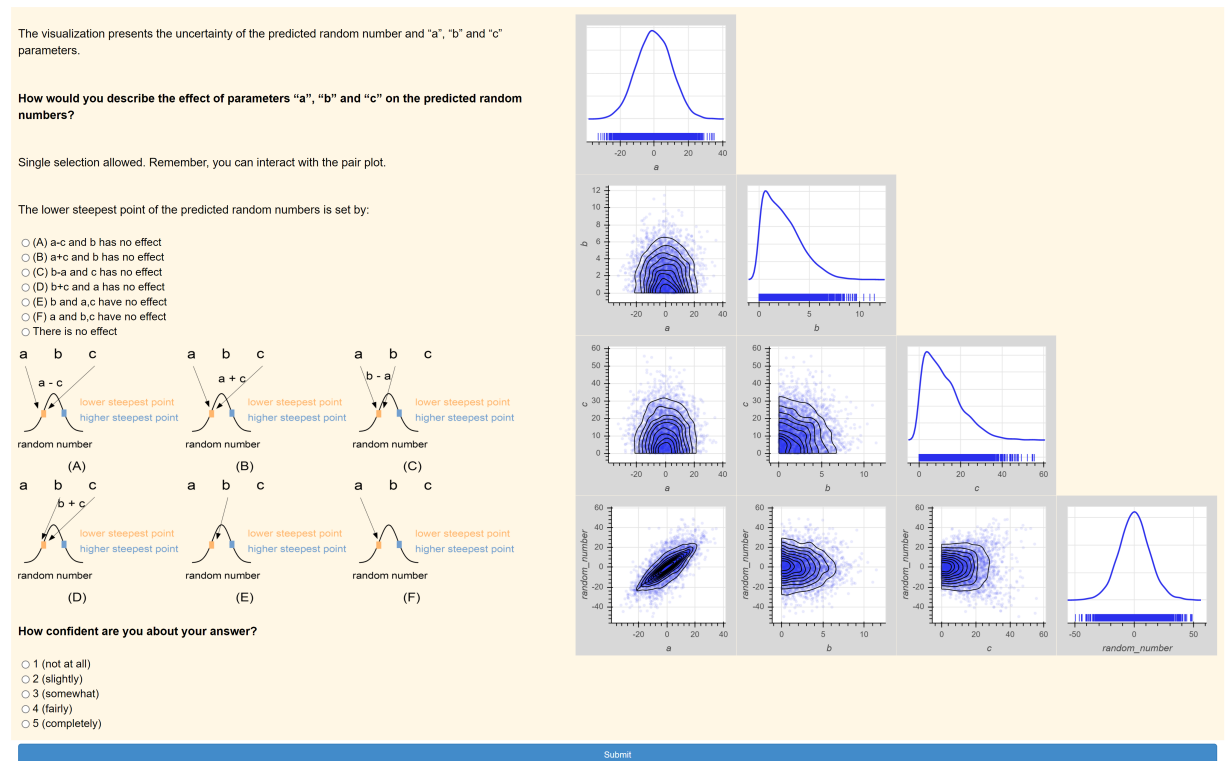


Figure C.10: Model 2 - Task t10 (T3).

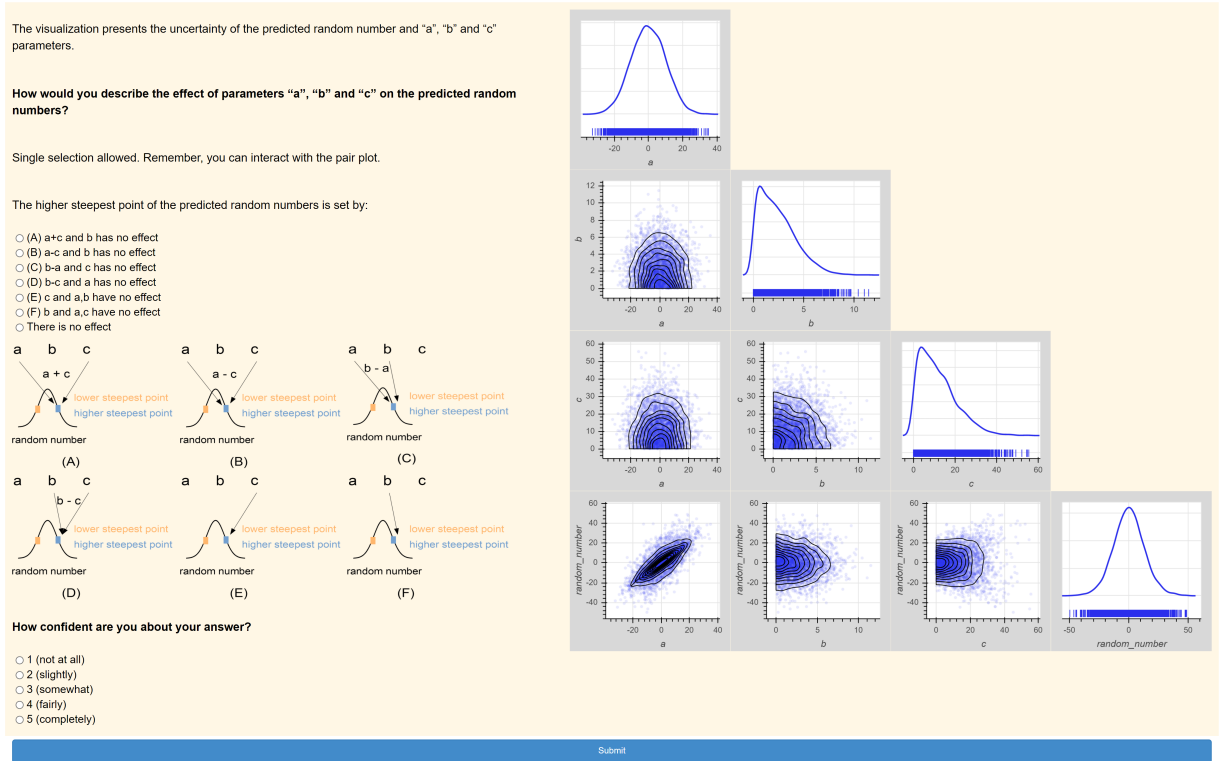


Figure C.11: Model 2 - Task t11 (T3).

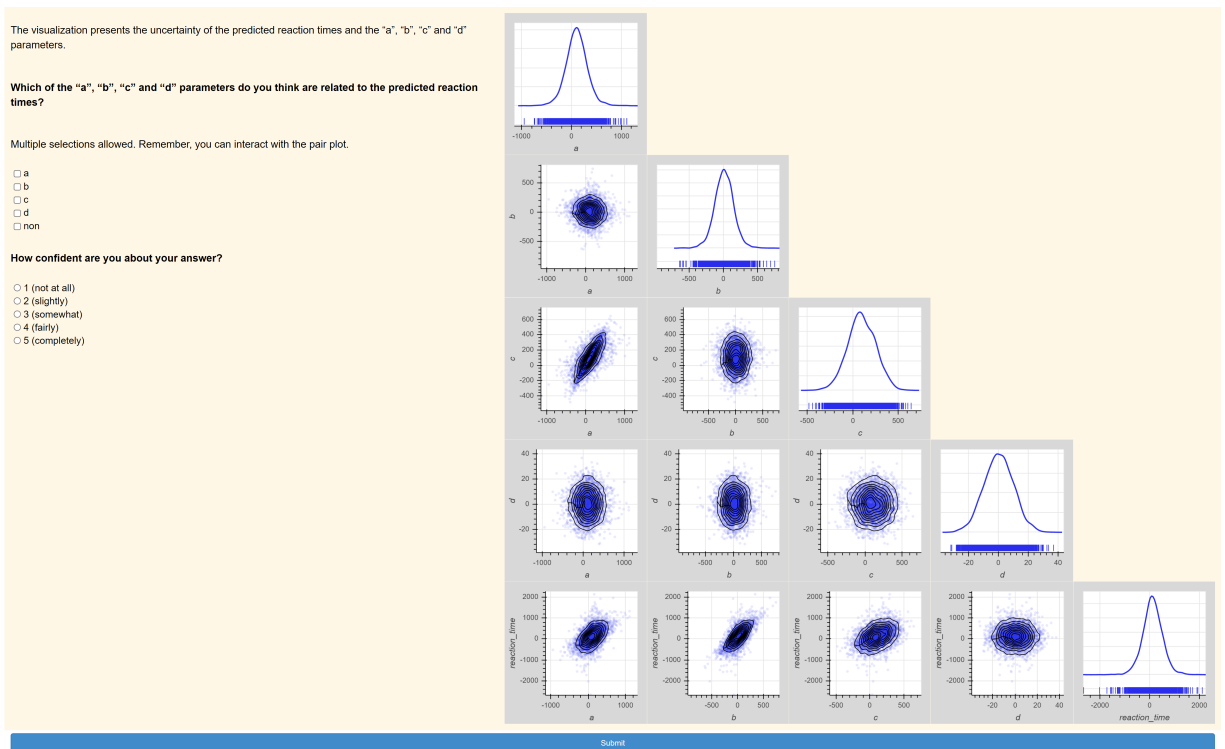


Figure C.12: Model 3 - Task t12 (T1).

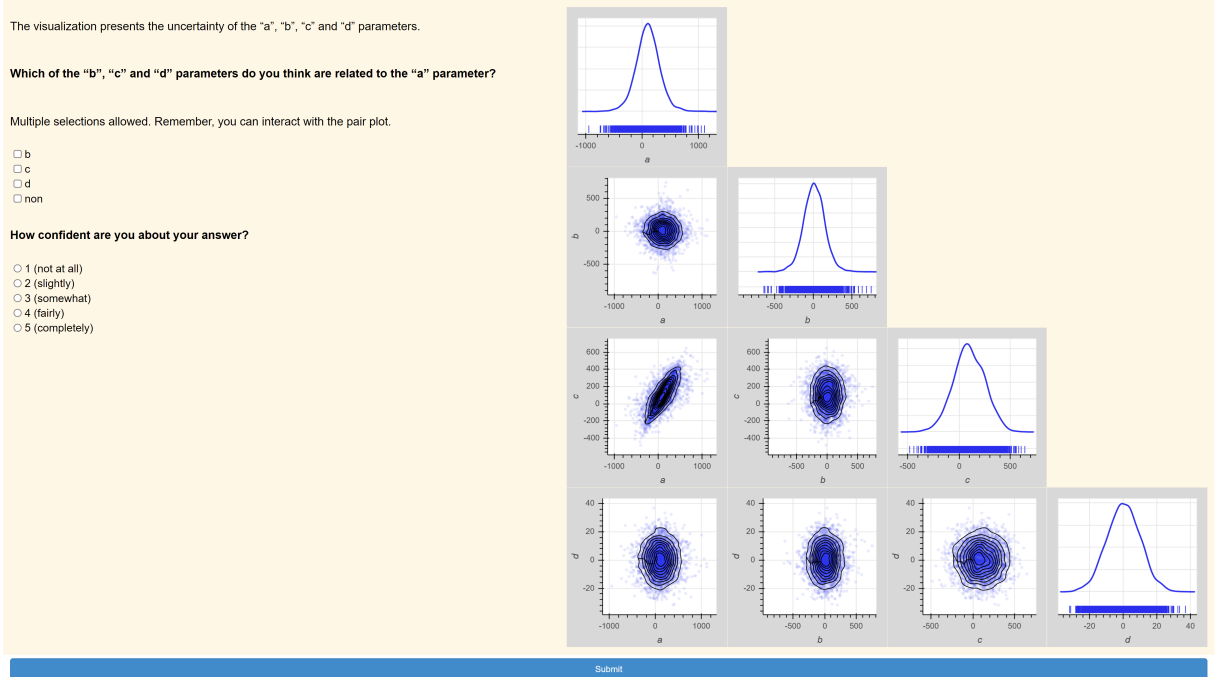


Figure C.13: Model 3 - Task t13 (T1).

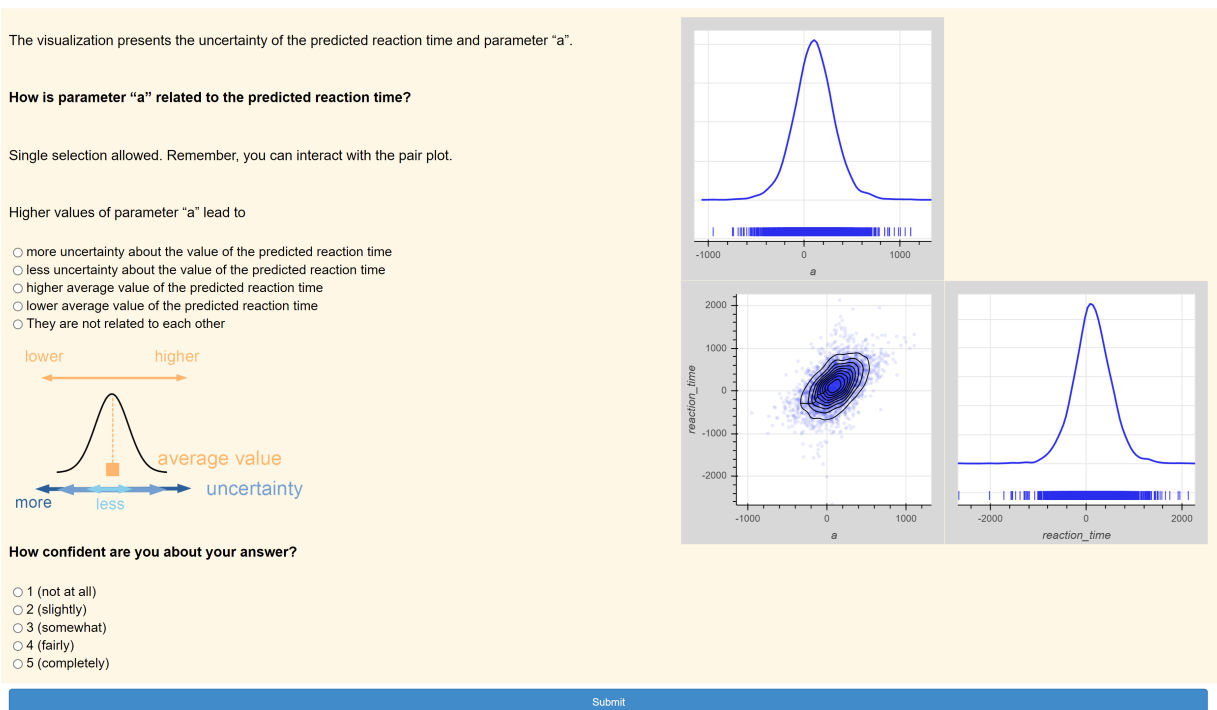


Figure C.14: Model 3 - Task t14 (T2).

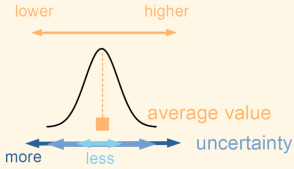
The visualization presents the uncertainty of the predicted reaction time and parameter "b".

How is parameter "b" related to the predicted reaction time?

Single selection allowed. Remember, you can interact with the pair plot.

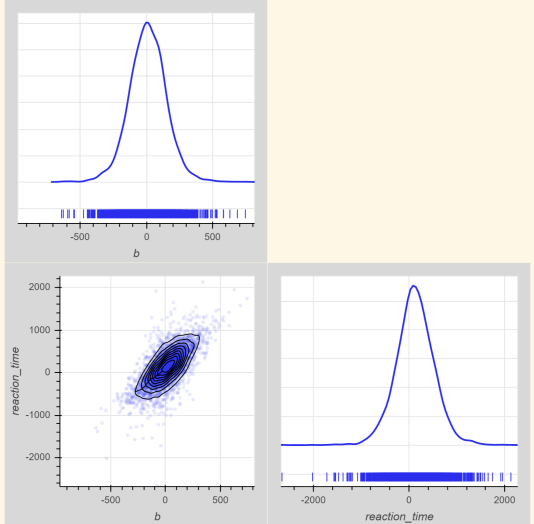
Higher values of parameter "b" lead to

- more uncertainty about the value of the predicted reaction time
- less uncertainty about the value of the predicted reaction time
- higher average value of the predicted reaction time
- lower average value of the predicted reaction time
- They are not related to each other



How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)



Submit

Figure C.15: Model 3 - Task t15 (T2).

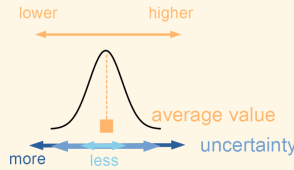
The visualization presents the uncertainty of the predicted reaction time and parameter "c".

How is parameter "c" related to the predicted reaction time?

Single selection allowed. Remember, you can interact with the pair plot.

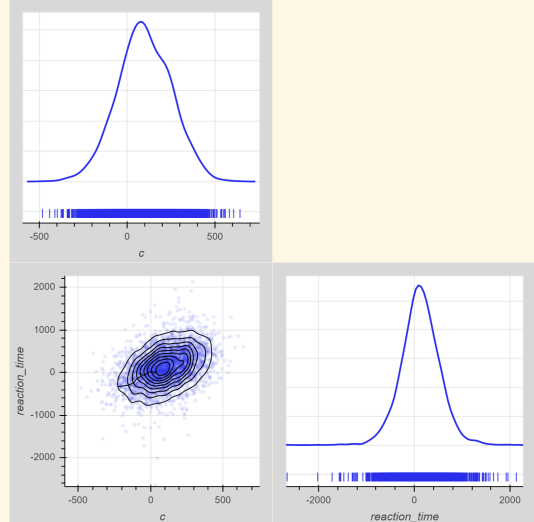
Higher values of parameter "c" lead to

- more uncertainty about the value of the predicted reaction time
- less uncertainty about the value of the predicted reaction time
- higher average value of the predicted reaction time
- lower average value of the predicted reaction time
- They are not related to each other



How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)



Submit

Figure C.16: Model 3 - Task t16 (T2).

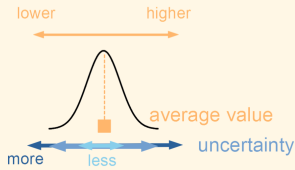
The visualization presents the uncertainty of the predicted reaction time and parameter "d".

How is parameter "d" related to the predicted reaction time?

Single selection allowed. Remember, you can interact with the pair plot.

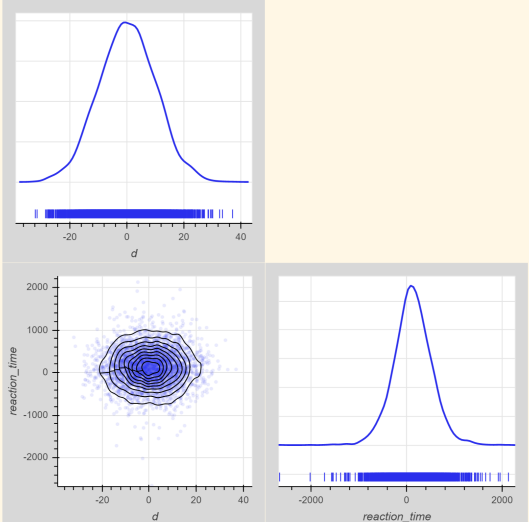
Higher values of parameter "d" lead to

- more uncertainty about the value of the predicted reaction time
- less uncertainty about the value of the predicted reaction time
- higher average value of the predicted reaction time
- lower average value of the predicted reaction time
- They are not related to each other



How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)



Submit

Figure C.17: **Model 3** - Task t17 (T2).

The visualization presents the uncertainty of the predicted reaction times and "a" and "c" parameters.

If the variable of predicted reaction times and the parameters "a" and "c" lie on a graph, what do you think is the structure of this graph?

Single selection allowed. Remember, you can interact with the pair plot.

(A) "a" sets the average value of reaction times and "c" sets the average value of "a"
 (B) "c" sets the average value of reaction times and "a" sets the average value of "c"
 (C) "a" sets the average value of reaction times and "c" doesn't affect reaction times
 (D) "c" sets the average value of reaction times and "a" doesn't affect reaction times
 There is no effect

(A)

(B)

(C)

(D)

reaction time

reaction time

reaction time

reaction time

How confident are you about your answer?

1 (not at all)
 2 (slightly)
 3 (somewhat)
 4 (fairly)
 5 (completely)

Submit

Figure C.18: Model 3 - Task t18 (T3).

The visualization presents the uncertainty of the predicted reaction times and "a" and "b" parameters. The dropdown menu on the right enables selection of the day of driving.

Select Day
 day
 0

How would you describe the effect of the "a" and "b" parameters and the day dimension on the predicted reaction times?

Single selection allowed. Remember, you can interact with the pair plot.

The average value of the predicted reaction times is set by:

- (A) $a + b \cdot \text{day}$
- (B) $b \cdot \text{day}$ and a has no effect
- (C) $a + b$ and day has no effect
- (D) $b + a \cdot \text{day}$
- (E) $a \cdot \text{day}$ and b has no effect
- (F) $a - b$ and day has no effect
- There is no effect

reaction time

(A)

reaction time

(B)

reaction time

(C)

reaction time

(D)

reaction time

(E)

reaction time

(F)

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Submit

Figure C.19: Model 3 - Task t19 (T3).

Appendix D

User Study 2

D.1 Generation of Synthetic Observations for the Insomnia-Anxiety-Tiredness Problem

Synthetic observations were generated for the insomnia-anxiety-tiredness problem based on the common cause model presented in Fig. D.1.

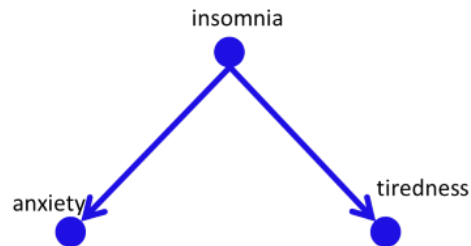


Figure D.1: Common cause model of the insomnia-anxiety-tiredness problem considered in the user study.

The Python code for generating the synthetic observations is shown below and can be found in [Taka, 2023c].

```
import numpy as np
import random

random.seed(10)
N = 7
insomnia = np.random.normal(size = N)
anxiety = np.random.normal(loc = 3*insomnia+1, size = N)
tiredness = np.random.normal(loc = 2.*insomnia, size = N)
```

The causal structure of the model is encoded in a Python dictionary as following:

```
causal_dag = {"insomnia":[],
```

```
"anxiety": ["insomnia"],
"tiredness": ["insomnia"]}
```

D.2 Participants' Training

The training part of the user study comprised of six pages presented to participants through the web interface of the user study. It consisted mainly of reading and comprehension and demonstration of the vicausi tool. The training was similar in all conditions; only the presented visualization mode of the vicausi tool varied among the static (SG), interaction (IG), and animated (AG) group. Fig. D.2-D.7 present the content of these six pages as presented in the SG.

Page 1 out of 6:

What will you do in the user study?

In this user study you will need to identify explanations of how some data is generated. These explanations will describe which variables cause (and affect) other variables. We will call these explanations *causal models*.

How is a causal model represented visually?

A causal model is represented by a directed graph (DAG). See the DAG of causal model 1 on the right. The graph shows the causal relations of the variables. Each variable is represented by a node (circle). When a variable causes (and affects) another, an arrow is drawn from that variable (the cause) to the variable that it causes (and affects) (the effect).

Causal model 1 explains the mechanism that generates insomnia, anxiety, and tiredness to patients. Having anxiety brings (causes) insomnia to a patient and having insomnia brings tiredness. Having tiredness does *not* cause anxiety, and having anxiety does not cause tiredness *directly* but *indirectly* through insomnia.

What do we investigate through this user study?

In this user study, we investigate if people can infer these explanations of how data is generated (the causal models) by observing visualizations of the data.

Click 'Next' to see how we can visualize the data.

Please confirm the following statement to move on:

I have read this page carefully

Next

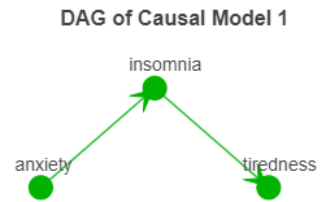


Figure D.2: Page 1 of training.

D.3 Tasks

Fig. D.8-D.23 present the screenshots of the user study tasks t1-t16.

Page 2 out of 6:

Visualization of data

Let's assume that we have some ratings of patients' insomnia, anxiety, and tiredness levels.

Is it possible to judge which variable causes which other variable based on these ratings?

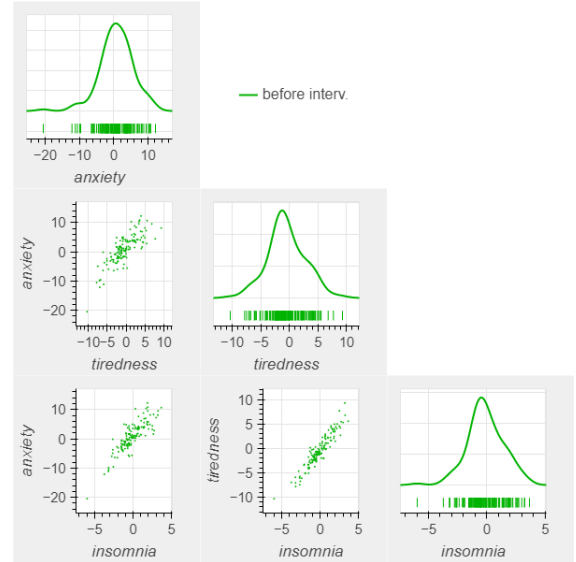
We plot these ratings (given on a continuous scale) in a *scatter plot matrix* on the right (scroll to the right if needed to see it). That's a grid (matrix) of *scatter plots*. Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart.

The *scatter plots* in the scatter plot matrix on the right present the levels of patients' insomnia, anxiety, and tiredness in pairs. For example, in the scatter plot of anxiety-insomnia, each marker point (dot) (x,y) represents the insomnia (x) and anxiety (y) value of a patient.

The diagonal of the scatter plot matrix presents the *KDE plots* of patients' anxiety, tiredness, and insomnia levels. These plots present the probability distribution of each variable as a curved line. The probability distribution expresses the probability of the different possible values of a variable. For example, anxiety might take a value within a range [-22,14] and the KDE plot at the top of the scatter plot matrix tells us how probable each value within this range is. The presented rug plot (vertical lines) below the probability distribution in the KDE plot presents the data-points as vertical lines. The denser the lines in an area of values is, the more probable this area of values is.

Can this data presented in the scatter plot matrix tell us which the mechanism that generated it is (which variable causes which other variable)?

Click 'Next' to see the answer.



Please confirm the following statement to move on:

I have read this page carefully

Next

Figure D.3: Page 2 of training.

Page 3 out of 6:

Can the data tell which the causal model of the data is?

Based on the scatter plots, all variables seem *positively correlated* (increasing one increases the other). Does this tell us which variable causes which other variable?

Let's take the pair of anxiety and insomnia (look at the scatter plot of anxiety-insomnia).

Is the value of *insomnia* increasing because of an increase in the value of *anxiety* or is the value of *anxiety* increasing because of an increase in the values of *insomnia*? Is anxiety that causes (and affects) insomnia or vice versa?

(We assume that the variables in this user study cannot cause each other at the same time, but only one of them might cause the other).

You might be surprised if I tell you that there is a third possible case: none of these variables causes the other. They might just appear correlated for other reasons (how this is possible will be explained in the example that follows).

The *observed data* of anxiety and insomnia alone cannot tell us if and which one causes the other. Based on the *observed data* we can only make assumptions about what the possible data generating mechanisms are, and usually there is more than one possibility. Let's take as an example the two causal models presented on the right of the scatter plot (scroll to the right if needed to see them). Any of these causal models could generate observations like the ones presented in the scatter plot matrix where all variables are correlated to each other.

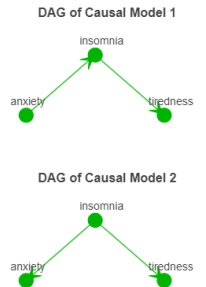
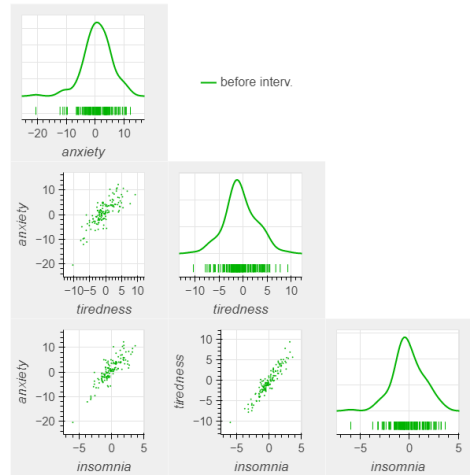
Causal model 1 tells us that any increase in *insomnia* is caused by an increase in *anxiety*, and any increase in *tiredness* is caused by an increase in *insomnia*. Although there is no direct link (arrow) between anxiety and tiredness, anxiety affects tiredness *indirectly* through insomnia (the pipe effect). So, based on this causal model, anxiety and tiredness would appear correlated.

Causal model 2 tells us that any increase in *anxiety* or *tiredness* is caused by an increase in *insomnia*. Although there is no link (arrow) between anxiety and tiredness, they both have a common cause, *insomnia*, which when increases causes an increase to tiredness and anxiety at the same time. So, if we plot the data of tiredness and anxiety in a scatter plot, each will seem to increase when the other increases, although none of them causes the other.

So, both causal models are plausible explanations of the process that generated this data.

Assuming that we know which the possible causal models of some data are, is there a way to tell which one actually generates the data.

Click 'Next' to see how we can do this.



Please confirm the following statement to move on:

I have read this page carefully

Next

Figure D.4: Page 3 of training.

Page 4 out of 6:

Intervention and Types of Intervention

The *observed data* of insomnia, anxiety, and tiredness showed so far occur naturally through a data generating mechanism described possibly by causal model 1 or 2. An experimenter, e.g., a sleep doctor in this context, could make an *intervention* externally on this data generating mechanism, and by collecting the new data that will be generated after the intervention make some inferences about the causal relations of the variables.

There are many possible intervention types an experimenter could apply but we will only study three in this user study. The first is the **atomic intervention** when the sleep doctor forces patients' insomnia, anxiety, or tiredness to be at a specific level through for example some appropriate treatment. The second is the **shift intervention** when the sleep doctor shifts the value of patients' insomnia, anxiety, or tiredness by a fixed amount x . The third is the **variance intervention** when the sleep doctor scales the variance (how far the values are spread out) of a variable by a fixed amount x (x^2 variance).

How can we use the new data generated after an intervention to make inferences about the causal relations of variables?

The main advantage of intervention is that it enables us to identify the variables that an intervened-on variable causes (or affects). A change in a variable-cause would affect only the variables caused (directly or indirectly) by this cause and leave the remaining variables unaffected. By an intervention, we apply controlled changes to the value of a variable and try to identify which variables are affected by it by spotting changes to the other variables. A trivial intervention example would be the removal of items from one's diet to understand which one makes them unhealthy. If their health improves after stopping to eat a specific food type, we can infer that this food was the cause of their health problem.

We can observe which variables are affected by an intervention by visualizing the new data after the intervention.

Click 'Next' to see how we can visualize the new data after an intervention.

Please confirm the following statement to move on:

I have read this page carefully

Next

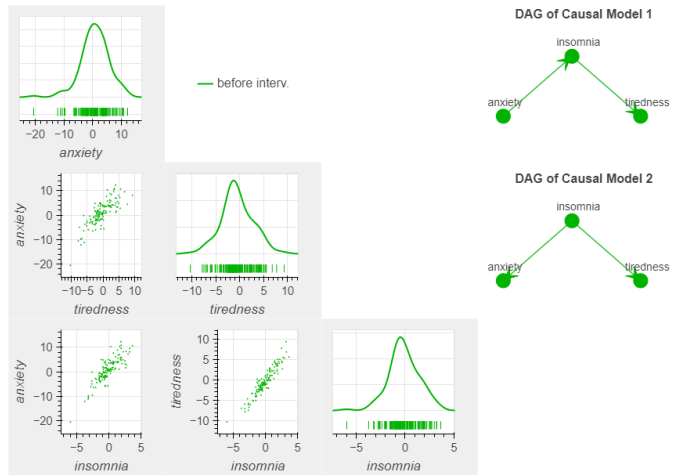


Figure D.5: Page 4 of training.

Page 6 out of 6:

Visualization of the new data after the intervention

Let's assume that the sleep doctor applies an **atomic intervention on insomnia**. Select the atomic intervention on insomnia above the scatter plot matrix on the right. We highlight the node of the variable that receives the intervention in the causal models' DAGs with an orange color. The data drawn in colors based on the given color-bar in the scatter plot matrix on the right presents the new data generated after intervening on patients' insomnia. The data of setting insomnia's level to a range of different values is shown.

What do we observe in the new data after the intervention?

Setting the level of insomnia to increasing values

1. causes an increase in patients' tiredness. There is a shift of tiredness's range of values to higher values in the scatter plot of tiredness-insomnia and tiredness's KDE plot. This means that *insomnia affects tiredness*.
2. does not affect anxiety. Anxiety's range of values remains essentially unaffected after the intervention in the scatter plot of anxiety-insomnia and anxiety's KDE plot (you should expect some noise though because you cannot get the exact same values every time you run an experiment). This means that *anxiety is not affected by insomnia*.
3. makes *tiredness and anxiety appear uncorrelated* after the intervention.

How could these 3 observations be used to tell which the causal model of the data is?

1. Both causal model 1 and 2 confirm the data in that insomnia affects (causes) tiredness (there is an arrow from insomnia to tiredness).
2. Causal model 1 confirms that anxiety does not change after the intervention because anxiety is the cause of (and not caused by) insomnia. In causal model 2 insomnia causes anxiety (there is an arrow from insomnia to anxiety) and we would expect to see anxiety changing with a change in insomnia.
3. Insomnia's value is completely controlled (set) externally in an atomic intervention. This means that any cause of insomnia will no more be able to affect the value of insomnia. In causal model 1, anxiety won't affect insomnia any more after the intervention because insomnia is manipulated externally. For this reason, the correlation of anxiety and insomnia breaks (the link between them breaks and is denoted by a dashed orange arrow in the causal models' DAG). This will make anxiety unable to affect tiredness through insomnia and their correlation will also break. In causal model 2 insomnia would affect both anxiety and tiredness, so we would expect to see a correlation between them after the intervention.

Would a shift or variance intervention result in similar observations?

Select the *shift intervention on insomnia* above the scatter plot matrix on the right. The new data after this intervention suggests the 1. *insomnia affects tiredness*, and 2. *anxiety is not affected by insomnia* but **not** the 3. *tiredness and anxiety appear uncorrelated* after the intervention.

Select the *variance intervention on insomnia* above the scatter plot matrix on the right. The new data after this intervention suggests the 1. *insomnia affects tiredness*, and 2. *anxiety is not affected by insomnia* but **not** the 3. *tiredness and anxiety appear uncorrelated* after the intervention.

The breaking of correlations between variables is an effect encountered only in the atomic interventions. A shift or variance intervention does not set the value of a variable externally but simply transforms its value while this is determined by the causes of the variable. The correlation between anxiety and tiredness does **not** break after these types of intervention like in the case of the atomic intervention (because anxiety keeps causing insomnia after an intervention of this type).

Click 'Next' to see some last instructions.

Please confirm the following statement to move on:

I have read this page carefully

Next

Select an intervention type on a variable to see the data from the intervention.

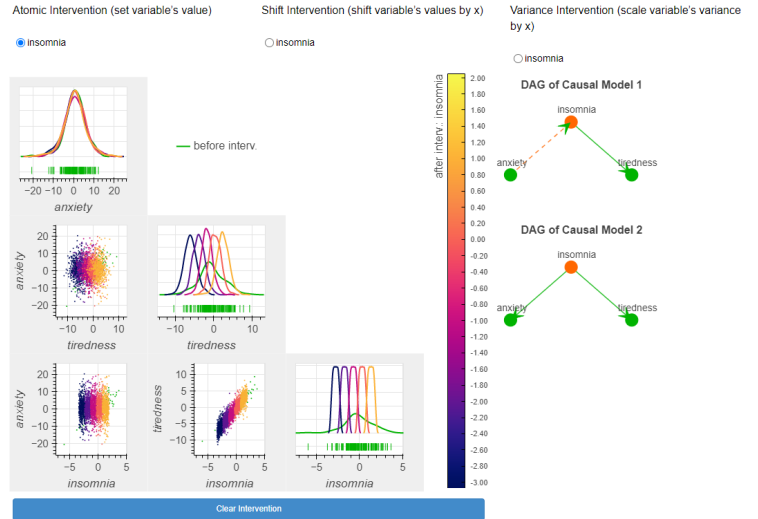


Figure D.6: Page 5 of training.

Page 6 out of 6:

The following two points summarize the main effects of interventions:

1. When a variable affects another, intervening on the variable will cause a change to the variable it affects and will leave the remaining variables unaffected.
2. If a link breaks between two variables after an atomic intervention, these two variables become uncorrelated. The rest variables might become uncorrelated too depending on how they were causally related (between them and with the remaining variables) before the intervention.

This is all you need to know for this user study

Please, pause here and ask any question you might have to the researcher.

There will be two phases in the tasks' part with each phase consisting of tasks of a different type.

When you are ready, click 'Next' to start with the tasks of the first type.

Select an intervention type on a variable to see the data from the intervention.



Please confirm the following statement to move on:

I have read this page carefully

Next

Figure D.7: Page 6 of training.

Question 1 out of 10:

Time is running...

A sleep doctor conducts an **atomic intervention** on patients' **tiredness** to set its level to a specific value.

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of setting patients' tiredness to a range of different levels can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

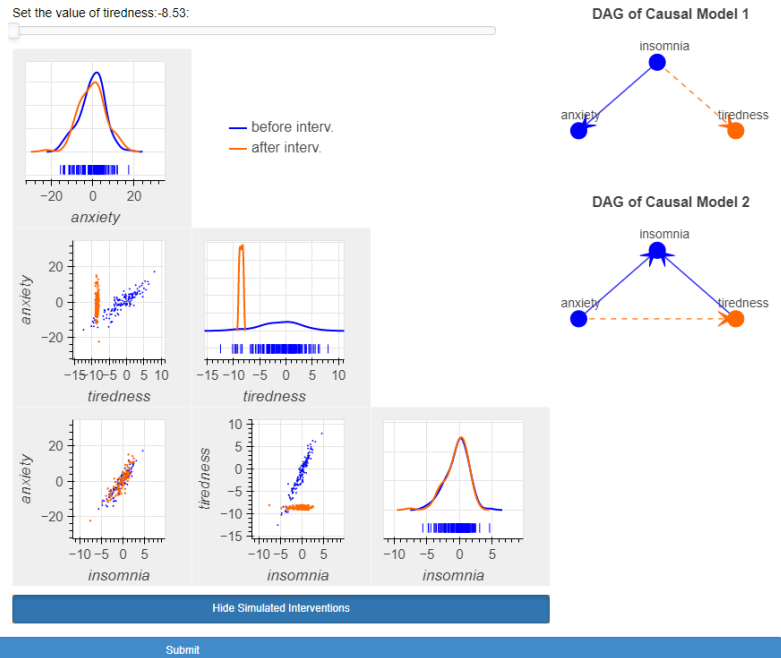


Figure D.8: Task t1 (TT1).

Question 2 out of 10:

Time is running...

A sleep doctor conducts an **atomic intervention** on patients' **insomnia** to set its level to a specific value.

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of setting patients' insomnia to a range of different levels can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

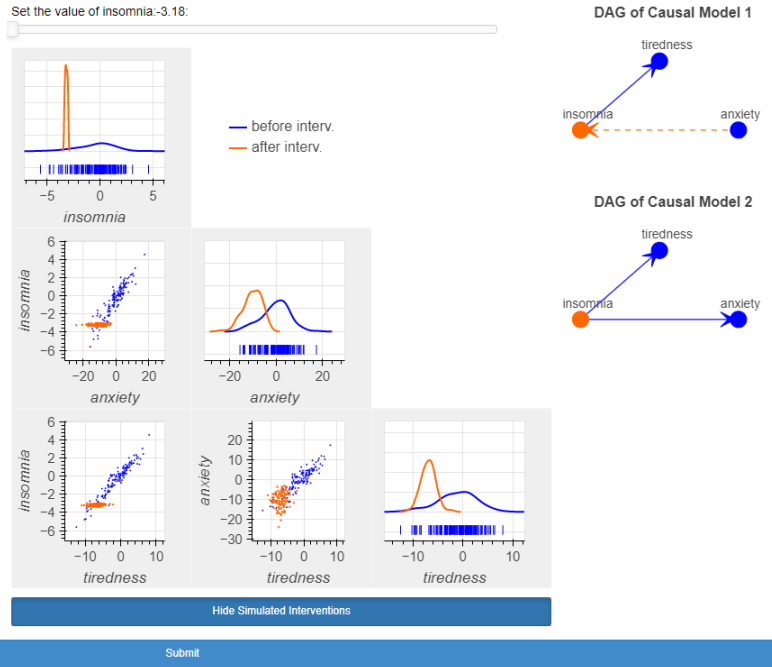


Figure D.9: Task t2 (TT1).

Question 3 out of 10:

Time is running...

A sleep doctor conducts an **atomic intervention** on patients' **anxiety** to set its level to a specific value.

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of setting patients' anxiety to a range of different levels can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

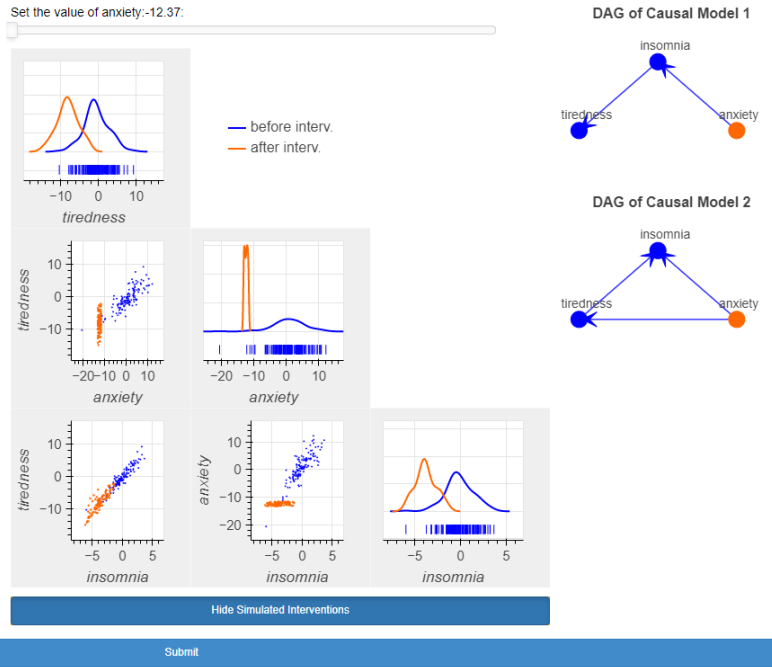


Figure D.10: Task t3 (TT1).

Question 4 out of 10:

Time is running...

A sleep doctor conducts an **atomic intervention** on patients' **tiredness** to set its level to a specific value.

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of setting patients' tiredness to a range of different levels can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

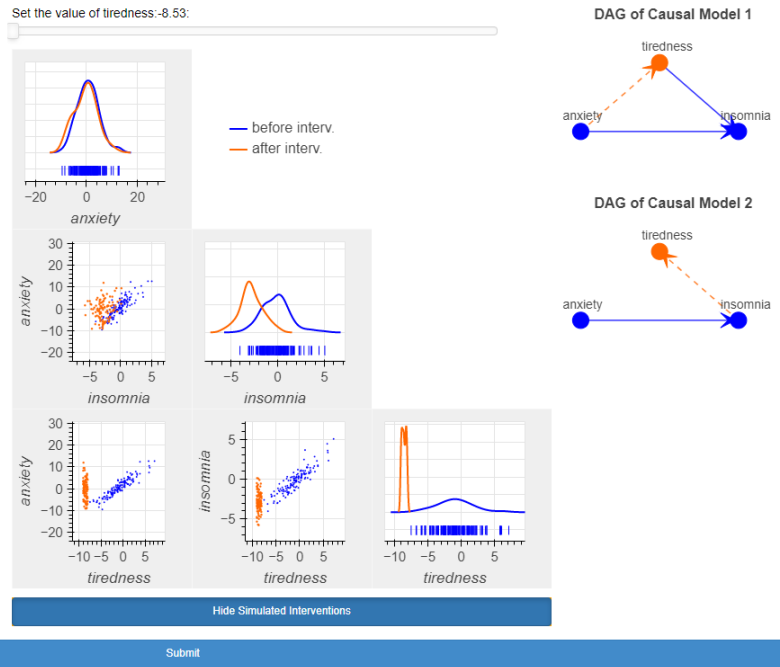


Figure D.11: Task t4 (TT1).

Question 5 out of 10:

Time is running...

A sleep doctor conducts an **atomic intervention** on patients' **anxiety** to set its level to a specific value.

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of setting patients' anxiety to a range of different levels can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

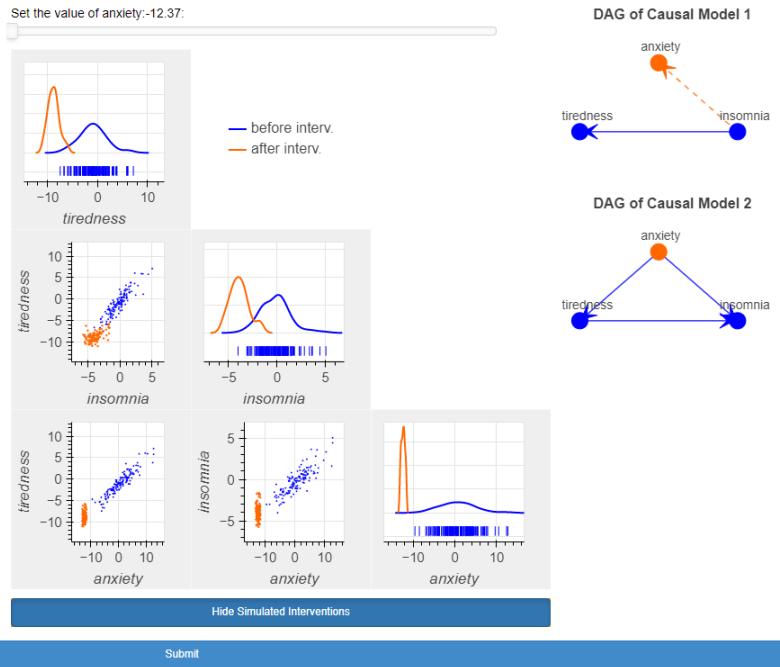


Figure D.12: Task t5 (TT1).

Question 6 out of 10:

Time is running...

A sleep doctor conducts a **shift intervention** on patients' **tiredness** to shift its values by a fixed amount x .

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of shifting patients' tiredness by a range of different values x can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

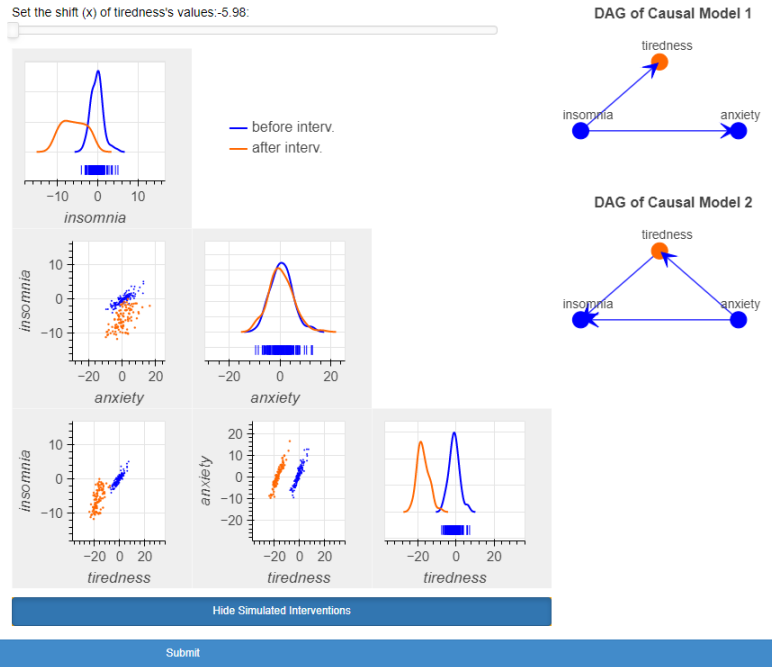


Figure D.13: Task t6 (TT1).

Question 7 out of 10:

Time is running...

A sleep doctor conducts a **shift intervention** on patients' **anxiety** to shift its values by a fixed amount x .

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of shifting patients' anxiety by a range of different values x can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

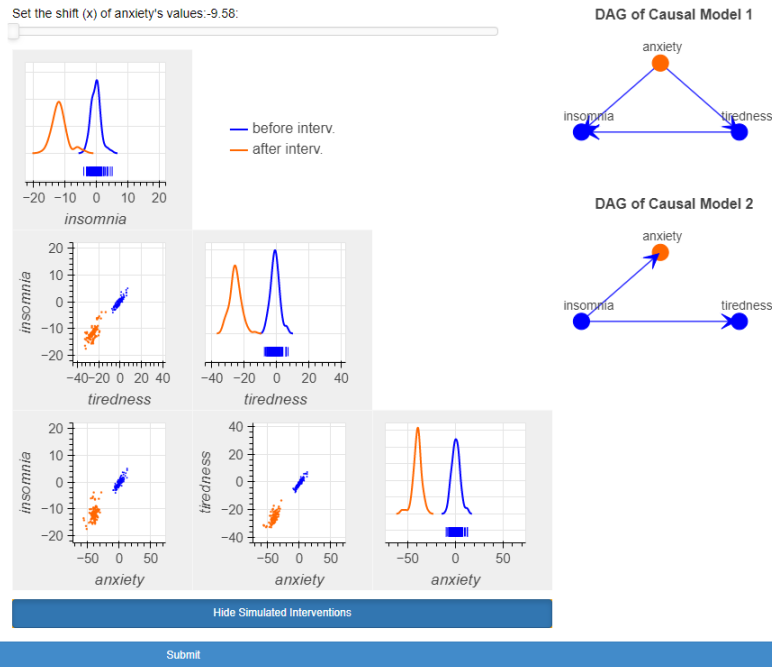


Figure D.14: Task t7 (TT1).

Question 8 out of 10:

Time is running...

A sleep doctor conducts a **variance intervention** on patients' **anxiety** to scale its variance by a fixed amount x .

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of scaling the variance of patients' anxiety by a range of different values x can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

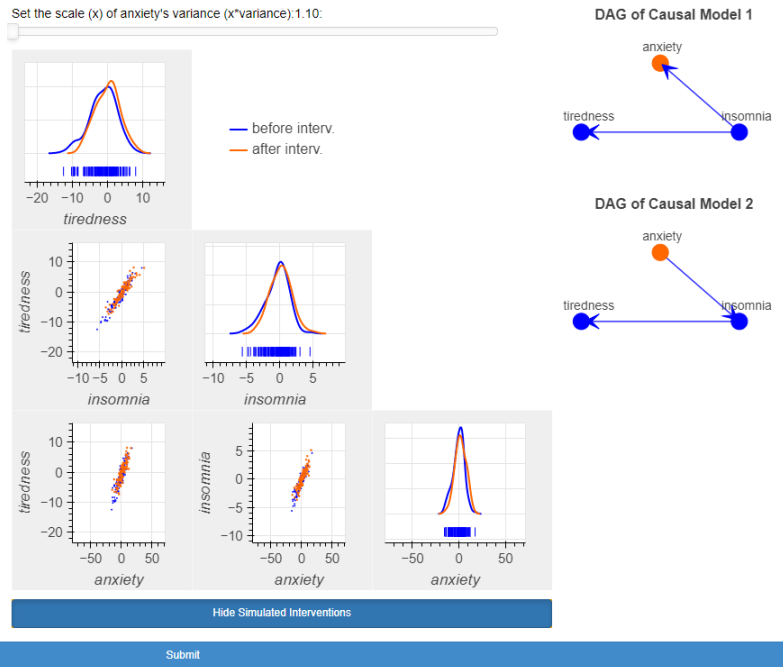


Figure D.15: Task t8 (TT1).

Question 9 out of 10:

Time is running...

A sleep doctor conducts a **variance intervention** on patients' **insomnia** to scale its variance by a fixed amount x .

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of scaling the variance of patients' insomnia by a range of different values x can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

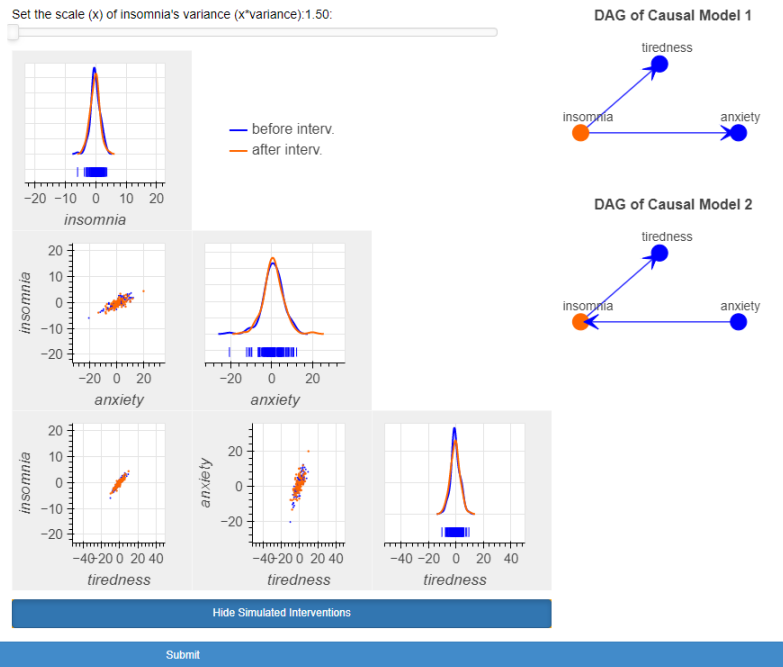


Figure D.16: Task t9 (TT1).

Question 10 out of 10:

Time is running...

A sleep doctor conducts a **variance intervention** on patients' **anxiety** to scale its variance by a fixed amount x .

The scatter plot matrix on the right shows patients' insomnia, anxiety, and tiredness data before and after the intervention. The data of scaling the variance of patients' anxiety by a range of different values x can be viewed.

Which is the causal model that generated this data?

- Causal Model 1
- Causal Model 2
- Both models are possible. This intervention isn't sufficient to judge

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

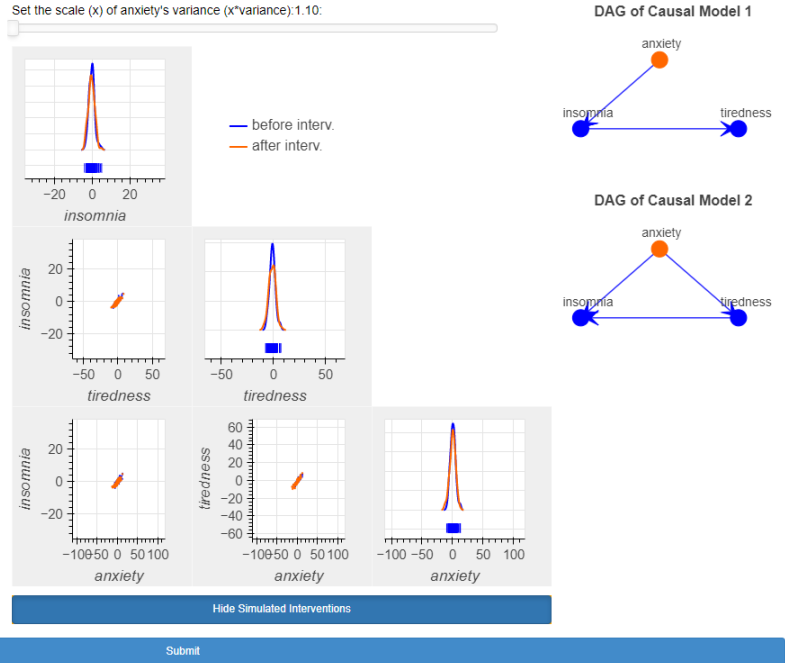


Figure D.17: Task t10 (TT1).

Question 1 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

- Atomic intervention on insomnia
- Atomic intervention on anxiety
- Atomic intervention on tiredness
- None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

- Causal Model 1
- Causal Model 2
- Causal Model 3

How confident are you about your answer?

- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

Atomic Intervention (set variable's value)

- insomnia
- anxiety
- tiredness

Set the value of \$:

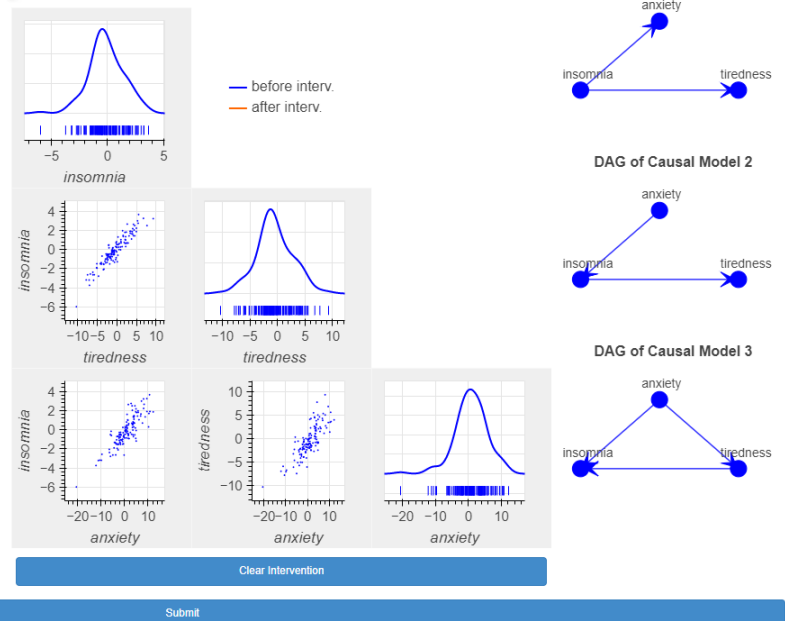


Figure D.18: Task t11 (TT2).

Question 2 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

- Atomic intervention on insomnia
- Atomic intervention on tiredness
- Atomic intervention on anxiety
- None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

- Causal Model 1
- Causal Model 2
- Causal Model 3

How confident are you about your answer?

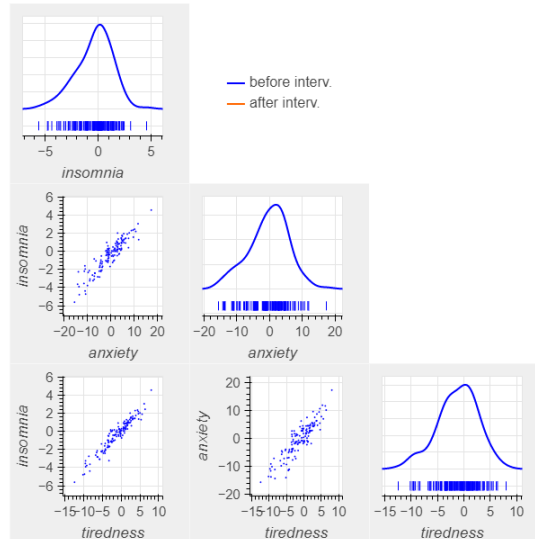
- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

Atomic Intervention (set variable's value)

- insomnia
- tiredness
- anxiety

Set the value of \$:



DAG of Causal Model 1



DAG of Causal Model 2



DAG of Causal Model 3



Clear Intervention

Submit

Figure D.19: Task t12 (TT2).

Question 3 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

- Shift intervention on anxiety
- Shift intervention on insomnia
- Shift intervention on tiredness
- None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

- Causal Model 1
- Causal Model 2
- Causal Model 3

How confident are you about your answer?

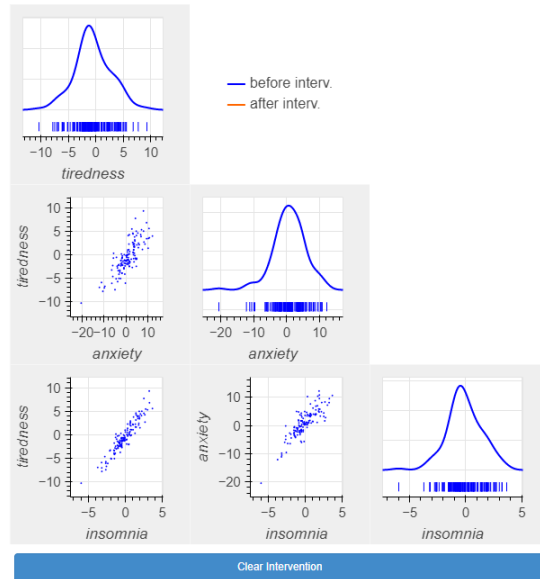
- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

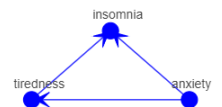
Shift Intervention (shift variable's values by x)

- anxiety
- insomnia
- tiredness

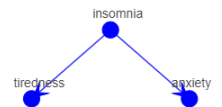
Set the shift (x) of \$s\$'s values:



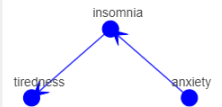
DAG of Causal Model 1



DAG of Causal Model 2



DAG of Causal Model 3



Submit

Figure D.20: Task t13 (TT2).

Question 4 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

- Shift intervention on anxiety
- Shift intervention on tiredness
- Shift intervention on insomnia
- None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

- Causal Model 1
- Causal Model 2
- Causal Model 3

How confident are you about your answer?

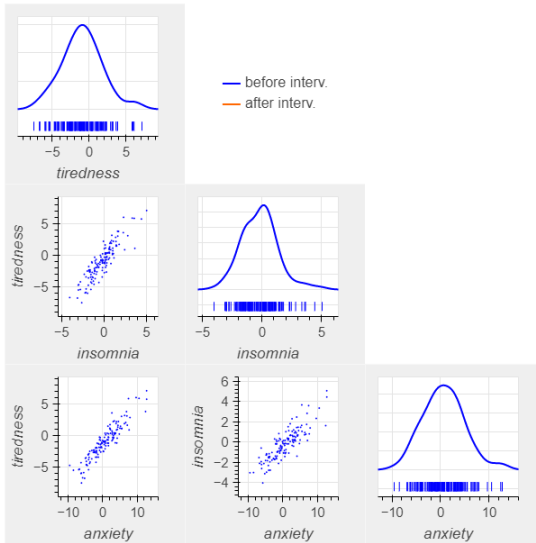
- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

Shift Intervention (shift variable's values by x)

- anxiety
- tiredness
- insomnia

Set the shift (x) of \$s\$'s values:



DAG of Causal Model 1



DAG of Causal Model 2



DAG of Causal Model 3

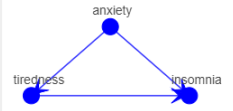


Figure D.21: Task t14 (TT2).

Question 5 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

- Variance intervention on tiredness
- Variance intervention on anxiety
- Variance intervention on insomnia
- None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

- Causal Model 1
- Causal Model 2
- Causal Model 3

How confident are you about your answer?

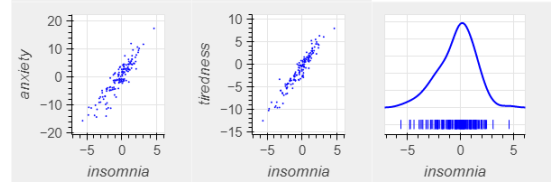
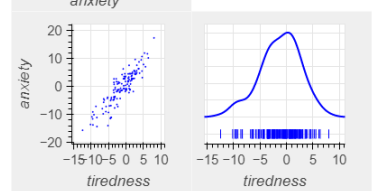
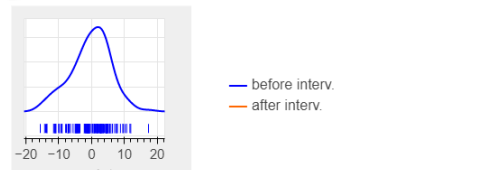
- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

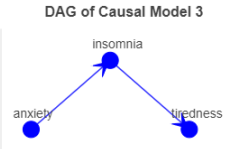
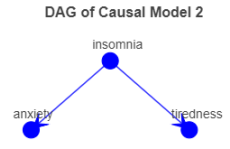
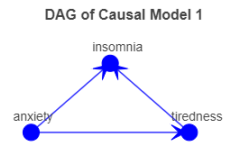
Variance Intervention (scale variable's variance by x)

- tiredness
- anxiety
- insomnia

Set the scale (x) of \$'s variance (x*variance):



Clear Intervention



Submit

Figure D.22: Task t15 (TT2).

Question 6 out of 6:

Time is running...

You want to design and run **one** interventional experiment, which will help you identify the causal model of the data.

Use the visualization on the right to observe the data from the possible interventions.

Which of the interventions below (if any) could be a candidate design for your experiment? Select all interventions (if any) that each alone is sufficient to reveal the causal model.

- Variance intervention on insomnia
- Variance intervention on tiredness
- Variance intervention on anxiety
- None of these interventions is sufficient to identify the causal model

Which is the causal model of the data?

- Causal Model 1
- Causal Model 2
- Causal Model 3

How confident are you about your answer?

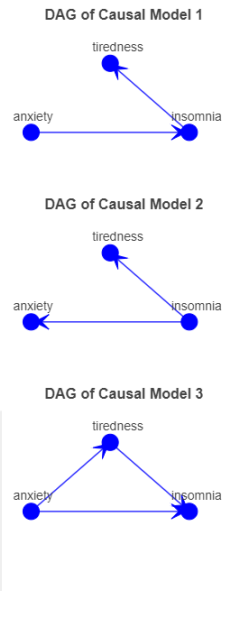
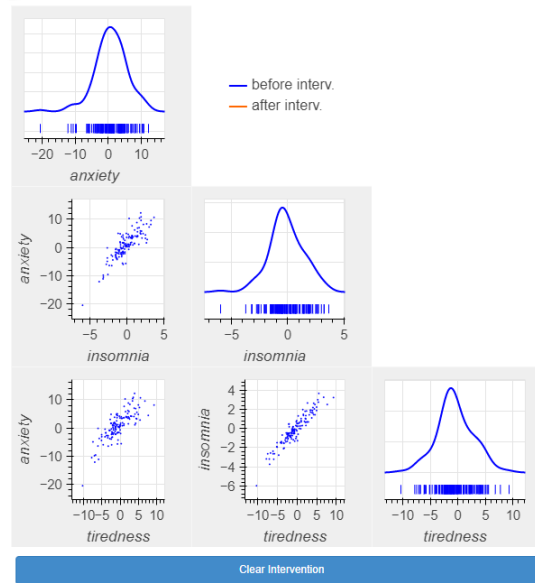
- 1 (not at all)
- 2 (slightly)
- 3 (somewhat)
- 4 (fairly)
- 5 (completely)

Select an intervention type on a variable to see the data from the intervention.

Variance Intervention (scale variable's variance by x)

- insomnia
- tiredness
- anxiety

Set the scale (x) of \$'s variance (x*variance):



Clear Intervention

Submit

Figure D.23: Task t16 (TT2).

Bibliography

- G. B. Alleman. Both aleatory and epistemic uncertainty create risk. https://herdingcats.typepad.com/my_weblog/2013/05/aleatory-and-epistemic-uncertainty-both-create-risk.html#:~:text=Aleatory%20uncertainty%20%2D%20is%20uncertainty%20that,of%20what%20we%20are%20observing., May 2013. Accessed January 1, 2023.
- J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902>.
- J. B. Arnold. Shrinkage and hierarchical models. https://jrnold.github.io/bayesian_notes/shrinkage-and-hierarchical-models.html.
- ArviZ Point Estimate Pairplot. https://arviz-devs.github.io/arviz/examples/plot_pair_point_estimate.html.
- E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1510507113>.
- T. Bayes and R. Price. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. doi: 10.1098/rstl.1763.0053. URL <https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1763.0053>.
- R. A. Becker and W. S. Cleveland. Brushing Scatterplots. *Technometrics*, 29(2):127–142, 1987. doi: 10.1080/00401706.1987.10488204. URL <https://www.tandfonline.com/doi/abs/10.1080/00401706.1987.10488204>.
- G. Belenky, N. J. Wesensten, D. R. Thorne, M. L. Thomas, H. C. Sing, D. P. Redmond, M. B. Russo, and T. J. Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, 12(1):1–12, 2003.
- S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4):389–396, 2005. doi: 10.1037/1082-989X.10.4.389. URL <https://psycnet.apa.org/doiLanding?doi=10.1037%2F1082-989X.10.4.389>.

- Y. Benjamini. Opening the box of a boxplot. *The American Statistician*, 42(4):257–262, 1988. ISSN 00031305. URL <http://www.jstor.org/stable/2685133>.
- J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946. ISSN 00994987. URL <http://www.jstor.org/stable/3002000>.
- M. Betancourt. Towards A Principled Bayesian Workflow. https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html#6_discussion, 2018.
- G. L. Brase. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23(3):369–381, 2009. doi: <https://doi.org/10.1002/acp.1460>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1460>.
- S. Breslav, A. Khan, and K. Hornbæk. Mimic: Visual Analytics of Online Micro-Interactions. In *Proc. AVI '14*, AVI '14, page 245–252, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327756. doi: 10.1145/2598153.2598168. URL <https://doi.org/10.1145/2598153.2598168>.
- F. J. Buchinsky and N. K. Chadha. To P or Not to P: Backing Bayesian Statistics. *Otolaryngology–Head and Neck Surgery*, 157(6):915–918, 2017. doi: 10.1177/0194599817739260. URL <https://doi.org/10.1177/0194599817739260>. PMID: 29192853.
- J. Bulbulia, U. Schjoedt, J. H. Shaver, R. Sosis, and W. J. Wildman. Causal inference in regression: advice to authors. *Religion, Brain & Behavior*, 11(4):353–360, 2021. doi: 10.1080/2153599X.2021.2001259. URL <https://doi.org/10.1080/2153599X.2021.2001259>.
- W. G. Cole. Understanding Bayesian Reasoning via Graphical Displays. In *Proc. CHI '89*, CHI '89, page 381–386, New York, NY, USA, 1989. ACM. ISBN 0897913019. doi: 10.1145/67449.67522. URL <https://doi.org/10.1145/67449.67522>.
- M. U. M. Collective. TheyDrawIt!: An Authoring Tool for Belief-Driven Visualization. <https://medium.com/multiple-views-visualization-research-explained/theydrawit-an-authoring-tool-for-belief-driven-visualization-b3267a001480>, July 2019.
- L. J. Colling and D. Szűcs. Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*, Nov 2018. ISSN 1878-5166. doi: 10.1007/s13164-018-0421-4. URL <https://doi.org/10.1007/s13164-018-0421-4>.
- Colorcet. Colorcet: Collection of perceptually accurate colormaps. <https://colorcet.holoviz.org/>.

- M. Correll and M. L. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12): 2142–2151, Dec 2014. ISSN 2160-9306. doi: 10.1109/TVCG.2014.2346298.
- L. Cosmides and J. Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1):1–73, 1996. ISSN 0010-0277. doi: [https://doi.org/10.1016/0010-0277\(95\)00664-8](https://doi.org/10.1016/0010-0277(95)00664-8). URL <https://www.sciencedirect.com/science/article/pii/0010027795006648>.
- P. Coyle. Why would I ever need Bayesian Statistics? <https://peadarcoyle.medium.com/why-would-i-ever-need-bayesian-statistics-4cf844c4a23a>, October 2018.
- G. Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966. URL <https://doi.org/10.1177/0956797613504966>. PMID: 24220629.
- O.-J. Dahl and K. Nygaard. Simula: An algol-based simulation language. *Commun. ACM*, 9(9):671–678, sep 1966. ISSN 0001-0782. doi: 10.1145/365813.365819. URL <https://doi.org/10.1145/365813.365819>.
- T. N. Dang, P. Murray, J. Aurisano, and A. G. Forbes. ReactionFlow: an interactive visualization tool for causality analysis in biological pathways. In *BMC Proc 9, S6*, 2015. doi: 10.1186/1753-6561-9-S6-S6. URL <https://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-9-S6-S6>.
- S. Depaoli, S. D. Winter, and M. Visser. The importance of prior sensitivity analysis in bayesian statistics: Demonstrations using an interactive shiny app. *Frontiers in Psychology*, 11, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.608045. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.608045>.
- P. Dragicevic, Y. Jansen, A. Sarma, M. Kay, and F. Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300295. URL <https://doi.org/10.1145/3290605.3300295>.
- C. Díaz and F. Inmaculada. Assessing students' difficulties with conditional probability and Bayesian reasoning. *International Electronic Journal of Mathematics Education*, 2(3):128–148, 2007. URL <https://www.iejme.com/article/assessing-students-difficulties-with-conditional-probability-and-bayesian-reasoning>.

- C. R. Ehlschlaeger, A. M. Shortridge, and M. F. Goodchild. Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23(4):387–395, 1997. ISSN 0098-3004. doi: [https://doi.org/10.1016/S0098-3004\(97\)00005-8](https://doi.org/10.1016/S0098-3004(97)00005-8). URL <http://www.sciencedirect.com/science/article/pii/S0098300497000058>. Exploratory Cartographic Visualisation.
- J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull. *Graphviz and Dynagraph — Static and Dynamic Graph Drawing Tools*, pages 127–148. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-642-18638-7. doi: 10.1007/978-3-642-18638-7_6. URL https://doi.org/10.1007/978-3-642-18638-7_6.
- N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Trans. Vis. Comput. Graphics*, 14(6): 1539–1148, 2008. doi: 10.1109/TVCG.2008.153.
- W. W. Esty and J. Banfield. The box-percentile plot. *Journal of Statistical Software*, 8(17):1–14, 2003. doi: 10.18637/jss.v008.i17. URL <https://www.jstatsoft.org/index.php/jss/article/view/v008i17>.
- B. J. Evans. Dynamic display of spatial data-reliability: Does it benefit the map user? *Computers & Geosciences*, 23(4):409–422, 1997. ISSN 0098-3004. doi: [https://doi.org/10.1016/S0098-3004\(97\)00011-3](https://doi.org/10.1016/S0098-3004(97)00011-3). URL <https://www.sciencedirect.com/science/article/pii/S0098300497000113>. Exploratory Cartographic Visualisation.
- J. Faith. Targeted projection pursuit for interactive exploration of high-dimensional data sets. In *2007 11th International Conference Information Visualization (IV '07)*, pages 286–292, July 2007. doi: 10.1109/IV.2007.107.
- E. D. Feigelson and G. J. Babu. *Response by William I. Newman et al.*, pages 161–162. Springer New York, New York, NY, 1992. ISBN 978-1-4613-9290-3. doi: 10.1007/978-1-4613-9290-3_17. URL https://doi.org/10.1007/978-1-4613-9290-3_17.
- M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173718. URL <https://doi.org/10.1145/3173574.3173718>.
- P. Fit. stan.fit.fit. <https://pystan.readthedocs.io/en/latest/reference.html#stan.fit.Fit>.

- M. Frigge, D. C. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50–54, 1989. ISSN 00031305. URL <http://www.jstor.org/stable/2685173>.
- J. Gabry and T. Mahr. bayesplot: Plotting for bayesian models. 2020. R package version 1.7.2. Available online at: <https://mc-stan.org/bayesplot> [Accessed October 14, 2020].
- J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019. doi: 10.1111/rssa.12378. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12378>.
- X. Ge, V. K. Raghu, P. K. Chrysanthis, and P. V. Benos. CausalMGM: an interactive web-based causal discovery tool. *Nucleic Acids Research*, 48(W1):W597–W602, 05 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa350. URL <https://doi.org/10.1093/nar/gkaa350>.
- A. Gelman. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*, 48(3):432–435, 2006. doi: 10.1198/004017005000000661. URL <https://doi.org/10.1198/004017005000000661>.
- A. Gelman. Prior Choice Recommendations). <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>, April 2020.
- A. Gelman. Epistemic and aleatoric uncertainty: The role of context. <https://statmodeling.stat.columbia.edu/2022/02/03/epistemic-and-aleatoric-uncertainty/>, February 2022. Accessed January 30, 2023.
- A. Gelman and D. B. Rubin. A single series from the gibbs sampler provides a false sense of security. *Bayesian Statistics 4 (eds. J. Bernardo et al.)*, pages 625–31, 1992. URL http://stat.columbia.edu/~gelman/research/published/false_sense.pdf.
- A. Gelman, D. Simpson, and M. Betancourt. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10), 2017. ISSN 1099-4300. doi: 10.3390/e19100555. URL <https://www.mdpi.com/1099-4300/19/10/555>.
- G. Gigerenzer. The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making*, 16(3):273–280, 1996. doi: 10.1177/0272989X9601600312. URL <https://doi.org/10.1177/0272989X9601600312>. PMID: 8818126.
- G. Gigerenzer and U. Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995. doi: 10.1037/0033-295x.102.4.684.

- C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL <https://www.frontiersin.org/article/10.3389/fgene.2019.00524>.
- N. D. Goodman and A. Stuhlmüller. The Design and Implementation of Probabilistic Programming Languages. <http://dippl.org>, 2014. Accessed: 2023-3-4.
- N. D. Goodman, V. K. Mansinghka, D. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: A language for generative models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 220–229, Arlington, Virginia, USA, 2008. AUAI Press. ISBN 0974903949.
- S. Greenland and J. Pearl. *Causal Diagrams*, pages 1–10. John Wiley & Sons, Ltd, 2017. ISBN 9781118445112. doi: <https://doi.org/10.1002/9781118445112.stat03732.pub2>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03732.pub2>.
- S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999. ISSN 10443983. URL <http://www.jstor.org/stable/3702180>.
- S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, Apr 2016. ISSN 1573-7284. doi: 10.1007/s10654-016-0149-3. URL <https://doi.org/10.1007/s10654-016-0149-3>.
- M. Greis, J. Hullman, M. Correll, M. Kay, and O. Shaer. Designing for uncertainty in HCI: When does uncertainty help? In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 593–600, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346566. doi: 10.1145/3027063.3027091. URL <https://doi.org/10.1145/3027063.3027091>.
- K. W. Haemer. Range-bar charts. *The American Statistician*, 2(2):23–23, 1948. doi: 10.1080/00031305.1948.10501576. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.1948.10501576>.
- H. Han, K. Lee, and F. Soylu. Simulating outcomes of interventions using a multipurpose simulation program based on the evolutionary causal matrices and markov chain. *Knowledge and Information Systems*, 57(3), dec 2018. ISSN 0219-3116. doi: 10.1007/s10115-017-1151-0. URL <https://doi.org/10.1007/s10115-017-1151-0>.

- J. Heer and G. Robertson. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, 2007. doi: 10.1109/TVCG.2007.70539.
- B. C. Herd and S. Miles. Detecting causal relationships in simulation models using intervention-based counterfactual analysis. *ACM Trans. Intell. Syst. Technol.*, 10(5), sep 2019. ISSN 2157-6904. doi: 10.1145/3322123. URL <https://doi.org/10.1145/3322123>.
- M. Hernán and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 1st edition, 2020. URL <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.
- J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998. doi: 10.1080/00031305.1998.10480559. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559>.
- U. Hoffrage and G. Gigerenzer. Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73:538–40, 1998.
- P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi: 10.1080/01621459.1986.10478354. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354>.
- M. D. Homan and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, Jan. 2014. ISSN 1532-4435.
- J. Hullman. Why Evaluating Uncertainty Visualization is Error Prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, page 143–151, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450348188. doi: 10.1145/2993901.2993919. URL <https://doi.org/10.1145/2993901.2993919>.
- J. Hullman. Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(01):130–139, jan 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2934287.
- J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS ONE*, 10(11):1–25, 11 2015. doi: 10.1371/journal.pone.0142444. URL <https://doi.org/10.1371/journal.pone.0142444>.
- J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha. Imagining replications: Graphical prediction discrete visualizations improve recall estimation of effect uncertainty. *IEEE Transactions*

- on *Visualization and Computer Graphics*, 24(1):446–456, Jan 2018. ISSN 2160-9306. doi: 10.1109/TVCG.2017.2743898.
- J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, Jan 2019. ISSN 2160-9306. doi: 10.1109/TVCG.2018.2864889.
- D. Hume. *An Enquiry Concerning Human Understanding*. Indianapolis/Cambridge: Hackett, reprinted and edited 1993 edition, 1748.
- F. Huszár. Causal Inference 2: Illustrating Interventions via a Toy Example. <https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/>, 2019.
- V. Hyvönen and T. Tolonen. Bayesian inference 2019. https://vioshyvo.github.io/Bayesian_inference/index.html, March 2019.
- D. Ibeling. Causal modeling with probabilistic simulation models, 2018. URL <https://arxiv.org/abs/1807.11139>.
- H. Ibrekk and M. G. Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, 7(4):519–529, 1987. doi: <https://doi.org/10.1111/j.1539-6924.1987.tb00488.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.1987.tb00488.x>.
- C. H. Jackson. Displaying uncertainty with shading. *The American Statistician*, 62(4):340–347, 2008. doi: 10.1198/000313008X370843. URL <https://doi.org/10.1198/000313008X370843>.
- R. G. Jarrett. A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191–193, 04 1979.
- D. G. Jenkins and P. F. Quintana-Ascencio. A solution to minimum sample size for regressions. *PLOS ONE*, 15(2):1–15, 02 2020. doi: 10.1371/journal.pone.0229345. URL <https://doi.org/10.1371/journal.pone.0229345>.
- L. L. Johnson, C. B. Borkowf, and P. A. Shaw. Chapter 21 - hypothesis testing. In J. I. Gallin and F. P. Ognibene, editors, *Principles and Practice of Clinical Research (Third Edition)*, pages 255 – 270. Academic Press, Boston, third edition edition, 2012. ISBN 978-0-12-382167-6. doi: <https://doi.org/10.1016/B978-0-12-382167-6.00021-7>. URL <http://www.sciencedirect.com/science/article/pii/B9780123821676000217>.

- S. Joslyn and J. LeClerc. Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4):308–315, 2013. ISSN 09637214. URL <http://www.jstor.org/stable/44318680>.
- S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(1):126–140, 2012. doi: 10.1037/a0025185. URL <https://psycnet.apa.org/record/2011-18824-001>.
- M. F. Jung, D. Sirkin, T. M. Gür, and M. Steinert. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15*, page 2201–2210, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702479. URL <https://doi.org/10.1145/2702123.2702479>.
- D. Kahneman and S. Frederick. *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, page 49–81. Cambridge University Press, 2002. doi: 10.1017/CBO9780511808098.004.
- D. Kahneman and A. Tversky. *Subjective Probability: A Judgment of Representativeness*, pages 25–48. Springer Netherlands, Dordrecht, 1974. ISBN 978-94-010-2288-0. doi: 10.1007/978-94-010-2288-0_3. URL https://doi.org/10.1007/978-94-010-2288-0_3.
- A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, Jan 2019. ISSN 2160-9306. doi: 10.1109/TVCG.2018.2864909.
- A. Kale, M. Kay, and J. Hullman. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(02):272–282, feb 2021. ISSN 1941-0506. doi: 10.1109/TVCG.2020.3030335.
- A. Kale, Y. Wu, and J. Hullman. Causal support: Modeling causal inferences with visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1150–1160, 2022. doi: 10.1109/TVCG.2021.3114824.
- M. Kay. tidybayes: Tidy data and geoms for Bayesian models. 2020. doi: 10.5281/zenodo.1308151. R package version 2.1.1.9000. Available online at: <http://mjskay.github.io/tidybayes/> [Accessed October 14, 2020].
- M. Kay, T. Kola, J. Hullman, and S. Munson. When(ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *ACM Human Factors in Computing*

- Systems (CHI)*, 2016a. URL <http://idl.cs.washington.edu/papers/when-ish-is-my-bus>.
- M. Kay, G. L. Nelson, and E. B. Hekler. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4521–4532, New York, NY, USA, 2016b. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858465. URL <https://doi.org/10.1145/2858036.2858465>.
- A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- A. Khan, S. Breslav, M. Glueck, and K. Hornbæk. Benefits of visualization in the Mammography Problem. *International Journal of Human-Computer Studies*, 83:94–113, 2015. ISSN 1071-5819. doi: <https://doi.org/10.1016/j.ijhcs.2015.07.001>. URL <https://www.sciencedirect.com/science/article/pii/S1071581915001081>.
- A. Khan, S. Breslav, and K. Hornbæk. Interactive Instruction in Bayesian Inference. *Human-Computer Interaction*, 33(3):207–233, 2018. doi: 10.1080/07370024.2016.1203264. URL <https://doi.org/10.1080/07370024.2016.1203264>.
- Y. Kim, M. Correll, and J. Heer. Designing animated transitions to convey aggregate operations. *Computer Graphics Forum*, 38(3):541–551, 2019. doi: <https://doi.org/10.1111/cgf.13709>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13709>.
- Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1375–1386, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025592. URL <http://doi.acm.org/10.1145/3025453.3025592>.
- Y.-S. Kim, K. Reinecke, and J. Hullman. Data Through Others' Eyes: The Impact of Visualizing Others' Expectations on Visualization Interpretation. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):760–769, Jan 2018. ISSN 1077-2626. doi: 10.1109/TVCG.2017.2745240.
- J. J. Koehler. The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(1):1–17, 1996. doi: 10.1017/S0140525X00041157.

- D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193.
- J. K. Kruschke. Shrinkage in multi-level hierarchical models. <https://doingbayesiandataanalysis.blogspot.com/2012/11/shrinkage-in-multi-level-hierarchical.html>, November 2012a. Accessed January 30, 2023.
- J. K. Kruschke. Graphical model diagrams in Doing Bayesian Data Analysis versus traditional convention. <http://doingbayesiandataanalysis.blogspot.com/2012/05/graphical-model-diagrams-in-doing.html>, 2012b. Accessed March 1, 2023.
- J. K. Kruschke. Diagrams for hierarchical models: New drawing tools. <http://doingbayesiandataanalysis.blogspot.com/2013/10/diagrams-for-hierarchical-models-new.html>, 2013. Accessed March 1, 2023.
- J. K. Kruschke. *Doing Bayesian Data Analysis (Second Edition)*. Academic Press, 2015. ISBN 978-0-12-405888-0. URL <https://sites.google.com/site/doingbayesiandataanalysis/home>.
- J. K. Kruschke. Make model diagrams for human comprehension and ease of programming. <http://doingbayesiandataanalysis.blogspot.com/2018/02/make-model-diagrams-for-human.html>, 2018a. Accessed March 1, 2023.
- J. K. Kruschke. Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1):155–177, 2018b. doi: 10.3758/s13423-017-1272-1. URL <https://doi.org/10.3758/s13423-017-1272-1>.
- J. K. Kruschke and T. M. Liddell. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206, 2018. doi: 10.3758/s13423-016-1221-4. URL <https://doi.org/10.3758/s13423-016-1221-4>.
- T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proc. of IUI '15*, IUI '15, page 126–137, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333061. doi: 10.1145/2678025.2701399. URL <https://doi.org/10.1145/2678025.2701399>.
- T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka. Picture: A probabilistic programming language for scene perception. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4390–4399, 2015. doi: 10.1109/CVPR.2015.7299068.

- R. Kumar, C. Carroll, A. Hartikainen, and O. A. Martin. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *The Journal of Open Source Software*, 2019. doi: 10.21105/joss.01143. URL <http://joss.theoj.org/papers/10.21105/joss.01143>.
- R. Kurzweil. Short probabilistic programming machine-learning code replaces complex programs for computer-vision tasks. <https://www.kurzweilai.net/short-probabilistic-programming-machine-learning-code-replaces-complex-programs-for-computer-vision-tasksf>, April 2015.
- L. Kwock, R. Taylor, Y. Lee, and D. Feng. Matching Visual Saliency to Confidence in Plots of Uncertain Data. *IEEE Transactions on Visualization and Computer Graphics*, 16(06):980–989, nov 2010. ISSN 1941-0506. doi: 10.1109/TVCG.2010.176.
- D. A. Lagnado and S. Sloman. The advantage of timely intervention. *J Exp Psychol Learn Mem Cogn.*, 30(4):856–76, July 2004. doi: 10.1037/0278-7393.30.4.856. URL <https://pubmed.ncbi.nlm.nih.gov/15238029/>.
- B. Lambert. *A Student's Guide to Bayesian Statistics*. SAGE Publications, 2018a. ISBN 9781526418265. URL <https://books.google.gr/books?id=CLZBDwAAQBAJ>.
- B. Lambert. A Student's Guide to Bayesian Statistics: The data for the problem questions in the book. https://benlambertdotcom.files.wordpress.com/2018/08/all_data.zip, 2018b.
- E. G. Larsen. Causality models: Campbell, Rubin and Pearl. <https://erikgahner.dk/2021/causality-models-campbell-rubin-and-pearl/>, 2021.
- F. Lattimore and D. Rohde. Replacing the do-calculus with bayes rule. <https://arxiv.org/abs/1906.07125>, 2019a.
- F. Lattimore and D. Rohde. Causal inference with bayes rule. <https://arxiv.org/abs/1910.01510>, 2019b.
- D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009. doi: 10.1002/sim.3680. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3680>.
- E. J. Ma. Reasoning about Shapes and Probability Distributions. <https://ericmj1.github.io/blog/2019/5/29/reasoning-about-shapes-and-probability-distributions/>, 2019.
- D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018. doi: <https://doi.org/10.1111/phc3.12470>. URL <https://doi.org/10.1111/phc3.12470>.

- [//compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12470](https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12470).
e12470 10.1111/phc3.12470.
- A. R. Martin and M. O. Ward. High Dimensional Brushing for Interactive Exploration of Multivariate Data. In *Proc. VIS '95, VIS '95*, page 271, USA, 1995. IEEE Computer Society. ISBN 0818671874.
- O. Martin. *Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ, 2nd Edition*. Packt Publishing, 2018. ISBN 9781789349665. URL <https://books.google.co.uk/books?id=1Z2BDwAAQBAJ>.
- J. A. McDonald. INTERACTIVE GRAPHICS FOR DATA ANALYSIS. Ph.d. dissertation, August 1982. URL <https://inspirehep.net/files/447e48e644b510fd6af7379cf9af2e8e>.
- R. McElreath. Statistical Rethinking: A Bayesian Course (with Code Examples in R/Stan/Python/Julia). https://github.com/rmcelreath/stat_rethinking_2020, December 2020a.
- R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition*. CRC Press, 2 edition, 2020b. URL <http://xcelab.net/rm/statistical-rethinking/>.
- R. McElreath. Science Before Statistics: Causal Inference. <https://www.youtube.com/watch?v=KNPYUVmY3NM&t=3531s>, September 2021.
- L. Micallef, P. Dragicevic, and J. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. Comput. Graphics*, 18(12):2536–2545, 2012.
- A. Mosca, A. Ottley, and R. Chang. Does Interaction Improve Bayesian Reasoning with Visualization? In *Proc. CHI '21, CHI '21*. ACM, May 2021. ISBN 9781450380966. doi: 10.1145/3411764.3445176. URL <http://dx.doi.org/10.1145/3411764.3445176>.
- P. MultiTrace. `pymc.backends.base.multitrace`. <https://www.pymc.io/projects/docs/en/stable/api/generated/pymc.backends.base.MultiTrace.html>.
- L. Nadav-Greenberg and S. L. Joslyn. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, 2009. doi: 10.1518/155534309X474460. URL <https://doi.org/10.1518/155534309X474460>.

- C. M. Newton. Graphics: From alpha to omega in data analysis. In P. C. WANG, editor, *Graphical Representation of Multivariate Data*, pages 59–92. Academic Press, 1978. ISBN 978-0-12-734750-9. doi: <https://doi.org/10.1016/B978-0-12-734750-9.50008-3>. URL <https://www.sciencedirect.com/science/article/pii/B9780127347509500083>.
- J. Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted)'. Master's thesis, Statistical Science, Vol. 5, pp. 463–472, 1923.
- Q. V. Nguyen, S. Simoff, Y. Qian, and M. L. Huang. Deep exploration of multidimensional data with linkable scatterplots. In *Proc. VINCI '16*, VINCI '16, page 43–50, New York, NY, USA, 2016. ACM. ISBN 9781450341493. doi: 10.1145/2968220.2968248. URL <https://doi.org/10.1145/2968220.2968248>.
- Q. V. Nguyen, N. Miller, D. Arness, W. Huang, M. L. Huang, and S. Simoff. Evaluation on interactive visualization data with scatterplots. *Visual Informatics*, 4(4):1–10, 2020. ISSN 2468-502X. doi: <https://doi.org/10.1016/j.visinf.2020.09.004>. URL <https://www.sciencedirect.com/science/article/pii/S2468502X20300358>.
- A. Ottley, B. Metevier, P. K. J. Han, and R. Chang. Visually Communicating Bayesian Statistics to Laypersons. Technical report, Tufts University, 2012. URL <http://www.cs.tufts.edu/~remco/publications/2012/Tufts2012-Bayes.pdf>.
- A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. J. Han, and R. Chang. Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE Trans. Vis. Comput. Graphics*, 22(1):529–538, 2016. doi: 10.1109/TVCG.2015.2467758.
- L. Padilla, M. Kay, and J. Hullman. *Uncertainty Visualization*, pages 1–18. John Wiley & Sons, Ltd, 2021a. ISBN 9781118445112. doi: <https://doi.org/10.1002/9781118445112.stat08296>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08296>.
- L. M. K. Padilla, M. Powell, M. Kay, and J. Hullman. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, 11, 2021b. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.579267. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.579267>.
- A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Education, 2002. ISBN 0073660116.

- J. Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993. ISSN 08834237. URL <http://www.jstor.org/stable/2245965>.
- J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444. URL <http://www.jstor.org/stable/2337329>.
- J. Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.
- J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), 2010. doi: doi:10.2202/1557-4679.1203. URL <https://doi.org/10.2202/1557-4679.1203>.
- J. Pearl and D. Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.
- C. Phelan, J. Hullman, M. Kay, and P. Resnick. Some prior(s) experience necessary: Templates for getting started with bayesian analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300709. URL <https://doi.org/10.1145/3290605.3300709>.
- M. Plummer. JAGS Version 4.3.0 user manual. https://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf, June 2017.
- D. Poole and F. Wood. *Probabilistic Programming Languages: Independent Choices and Deterministic Systems*, page 691–712. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL <https://doi.org/10.1145/3501714.3501753>.
- K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing summary statistics and uncertainty. In *Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'10, page 823–832, Chichester, GBR, 2010. The Eurographs Association & John Wiley & Sons, Ltd. doi: 10.1111/j.1467-8659.2009.01677.x. URL <https://doi.org/10.1111/j.1467-8659.2009.01677.x>.
- M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*, 5(2):79–83, 2012.
- J. Quddus. The Future of Artificial Intelligence Part 1 – Probabilistic Programming Languages. <https://methods.co.uk/blog/the-future-of-artificial-intelligence-part-1-probabilistic-programming-languages/>, April 2019.

- A. Rotondi, P. Pedroni, and A. Pievatolo. *Probability, Statistics and Simulation With Application Programs Written in R*. Springer Cham, 2022. ISBN 978-3-031-09429-3. URL <https://doi-org.ezproxy.lib.gla.ac.uk/10.1007/978-3-031-09429-3>.
- D. Rubin. Estimating causal effects of treatments in experimental and observational studies. *ETS Research Bulletin Series*, 1972(2):i–31, 1972. doi: <https://doi.org/10.1002/j.2333-8504.1972.tb00631.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1972.tb00631.x>.
- D. B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981. doi: 10.3102/10769986006004377. URL <https://doi.org/10.3102/10769986006004377>.
- J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.
- K. Sankaran and S. W. Holmes. Interactive visualization of hierarchically structured data. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 27 3:553–563, 2018.
- A. Sarma and M. Kay. Prior setting in practice: Strategies and rationales used in choosing prior distributions for bayesian analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376377. URL <https://doi.org/10.1145/3313831.3376377>.
- J. Sekhon. The oxford handbook of political methodology. <http://sekhon.berkeley.edu/papers/SekhonOxfordHandbook.pdf>, 2007.
- J. Sekhon. 271 The Neyman—Rubin Model of Causal Inference and Estimation Via Matching Methods. In *The Oxford Handbook of Political Methodology*. Oxford University Press, 08 2008. ISBN 9780199286546. doi: 10.1093/oxfordhb/9780199286546.003.0011. URL <https://doi.org/10.1093/oxfordhb/9780199286546.003.0011>.
- B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi: 10.1109/VL.1996.545307.
- P. E. Shrout and J. L. Rodgers. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, 69(1):487–510, 2018. doi: 10.1146/annurev-psych-122216-011845. URL <https://doi.org/10.1146/annurev-psych-122216-011845>. PMID: 29300688.

- G. L. Silva, P. Soares, S. Marques, M. I. Dias, M. M. Oliveira, and J. G. Borges. A bayesian modelling of wildfires in portugal. In J.-P. Bourguignon, R. Jeltsch, A. A. Pinto, and M. Viana, editors, *Dynamics, Games and Science*, pages 723–733, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16118-1.
- B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on statistics and applied probability (Series). London ; New York : Chapman and Hall, 1st edition, 1986. URL <https://adams.marmot.org/Record/.b17636802>.
- E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951. ISSN 00359246. URL <http://www.jstor.org/stable/2984065>.
- S. Sinharay and H. S. Stern. Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111(1):209 – 221, 2003. ISSN 0378-3758. doi: [https://doi.org/10.1016/S0378-3758\(02\)00303-8](https://doi.org/10.1016/S0378-3758(02)00303-8). URL <http://www.sciencedirect.com/science/article/pii/S0378375802003038>. Special issue I: Model Selection, Model Diagnostics, Empirical Bayesian and Hierarchical Bayesian.
- O. Sofrygin, R. Neugebauer, and M. J. van der Laan. Conducting simulations in causal inference with networks-based structural equation models, 2017. URL <https://arxiv.org/abs/1705.10376>.
- M. E. Spear. *Charting Statistics*. McGraw-Hill, 1952.
- D. Speegle and B. Clair. *Probability, Statistics, and Data : A Fresh Approach Using R*. CRC Press LLC, 2021. ISBN 9781000504514. URL https://mathstat.slu.edu/~speegle/_book/.
- D. Spiegelhalter and K. Rice. Bayesian statistics. *Scholarpedia*, 4(8):5230, 2009. doi: 10.4249/scholarpedia.5230. revision #185711.
- D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. WinBUGS Version 2.0 Users Manual., 2003. Available online at: <https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf> [Accessed February 17, 2022].
- D. J. Spiegelhalter and H. Riesch. Don’t know, can’t know: embracing deeper uncertainties when analysing risks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956):4730–4750, 2011. doi: 10.1098/rsta.2011.0163. URL <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0163>.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

- Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. <https://mc-stan.org>.
- Stan Development Team. shinystan: Interactive visual and numerical diagnostics and posterior analysis for bayesian models. 2017. R package version 2.5.0. Available online at: <http://mc-stan.org/shinystan/> [Accessed October 14, 2020].
- Stanford Encyclopedia of Philosophy. Causal models. <https://plato.stanford.edu/entries/causal-models/>, 2018.
- S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637. URL <https://doi.org/10.1177/1745691616658637>.
- W. A. Stock and J. T. Behrens. Box, line, and midgap plots: Effects of display characteristics on the accuracy and bias of estimates of whisker length. *Journal of Educational Statistics*, 16(1):1–20, 1991. ISSN 03629791. URL <http://www.jstor.org/stable/1165096>.
- E. Taka. Interactive Visualizations of Probabilistic Models. <https://github.com/evdoxiataka/ipme>, 2020a.
- E. Taka. Interactive Pair Plot. <https://github.com/evdoxiataka/ipme/releases/tag/ipp>, 2020b.
- E. Taka. Talk at PyMCon 2020: Automatic transformation of Bayesian probabilistic models into interactive visualizations. <https://www.youtube.com/watch?v=\=2hadiSJRAJI>, 2020c.
- E. Taka. Thesis examples. https://github.com/evdoxiataka/thesis_examples, February 2023a. Accessed March 1, 2023.
- E. Taka. Data and Analysis Code of User Study. https://github.com/evdoxiataka/simulated_interventions_study_analysis, 2023b.
- E. Taka. Probabilistic Simulator of Interventions. https://github.com/evdoxiataka/simulated_interventions_pipeline, 2023c.
- E. Taka. Visualizer of Causal Assumptions and Uncertainty-Aware Simulations of Interventions. <https://github.com/evdoxiataka/vicausi>, 2023d.
- E. Taka, S. Stein, and J. H. Williamson. Increasing interpretability of Bayesian probabilistic programming models through interactive representations. *Front. Comput. Sci.*, 2:52, 2020. ISSN 2624-9898. doi: 10.3389/fcomp.2020.567344. URL <https://www.frontiersin.org/article/10.3389/fcomp.2020.567344>.

- E. Taka, S. Stein, and J. H. Williamson. Does interacting help users better understand the structure of probabilistic models?, 2022. February 17, 2022. Distributed by University of Glasgow Enlighten Repository. <http://dx.doi.org/10.5525/gla.researchdata.1248>.
- E. Taka, S. Stein, and J. H. Williamson. Does interactive conditioning help users better understand the structure of probabilistic models? *IEEE Transactions on Visualization and Computer Graphics*, pages 1–12, 2023. doi: 10.1109/TVCG.2022.3231967.
- J. Textor and M. Gilthorpe. The Table 2 Fallacy. <http://dagitty.net/learn/graphs/table2-fallacy.html>.
- J. Tian and J. Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence. ISBN 0262511290.
- D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.
- J. Tsai, S. Miller, and A. Kirlik. Interactive Visualizations to Improve Bayesian Reasoning. *Proc. Hum. Factors Ergonom. Society Annu. Meeting*, 55:385–389, 09 2011. doi: 10.1177/1071181311551079.
- J. W. Tukey. *Exploratory Data Analysis*. AddisonWesley, 1977.
- A. Tversky and D. Kahneman. Belief in the law of small numbers. *Psychological Bulletin*, 76(2): 105–110, 1971. doi: 10.1037/h0031322. URL <https://psycnet.apa.org/record/1972-01934-001>.
- A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973. doi: 10.1016/0010-0285(73)90033-9. URL [https://doi.org/10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9).
- A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL <https://www.science.org/doi/abs/10.1126/science.185.4157.1124>.
- B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002. ISSN 1071-5819. doi: <https://doi.org/10.1006/ijhc.2002.1017>. URL <https://www.sciencedirect.com/science/article/pii/S1071581902910177>.
- B. Victor. Simulation as a Practical Tool. <http://worrydream.com/\#\!/SimulationAsAPracticalTool>, 2009.

- B. Victor. Explorable Explanations. <http://worrydream.com/ExplorableExplanations/>, 2011a.
- B. Victor. Up and Down the Ladder of Abstraction. <http://worrydream.com/\#!/LadderOfAbstraction>, 2011b.
- B. Victor. Scrubbing Calculator. <http://worrydream.com/\#!/ScrubbingCalculator>, 2011c.
- J. Wang and K. Mueller. The visual causality analyst: An interactive interface for causal reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 22:1–1, 11 2015. doi: 10.1109/TVCG.2015.2467931.
- J. Wang and K. Mueller. Visual causality analysis made practical. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 151–161, 2017. doi: 10.1109/VAST.2017.8585647.
- C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, Amsterdam, 3 edition, 2012. ISBN 978-0-12-381464-7. URL <http://www.sciencedirect.com/science/book/9780123814647>.
- D. Westreich and S. Greenland. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*, 177(4):292–298, 01 2013. ISSN 0002-9262. doi: 10.1093/aje/kws412. URL <https://doi.org/10.1093/aje/kws412>.
- Wikipedia:ReplicationCrisis. Reprication crisis. https://en.wikipedia.org/wiki/Replication_crisis. Accessed February 1, 2023.
- L. Wilkinson. Dot plots. *The American Statistician*, 53(3):276–281, 1999. doi: 10.1080/00031305.1999.10474474. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474474>.
- J. H. Williamson. arviz_json. https://github.com/johnhw/arviz_json, 2019.
- S. Witty, A. Lew, D. Jensen, and V. Mansinghka. Bayesian causal inference via probabilistic program synthesis, 2019. URL <https://arxiv.org/abs/1910.14124>.
- S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921. URL <https://naldc.nal.usda.gov/catalog/IND43966364>.
- S. Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215, 1934.

- X. Xie, F. Du, and Y. Wu. A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications. *IEEE transactions on visualization and computer graphics*, PP, October 2020. ISSN 1077-2626. doi: 10.1109/tvcg.2020.3028957. URL <https://doi.org/10.1109/TVCG.2020.3028957>.
- C.-H. E. Yen, A. Parameswaran, and W.-T. Fu. An exploratory user study of visual causality analysis. *Computer Graphics Forum*, 38(3):173–184, 2019. doi: <https://doi.org/10.1111/cgf.13680>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13680>.
- M. Yi. A Complete Guide to Violin Plots. <https://chartio.com/learn/charts/violin-plot-complete-guide/>.