



UNIVERSITY OF AMSTERDAM

## UvA-DARE (Digital Academic Repository)

### Bayes factor model comparison for psychological science

Gronau, Q.F.

**Publication date**

2021

**Document Version**

Final published version

[Link to publication](#)

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

Gronau, Q. F. (2021). *Bayes factor model comparison for psychological science*.

**General rights**

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

**Disclaimer/Complaints regulations**

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

# Bayes Factor Model Comparison for Psychological Science

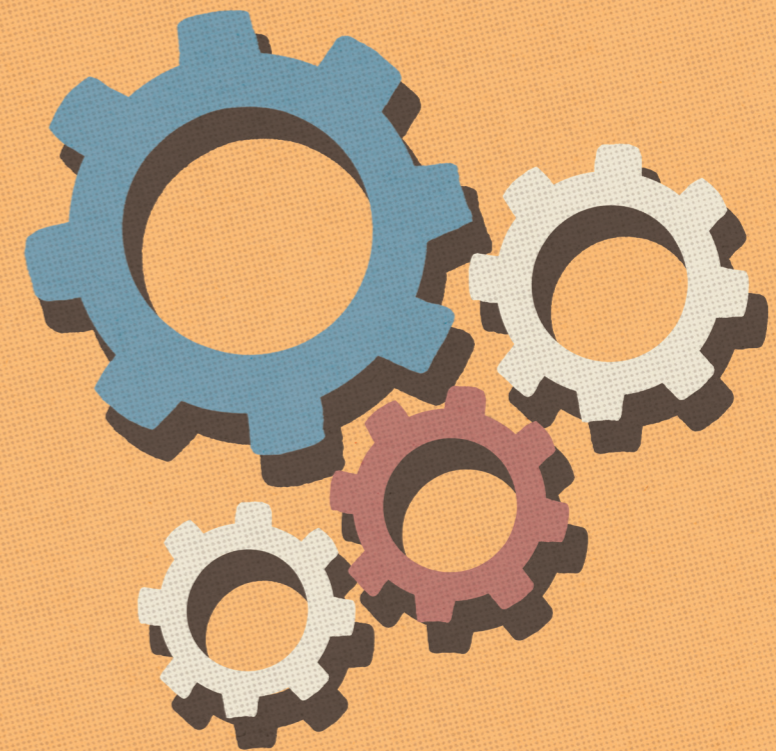
In this dissertation it is argued that rival scientific models should be compared by treating them as competing forecasters and assessing their relative predictive adequacy using the Bayes factor. The first part of the dissertation is concerned with bridge sampling, a computational procedure for estimating the marginal likelihood – the key quantity for computing Bayes factors. The second part of the dissertation is concerned with Bayesian methods for meta-analyzing a set of studies. One central concept of this part is the idea to combine several forecasters using Bayesian model averaging (BMA). The third part of the dissertation introduces Bayesian approaches to a number of standard statistical tests. A central idea of this part is the incorporation of prior knowledge into the analyses to make the models' forecasts more precise.

Quentin F. Gronau

Bayes Factor Model Comparison for Psychological Science

Quentin F. Gronau

# Bayes Factor Model Comparison for Psychological Science



Quentin F. Gronau

# **Bayes Factor Model Comparison for Psychological Science**

Quentin Frederik Gronau

ISBN 978-94-6421-315-7

This publication is typeset in L<sup>A</sup>T<sub>E</sub>X using the Memoir class

Printed by Ipskamp Printing B.V., Enschede

Cover by Viktor Beekman, @viktordepictor

Copyright © 2020 by Quentin Frederik Gronau

All rights reserved

The research of this doctoral thesis received financial assistance from the Netherlands Organisation for Scientific Research (NWO; 406.16.528).

# Bayes Factor Model Comparison for Psychological Science

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

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

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

in het openbaar te verdedigen in de Aula der Universiteit

op vrijdag 25 juni 2021, te 11.00 uur

door Quentin Frederik Gronau

geboren te Frankfurt am Main

## Promotiecommissie

Promotor:	Prof. dr. E. M. Wagenmakers	Universiteit van Amsterdam
Copromotores:	Dr. D. Matzke	Universiteit van Amsterdam
	Dr. M. Marsman	Universiteit van Amsterdam
Overige leden:	Prof. dr. M. D. Lee	University of California Irvine
	Prof. dr. A. J. Heathcote	University of Tasmania
	Dr. H. Singmann	University College London
	Prof. dr. D. Borsboom	Universiteit van Amsterdam
	Prof. dr. H. L. J. van der Maas	Universiteit van Amsterdam
	Dr. C. E. Stevenson	Universiteit van Amsterdam
	Dr. L. J. Waldorp	Universiteit van Amsterdam
Faculteit:	Faculteit der Maatschappij- en Gedragwetenschappen	

*Für meine Familie*



---

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	A Competition Between Forecasters . . . . .	1
1.2	Treating Scientific Models as Forecasters . . . . .	3
1.3	Chapter Outline . . . . .	7
<b>I</b>	<b>Bridge Sampling</b>	<b>11</b>
<b>2</b>	<b>A Tutorial on Bridge Sampling</b>	<b>13</b>
2.1	Introduction . . . . .	13
2.2	Four Sampling Methods to Approximate the Marginal Likelihood .	16
2.3	Case Study: Bridge Sampling for Reinforcement Learning Models .	35
2.4	Discussion . . . . .	44
2.A	The Bridge Sampling Estimator as a General Case of Methods 1 – 3	47
2.B	Bridge Sampling Implementation: Avoiding Numerical Issues . . .	49
2.C	Correcting for the Probit Transformation . . . . .	50
2.D	Details on the Application of Bridge Sampling to the Individual- Level EV Model . . . . .	52
2.E	Details on the Application of Bridge Sampling to the Hierarchical EV Model . . . . .	52
<b>3</b>	<b>A Simple Method for Comparing Complex Models: Bayesian Model Comparison for Hierarchical Multinomial Processing Tree Models using Warp-III Bridge Sampling</b>	<b>55</b>
3.1	Introduction . . . . .	55
3.2	Multinomial Processing Trees . . . . .	58
3.3	Warp-III Bridge Sampling for MPTs . . . . .	63
3.4	Empirical Examples . . . . .	68
3.5	Discussion . . . . .	77
<b>4</b>	<b>Computing Bayes Factors for Evidence-Accumulation Models Using Warp-III Bridge Sampling</b>	<b>81</b>
4.1	Introduction . . . . .	81
4.2	Simple Monte Carlo Sampling . . . . .	86
4.3	Warp-III Bridge Sampling . . . . .	87
4.4	Simulation Study I: Nested Model Comparison for the Single-Participant Case . . . . .	90

4.5	Simulation Study II: Nested and Non-nested Model Comparison for the Hierarchical Case . . . . .	96
4.6	Simulation Study III: Estimating Equivocal Bayes Factors for the Hierarchical Case . . . . .	100
4.7	Discussion . . . . .	102
4.8	Conclusion . . . . .	108
4.A	Savage-Dickey Density Ratio . . . . .	109
4.B	Reversible Jump Markov Chain Monte Carlo . . . . .	109
<b>5</b>	<b>Bayesian Inference for Multidimensional Scaling Representations with Psychologically-Interpretable Metrics</b>	<b>113</b>
5.1	Introduction . . . . .	114
5.2	MDS Model Identifiability . . . . .	117
5.3	Bayesian MDS Inference via DE-MCMC . . . . .	124
5.4	Bayesian Model Comparison via Bridge Sampling . . . . .	124
5.5	Applications . . . . .	127
5.6	Discussion . . . . .	133
5.A	The Ordering Heuristic . . . . .	138
5.B	Transformation Ordered Vector (0-1 Bounded) . . . . .	139
<b>6</b>	<b>bridgesampling: An R Package for Estimating Normalizing Constants</b>	<b>141</b>
6.1	Introduction . . . . .	141
6.2	Bridge Sampling: The Algorithm . . . . .	145
6.3	Toy Example: Bayesian $T$ -test . . . . .	151
6.4	A “Black Box” Stan Interface . . . . .	155
6.5	Discussion . . . . .	164
<b>II</b>	<b>Multi-Model Meta-Analysis</b>	<b>169</b>
<b>7</b>	<b>Bayesian Mixture Modeling of Significant <math>P</math> Values: A Meta-Analytic Method to Estimate the Degree of Contamination from <math>\mathcal{H}_0</math></b>	<b>171</b>
7.1	Introduction . . . . .	172
7.2	A Bayesian Mixture Model for Significant $P$ Values . . . . .	173
7.3	Estimating the Model and Interpreting the Results . . . . .	176
7.4	Example 1: 587 $T$ -Test $P$ Values . . . . .	178
7.5	Example 2: Social Priming Studies and Yoked Controls . . . . .	179
7.6	Challenges and Limitations . . . . .	182
7.7	Concluding Comments . . . . .	183
7.A	Prior Sensitivity Analysis for Example 2: Social Priming Studies and Yoked Controls . . . . .	185
<b>8</b>	<b>A Primer on Bayesian Model-Averaged Meta-Analysis</b>	<b>189</b>
8.1	Introduction . . . . .	189
8.2	Bayesian Meta-Analysis . . . . .	191
8.3	Example: Testing the Self-Concept Maintenance Theory . . . . .	201

8.4	Discussion . . . . .	207
8.A	Changing the Prior Probabilities of the Hypotheses . . . . .	209
<b>9</b>	<b>A Bayesian Model-Averaged Meta-Analysis of the Power Pose Effect with Informed and Default Priors: The Case of Felt Power</b>	<b>213</b>
9.1	Introduction . . . . .	214
9.2	Method . . . . .	216
9.3	Results . . . . .	221
9.4	Discussion . . . . .	226
9.A	Robustness Check: Different Priors for the Between-Study Heterogeneity . . . . .	228
<b>III</b>	<b>Hypothesis Testing</b>	<b>231</b>
<b>10</b>	<b>Bayesian Evidence Accumulation in Experimental Mathematics: A Case Study of Four Irrational Numbers</b>	<b>233</b>
10.1	Introduction . . . . .	233
10.2	Bayes Factors to Quantify Evidence for General Laws . . . . .	234
10.3	The Normality of Irrational Numbers . . . . .	236
10.4	A Bayes Factor Multinomial Test for Normality . . . . .	237
10.5	Alternative Analysis . . . . .	242
10.6	Discussion and Conclusion . . . . .	246
10.A	Limit of the Difference Between the Log Bayes Factors . . . . .	250
<b>11</b>	<b>Informed Bayesian <math>T</math>-Tests</b>	<b>251</b>
11.1	Introduction . . . . .	251
11.2	Theory . . . . .	254
11.3	Practice . . . . .	257
11.4	Concluding Comments . . . . .	260
<b>12</b>	<b>Informed Bayesian Inference for the A/B Test</b>	<b>263</b>
12.1	Introduction . . . . .	263
12.2	Implementation Details . . . . .	266
12.3	Example: Effectiveness of Resilience Training . . . . .	272
12.4	Concluding Comments . . . . .	280
12.A	Interpretation of the Parameters . . . . .	282
12.B	Prior Elicitation: Implied Distributions . . . . .	282
12.C	Laplace Approximation Details . . . . .	287
12.D	Example: Effectiveness of Resilience Training (Default Analysis) . . . . .	292
12.E	Progesterone in Women with Bleeding in Early Pregnancy: Absence of Evidence, Not Evidence of Absence . . . . .	298
<b>13</b>	<b>Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection</b>	<b>301</b>
13.1	Introduction . . . . .	302
13.2	Bayesian Leave-One-Out Cross-Validation . . . . .	304
13.3	Example 1: Induction . . . . .	305

13.4 Example 2: Chance . . . . .	308
13.5 Example 3: Nullity of a Normal Mean . . . . .	311
13.6 Closing Comments . . . . .	314
13.A Derivation Example 1 – Induction . . . . .	316
13.B Derivation Example 2 – Chance . . . . .	316
13.C Derivation Example 3 – Nullity of a Normal Mean . . . . .	317
<b>14 Rejoinder: More Limitations of Bayesian Leave-One-Out Cross-Validation</b>	<b>319</b>
14.1 Mathematical Psychology: An Epistemic Enterprise . . . . .	320
14.2 Rejoinder to Vehtari, Simpson, Yao, & Gelman . . . . .	321
14.3 Rejoinder to Navarro . . . . .	330
14.4 Rejoinder to Shiffrin & Chandramouli . . . . .	332
14.5 Concluding Remarks . . . . .	334
14.A Jevons (1874) on Bayesian Model Averaging . . . . .	335
14.B Coherence of BMA and Bayesian Parameter Inference . . . . .	337
<b>IV Conclusion</b>	<b>341</b>
<b>15 Summary and Future Directions</b>	<b>343</b>
15.1 Part I: Bridge Sampling . . . . .	343
15.2 Part II: Multi-Model Meta-Analysis . . . . .	346
15.3 Part III: Hypothesis Testing . . . . .	348
15.4 General Conclusion . . . . .	350
<b>References</b>	<b>355</b>
<b>Nederlandse Samenvatting</b>	<b>395</b>
Deel I: Bridge Sampling . . . . .	395
Deel II: Multi-Model Meta-Analyse . . . . .	396
Deel III: Hypothese Toetsing . . . . .	397
<b>Acknowledgements</b>	<b>399</b>
<b>Publications</b>	<b>401</b>

# Introduction







---

## 1.1 A Competition Between Forecasters

Suppose you are interested in determining which of four weather forecasters, A, B, C, and D, is the most accurate. To this end, you consider their forecasts for three consecutive days. For simplicity, the forecasters predict only three different types of weather: rain, clouds, or sun. The predictions of each of the four forecasters are displayed in Table 1.1. With these probabilistic predictions in hand, it is straightforward to assess the accuracy of the four forecasters: we check how well the forecasters have predicted the weather on the three days of interest. Specifically, we consider how likely the actually observed weather is according to their forecasts. Naturally, the forecaster that has predicted the observed weather best is the most accurate.

On Day 1 there are clouds. Forecaster A has assigned 60% to this outcome, forecaster B 35%, forecaster C 55%, and forecaster D 33.3%. Since forecaster A has assigned the highest probability to the observed weather they are the most accurate for Day 1. On Day 2 it rains. Forecaster A has assigned 40% to this outcome, forecaster B 50%, forecaster C 40%, and forecaster D 33.3%. Based on this information we can update our knowledge about the accuracy of the four forecasters. Specifically, we can consider the probability that each forecaster has assigned to the observed data sequence “clouds”  $\rightarrow$  “rain”. Forecaster A has assigned  $60\% \times 40\% = 24\%$  to this sequence, forecaster B  $35\% \times 50\% = 17.5\%$ , forecaster C  $55\% \times 40\% = 22\%$ , and forecaster D  $33.3\% \times 33.3\% = 11.1\%$ . Therefore, although forecaster B has predicted the weather on Day 2 best, when taking into account all available data (i.e., the weather on Day 1 and Day 2), forecaster A is still the most accurate, followed by C, B, and D. On Day 3 the sun is shining. Forecaster A has assigned 70% to this outcome, forecaster B 40%, forecaster C 20%, and forecaster D 33.3%. Again, we can update our knowledge based on the new observation. Forecaster A had assigned 24% to the observed weather sequence on the first two days. Updating this probability with the information from Day 3 reveals that forecaster A has assigned  $24\% \times 70\% = 16.8\%$  to the observed weather sequence “clouds”  $\rightarrow$  “rain”  $\rightarrow$  “sun”. Forecaster B has assigned  $17.5\% \times 40\% = 7\%$  to this sequence, forecaster C  $22\% \times 20\% = 4.4\%$ , and forecaster D has assigned

Table 1.1: Predictions of four weather forecasters, A, B, C and D, for three consecutive days. The bold numbers correspond to the weather that actually occurred on that day.

			
<b>Weather on Day 1:</b> 			
Predictions of forecaster A	25%	<b>60%</b>	15%
Predictions of forecaster B	55%	<b>35%</b>	10%
Predictions of forecaster C	25%	<b>55%</b>	20%
Predictions of forecaster D	33.3%	<b>33.3%</b>	33.3%
<b>Weather on Day 2:</b> 			
Predictions of forecaster A	<b>40%</b>	50%	10%
Predictions of forecaster B	<b>50%</b>	35%	15%
Predictions of forecaster C	<b>40%</b>	35%	25%
Predictions of forecaster D	<b>33.3%</b>	33.3%	33.3%
<b>Weather on Day 3:</b> 			
Predictions of forecaster A	10%	20%	<b>70%</b>
Predictions of forecaster B	5%	55%	<b>40%</b>
Predictions of forecaster C	5%	75%	<b>20%</b>
Predictions of forecaster D	33.3%	33.3%	<b>33.3%</b>

$11.1\% \times 33.3\% = 3.7\%$  to this weather sequence. Therefore, based on these three days, we conclude that forecaster A is the most accurate, followed by forecaster B, forecaster C, and forecaster D. Note that we could naturally keep updating our knowledge about the accuracy of the forecasters. Specifically, we could obtain their predictions for future days and then check how likely the observed weather is given their forecasts.

We can not only assess who is the best forecaster for these three days, but we can also gauge how much better, say, forecaster A is compared to forecaster B. Specifically, the observed weather sequence is predicted  $16.8\%/7\% = 2.4$  times better by forecaster A than by forecaster B. Similarly, the observed weather sequence is predicted  $16.8\%/4.4\% = 3.8$  times better by forecaster A than by forecaster C. Finally, the observed weather sequence is predicted  $16.8\%/3.7\% = 4.5$  times better by forecaster A than by forecaster D. By transitivity, it follows that the observed weather sequence is predicted 1.6 times better by forecaster B than by forecaster C, 1.9 times better by forecaster B than by forecaster D, and 1.2 times better by forecaster C than by forecaster D. The factor by which one forecaster outpredicts another one is known in statistics as the *Bayes factor* (Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995), and it is a central part of this dissertation.

Another observation is that the predictions of forecaster D are trivial: on every day, D assigns equal probability to each of the three possible outcomes. Naively, one might think that this is a good strategy for performing reasonably well, since every possible outcome receives a decent probability. However, since the only

predictions that matter for assessing the quality of the forecasters are the ones for the observed weather, the vague predictions of forecaster D suffer a penalty compared to the more risky, precise predictions of the other forecasters. For instance, forecaster A has assigned 70% to sunshine on Day 3 and is rewarded for this precise prediction since it turns out to be true. Note, however, that this is only the case when these predictions are correct. For instance, forecaster C also made a relatively precise prediction for Day 3 (i.e., 75% chance of clouds). However, this precise prediction is not rewarded since the observed weather is sunshine. In fact, this precise but incorrect prediction results in forecaster C losing his second place to forecaster B. In sum, more risky, precise predictions are rewarded compared to vague predictions, but only in case they turn out to be true.

## 1.2 Treating Scientific Models as Forecasters

In science, researchers often aim to compare different accounts (i.e., models) of the world. As showcased in the previous section, when assessing who is the most accurate weather forecaster one just needs predictions and data for checking these predictions – nothing more. In science we can treat models as forecasters. Based on the forecasts of a number of competing models of interest, we can assess their relative predictive adequacy for observed data. This approach to comparing competing scientific accounts of the world is naturally implemented using Bayesian statistics (e.g., Dawid, 1984). Specifically, Bayesian statistics allows one to update one’s beliefs about the adequacy of competing accounts of the world by means of observed data. In Bayesian statistics predictive performance is the tool by which we learn, but prediction does not necessarily need to be the ultimate goal. Typically scientists are interested not only in predicting data but also in explaining phenomena. Nevertheless, to explain phenomena one typically compares different accounts of the world and using Bayesian statistics this is naturally accomplished by means of comparing predictions.

When forecasts are provided directly as in the weather forecast example, assessing their predictive adequacy based on observed data is straightforward. However, for scientific models we often need to work to see what the models actually predict. Specifically, when comparing models it can be challenging to find out how much probability exactly a model has assigned to the observed data. The reason is that scientific models typically feature parameters, often denoted by  $\theta$ , adjustable quantities that affect what data patterns a model can predict.

As a concrete example, we consider a simplified version of the exponential decay model for describing the relationship between memory retention and time (Lee & Wagenmakers, 2013; Shiffrin, Lee, Kim, & Wagenmakers, 2008). In a typical memory retention experiment participants are presented with a list of items and subsequently their ability to remember items from the list is tested after different periods of time have elapsed. The simplified exponential decay model features two parameters: (1)  $\alpha$  which corresponds to the rate of decay of information, and (2)  $\beta$  which corresponds to a baseline level of remembering. Specifically, the model stipulates that the probability of remembering an item after time  $t$  is

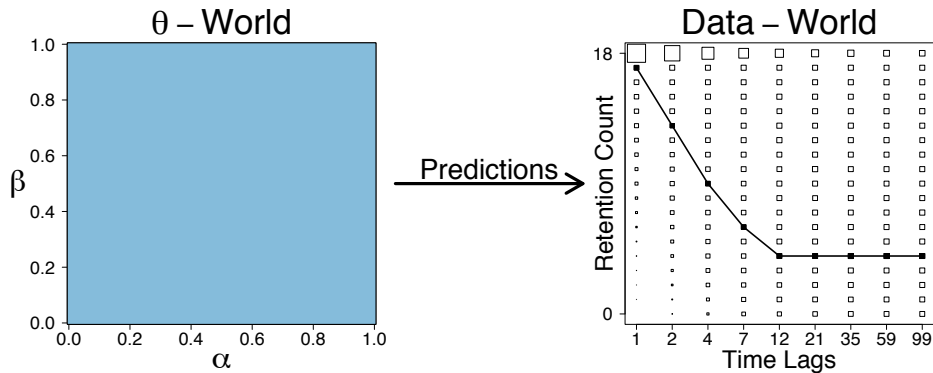


Figure 1.1: Transitioning from the  $\theta$ -world to the data-world to obtain the model’s predictions for the memory retention example. In the  $\theta$ -world, all possible combinations of  $\alpha$  and  $\beta$  are deemed equally plausible a priori. This results in very vague predictions for the data-world. The probabilities that are assigned to the different retention counts for each time lag are represented by the size of the corresponding squares. The superimposed black symbols that are connected by a line display the data of fictitious Participant 2 from Shiffrin et al. (2008). Available at <https://tinyurl.com/yyn7e2o9> under CC license <https://creativecommons.org/licenses/by/2.0/>.

$$\exp(-\alpha t) + \beta.^1$$

To assess the accuracy of the model as a forecaster, we need to find out what data the model predicts. Based on observed data we can then check, just as in the weather forecast example, how much probability the model has assigned to these observed data. For concreteness, suppose we are interested in an experiment that presents participants with 18 items and tests their ability to remember these items after 1, 2, 4, 7, 12, 21, 35, 59, and 99 seconds. Note that for different values of the parameters  $\alpha$  and  $\beta$  the model specifies a different exponential function of memory retention and hence also predicts different data. To determine what data the model predicts as a whole we need to consider the predictions for all possible combinations of  $\alpha$  and  $\beta$  and weight them by how plausible these specific combinations of  $\alpha$  and  $\beta$  are deemed a priori.

Figure 1.1 illustrates this process. The left part of the figure displays what can be called the *parameter-world* or  $\theta$ -world. Here we assume that  $\alpha$  and  $\beta$  can take values between 0 and 1 and each possible combination of  $\alpha$  and  $\beta$  is equally plausible a priori. To obtain the model’s predictions we need to transition from the  $\theta$ -world to what can be called the *data-world*. Specifically, for each possible combination of  $\alpha$  and  $\beta$  we need to determine the resulting predictions and then take a weighted average of all of these predictions. The averaging weights correspond to how plausible each specific combination of  $\alpha$  and  $\beta$  is deemed a

<sup>1</sup>To make sure that this yields a probability, the restriction is imposed that the resulting value cannot be smaller than 0 or larger than 1.

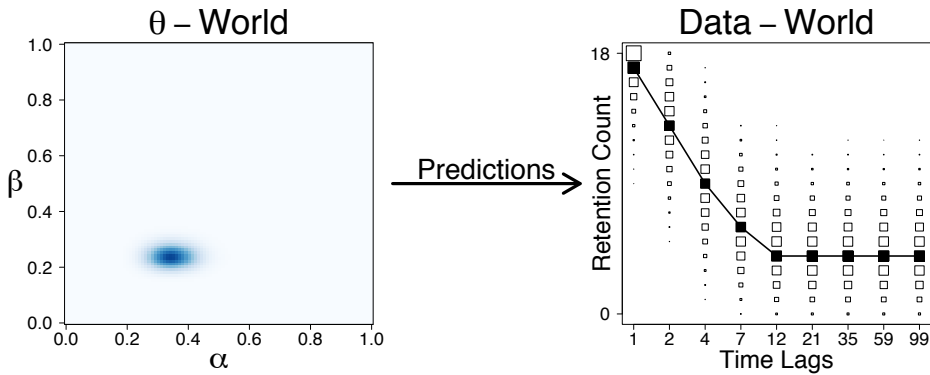


Figure 1.2: Transitioning from the  $\theta$ -world to the data-world to obtain the model’s predictions for the memory retention example. In the  $\theta$ -world, certain combinations of  $\alpha$  and  $\beta$  are deemed more plausible than others based on the data of the three fictitious participants from Shiffrin et al. (2008). This results in more precise predictions for the data-world. The probabilities that are assigned to the different retention counts for each time lag are represented by the size of the corresponding squares. The superimposed black symbols that are connected by a line display the data of fictitious Participant 2 from Shiffrin et al. (2008). Available at <https://tinyurl.com/y4o2q4d7> under CC license <https://creativecommons.org/licenses/by/2.0/>.

priori. The resulting predictions are displayed in the right part of Figure 1.1. It is apparent that these predictions are very vague. Specifically, for many time lags the different possible numbers of remembered items (i.e., retention count) are all assigned a similar probability. Only for the first time lags, the model makes more precise predictions which are that a higher retention count is more likely than a lower retention count. To determine the accuracy of the model’s predictions one needs observed data. As an example, the superimposed black symbols that are connected by a line in the right part of Figure 1.1 display the data of fictitious Participant 2 from Shiffrin et al. (2008). It is apparent that the model has assigned a decent probability to the observed retention curve, however, the model has clearly also assigned a similar probability to a number of (very) different possible retention curves.

To make the model predictions more precise we can incorporate prior knowledge about plausible values for  $\alpha$  and  $\beta$ , for instance, by considering previous experimental data. To demonstrate how the incorporation of prior knowledge can result in more precise predictions, we use the data of the three fictitious participants from Shiffrin et al. (2008). The left part of Figure 1.2 displays again the  $\theta$ -world. However, this time certain combinations of  $\alpha$  and  $\beta$  are assigned more plausibility than others based on what we have learned about these parameters from the data of the three fictitious participants. The right part of Figure 1.2 displays the resulting predictions. Clearly, the predictions are much more precise

than before. When assessing the accuracy of the model as a forecaster, these more precise predictions will be advantageous in case the observed data indeed correspond to the retention counts that are assigned more probability than others. For instance, the superimposed black symbols that are connected by a line again correspond to the data of fictitious Participant 2. It is apparent that the model has predicted this observed retention curve very well. However, the refined knowledge about what values of the parameters are more plausible than others is partially based on exactly these data of Participant 2, so it is not valid to use these data a second time for assessing the accuracy of the resulting predictions. Instead, these predictions must be tested based on new data.

In practice there are typically several participants which complicates transitioning from the  $\theta$ -world to the data-world and determining how much probability a model has assigned to the observed data of all participants simultaneously. Furthermore, transitioning from the  $\theta$ -world to the data-world is also more challenging for complex models that feature many parameters. A substantial part of this dissertation is concerned with computational procedures that make it easier to transition from the  $\theta$ -world to the data-world to obtain a model's predictions. Specifically, these procedures provide an estimate of the *marginal likelihood*, the probability of the data given a model, which allows researchers to assess how well a model has predicted observed data. Based on this quantity researchers can compare different accounts of the world, just as we compared different weather forecasters in the introductory example, using the Bayes factor. For instance, in the memory retention example, one could obtain the predictions for a competing model of memory retention (e.g., a model that specifies a power function) and then compare the predictive adequacy of the two models for observed data using the Bayes factor.

The weather forecast example and also the memory retention example illustrated that it may be advantageous to make more precise predictions since they are rewarded in case they are accurate. A few chapters of this dissertation are concerned with providing statistical procedures to researchers that allow them to make their hypotheses more precise by incorporating prior information about the quantities of interest.

Sometimes when comparing different scientific accounts of the world, there may not be a model that is clearly favored by the data. For instance, in the weather forecast example, the observed weather sequence was predicted only about 2.4 times better by forecaster A than by forecaster B. Suppose you are interested in obtaining an accurate weather forecast for a new day. In this case, it may be prudent to take into account not only the predictions of forecaster A, but to consider the predictions of all forecasters. Specifically, we can obtain a combined, averaged prediction for the new day by weighting each forecaster's predictions by how well they have done so far. Concretely, we want to take into account predictions from A, B, C, and D, but we want to trust the predictions of A more than the ones of B, C, and D since they have performed better so far. This approach is naturally implemented using Bayesian model averaging (e.g., Hoeting, Madigan, Raftery, & Volinsky, 1999). A few chapters of this dissertation provide concrete applications of this procedure for taking into account model uncertainty to prevent overconfident conclusions that one could obtain by trusting a single

forecaster.

## 1.3 Chapter Outline

### 1.3.1 Part I: Bridge Sampling

The first part of the dissertation is concerned with *bridge sampling*, a computational procedure that facilitates the transition from the  $\theta$ -world to the data-world to obtain a model's predictions for observed data. Specifically, bridge sampling yields an estimate of a model's marginal likelihood, the probability of the data given a model.

Chapter 2 is a tutorial on bridge sampling. The method is introduced by comparing it with three other Monte Carlo sampling procedures for estimating the marginal likelihood in a simple beta-binomial example. The feasibility of the approach in practice is demonstrated using single-participant and hierarchical versions of a reinforcement learning model. It is argued that bridge sampling is an attractive method for comparing models in mathematical psychology where researchers are often interested in comparing a limited set of possibly non-nested models that are implemented in a hierarchical fashion and may have many parameters.

Chapter 3 applies an advanced version of bridge sampling called *Warp-III* for comparing hierarchical multinomial processing tree (MPT) models. This version of bridge sampling accounts for potential skewness in the posterior distribution and can thereby provide more precise estimates of the marginal likelihood of the models. The first example demonstrates how this procedure can be used to assess which model parameters differ across trials. Specifically, similar to the idea of combining the predictions of several weather forecasters, Bayesian model averaging is used to assess which parameters vary across trials. The second example reanalyzes data that have been used to compare two non-nested MPT models concerning the illusory truth effect.

Chapter 4 applies Warp-III bridge sampling for computing the marginal likelihood of evidence-accumulation models. Specifically, using the Linear Ballistic Accumulator (LBA) model it is demonstrated that the combination of differential evolution Markov chain Monte Carlo (DE-MCMC) and Warp-III bridge sampling provides precise estimates of the marginal likelihood for both single-participant and hierarchical versions of the LBA. An easy-to-use software implementation is provided that allows researchers to estimate the marginal likelihood for many evidence-accumulation models in a straightforward manner. The chapter concludes with a series of recommendations for applying Warp-III bridge sampling in practical applications.

Chapter 5 applies Bayesian methods to multidimensional scaling (MDS) models for inferring the appropriate number of dimensions and the metric structure of the space used to measure distance. Specifically, priors are defined for making the model identifiable under metrics corresponding to psychologically separable and psychologically integral stimulus domains. DE-MCMC is used in combination with Warp-III bridge sampling to make inference about the model parameters, to identify the appropriate number of dimensions, and to infer the appropriate

metric of the latent space. Using five existing data sets, it is demonstrated that the procedure provides sensible results. The chapter also discusses a number of remaining technical challenges that need to be addressed before the method can be applied generally in a straightforward fashion.

Chapter 6 introduces **bridgesampling**, an R package for estimating the marginal likelihood (or, more generally, normalizing constants) using bridge sampling in a generic and easy-to-use fashion. In combination with the Bayesian sampling software **Stan** (Carpenter et al., 2017), the R package can provide automatic estimates of the marginal likelihood. The package functionality is demonstrated using three examples.

### 1.3.2 Part II: Multi-Model Meta-Analysis

The second part of the dissertation is concerned with methods for meta-analyzing a set of studies. The idea of combining several forecasters using Bayesian model averaging is applied in a few chapters of this part.

Chapter 7 proposes a Bayesian mixture model for meta-analyzing the distribution of significant  $p$  values of a set of studies. Specifically, the mixture model estimates the proportion of significant results that originate from the null hypothesis of no effect, and it also provides an estimate of the probability that each specific  $p$  value originates from the null hypothesis. The procedure is demonstrated using two examples. A web application is provided to enable researchers to apply the method in a straightforward manner to any set of significant  $p$  values.

Chapter 8 is a primer on Bayesian model-averaged meta-analysis. This procedure applies the idea of combining several forecasters to avoid an all-or-none decision between a fixed-effect and a random-effects meta-analysis model. Specifically, this approach combines four Bayesian meta-analysis models according to their plausibility in light of the observed data: (1) fixed-effect null hypothesis, (2) fixed-effect alternative hypothesis, (3) random-effects null hypothesis, and (4) random-effects alternative hypothesis. This procedure allows researchers to address, in a principled manner, the two key questions “Is the overall effect non-zero?” and “Is there between-study variability in effect size?”. The method is illustrated with an example concerning the self-concept maintenance theory.

Chapter 9 applies the Bayesian model-averaged meta-analysis introduced in Chapter 8 to a set of six preregistered studies concerning the effect of power posing. Specifically, the analysis focuses on the effect of power posing on felt power. The meta-analysis yields very strong evidence for an effect of power posing on felt power. However, the evidence is only moderate when one takes into account only participants that were unfamiliar with the effect.

### 1.3.3 Part III: Hypothesis Testing

The third part of the dissertation is concerned with hypothesis testing. Specifically, Bayesian approaches to a number of standard statistical tests are presented and it is demonstrated how these can be used to address questions of interest. A recurring theme is the ability to incorporate prior knowledge into the analyses which helps make the hypotheses more precise and can thus yield tests that are

more diagnostic and correspond closer to what researchers actually want to test. Just as in the weather forecast example, these more precise predictions will be rewarded when comparing different models in case they turn out to be true.

Chapter 10 illustrates how Bayesian inference can be used to quantify the evidence in favor of a general law based on finite data. Concretely, the chapter focuses on quantifying evidence in favor of the hypothesis that certain fundamental constants (i.e.,  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ ) are normal. Specifically, Bayesian inference is used to test the more restricted hypothesis that each digit in the constants' decimal expansions occurs equally often. For all four constants the evidence in favor of the general law is overwhelming.

Chapter 11 proposes the use of a flexible  $t$ -prior for effect size in the Bayesian  $t$ -test. This prior allows researchers to incorporate advance knowledge into the analysis to make their predictions more precise. Furthermore, this prior specification contains previous subjective, but also objective Bayesian  $t$ -test versions. Two measures for informed prior distributions are proposed that quantify the departure from the objective Bayes factor desiderata of predictive matching and information consistency. The approach is illustrated using an example concerning the facial feedback hypothesis that features an expert prior elicitation effort.

Chapter 12 introduces **abtest**, an R package for conducting Bayesian A/B tests. The implemented approach is based on work by Kass and Vaidyanathan (1992) and allows researchers to monitor the evidence for the hypotheses that the treatment has either a positive effect, a negative effect, or, crucially, no effect. This method also enables one to incorporate expert knowledge about the relative prior plausibility of the rival hypotheses as well as about the expected size of the effect.

Chapter 13 discusses Bayesian leave-one-out cross-validation (LOO), an alternative method for comparing competing models. Several limitations of this approach are demonstrated using concrete examples and it is concluded that LOO is not a panacea for model selection.

Chapter 14 is a rejoinder to three commentaries on Chapter 13. Each of the commentaries is addressed and additional limitations of methods that are based on LOO (such as Bayesian stacking) are identified. These methods are contrasted with approaches that consistently use Bayes' rule for both parameter estimation and model comparison. It is concluded that LOO-based methods do not align satisfactorily with the epistemic goal of mathematical psychology.



Part I

Bridge Sampling



---

# A Tutorial on Bridge Sampling

---

## Abstract

The marginal likelihood plays an important role in many areas of Bayesian statistics such as parameter estimation, model comparison, and model averaging. In most applications, however, the marginal likelihood is not analytically tractable and must be approximated using numerical methods. Here we provide a tutorial on bridge sampling (Bennett, 1976; Meng & Wong, 1996), a reliable and relatively straightforward sampling method that allows researchers to obtain the marginal likelihood for models of varying complexity. First, we introduce bridge sampling and three related sampling methods using the beta-binomial model as a running example. We then apply bridge sampling to estimate the marginal likelihood for the Expectancy Valence (EV) model – a popular model for reinforcement learning. Our results indicate that bridge sampling provides accurate estimates for both a single participant and a hierarchical version of the EV model. We conclude that bridge sampling is an attractive method for mathematical psychologists who typically aim to approximate the marginal likelihood for a limited set of possibly high-dimensional models.

## 2.1 Introduction

Bayesian statistics has become increasingly popular in mathematical psychology (Andrews & Baguley, 2013; Bayarri, Benjamin, Berger, & Sellke, 2016; Poirier, 2006; Vanpaemel, 2016; Verhagen, Levy, Millsap, & Fox, 2015; Wetzels et al., 2016). The Bayesian approach is conceptually simple, theoretically coherent, and

---

This chapter is published as Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–97. doi: <https://doi.org/10.1016/j.jmp.2017.09.005>. Also available as *arXiv preprint*: <https://arxiv.org/abs/1703.05984>

easily applied to relatively complex problems. These problems include, for instance, hierarchical modeling (Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Matzke & Wagenmakers, 2009; Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder et al., 2007) or the comparison of non-nested models (Lee, 2008; Pitt, Myung, & Zhang, 2002; Shiffrin et al., 2008). Three major applications of Bayesian statistics concern parameter estimation, model comparison, and Bayesian model averaging. In all three areas, the marginal likelihood – that is, the probability of the observed data given the model of interest – plays a central role (see also Gelman & Meng, 1998).

First, in parameter estimation, we consider a single model and aim to quantify the uncertainty for a parameter of interest  $\theta$  after having observed the data  $y$ . This is realized by means of a posterior distribution that can be obtained using Bayes’ theorem:

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{\int p(y | \theta') p(\theta') d\theta'} = \frac{\overbrace{p(y | \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(y)}_{\text{marginal likelihood}}}. \quad (2.1)$$

Here, the marginal likelihood of the data  $p(y)$  ensures that the posterior distribution is a proper probability density function (PDF) in the sense that it integrates to 1. This illustrates why in parameter estimation the marginal likelihood is referred to as a normalizing constant.

Second, in model comparison, we consider  $m$  ( $m \in \mathbb{N}$ ) competing models, and are interested in the relative plausibility of a particular model  $\mathcal{M}_i$  ( $i \in \{1, 2, \dots, m\}$ ) given the prior model probability and the evidence from the data  $y$  (see three special issues on this topic in the *Journal of Mathematical Psychology*: J. Mulder & Wagenmakers, 2016; Myung, Forster, & Browne, 2000a; Wagenmakers & Waldorp, 2006a). This relative plausibility is quantified by the so-called posterior model probability  $p(\mathcal{M}_i | y)$  of model  $\mathcal{M}_i$  given the data  $y$  (Berger & Molina, 2005):

$$p(\mathcal{M}_i | y) = \frac{p(y | \mathcal{M}_i) p(\mathcal{M}_i)}{\sum_{j=1}^m p(y | \mathcal{M}_j) p(\mathcal{M}_j)}, \quad (2.2)$$

where the denominator is the sum of the marginal likelihood times the prior model probability of all  $m$  models. In model comparison, the marginal likelihood for a specific model is also referred to as the model evidence (Didelot, Everitt, Johansen, & Lawson, 2011), the integrated likelihood (Kass & Raftery, 1995), the predictive likelihood of the model (Gamerman & Lopes, 2006, Chapter 7), the predictive probability of the data (Kass & Raftery, 1995), or the prior predictive density (Ntzoufras, 2009). Note that conceptually the marginal likelihood of Equation 2.2 is the same as the marginal likelihood of Equation 2.1. However, for the latter equation we dropped the model index because in parameter estimation we consider only one model.

If only two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  are considered, Equation 2.2 can be used to quantify the relative posterior model plausibility of model  $\mathcal{M}_1$  compared to model  $\mathcal{M}_2$ . This relative plausibility is given by the ratio of the posterior probabilities

of both models, and is referred to as the posterior model odds:

$$\underbrace{\frac{p(\mathcal{M}_1 | y)}{p(\mathcal{M}_2 | y)}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{prior odds}} \times \underbrace{\frac{p(y | \mathcal{M}_1)}{p(y | \mathcal{M}_2)}}_{\text{Bayes factor}}. \quad (2.3)$$

Equation 2.3 illustrates that the posterior model odds are the product of two factors: The first factor is the ratio of the prior probabilities of both models – the prior model odds. The second factor is the ratio of the marginal likelihoods of both models – the so-called Bayes factor (Etz & Wagenmakers, 2017; Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016a, 2016b; Robert, 2016). The Bayes factor plays an important role in model comparison and is referred to as the “standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis & Raftery, 1997, p. 648) and “the primary tool used in Bayesian inference for hypothesis testing and model selection” (Berger, 2006, p. 378).

Third, the marginal likelihood plays an important role in Bayesian model averaging (BMA; Hoeting et al., 1999) where aspects of parameter estimation and model comparison are combined. As in model comparison, BMA considers several models; however, it does not aim to identify a single best model. Instead it fully acknowledges model uncertainty. Model-averaged parameter inference can be obtained by combining, across all models, the posterior distribution of the parameter of interest weighted by each model’s posterior model probability, and as such depends on the marginal likelihood of the models. This procedure assumes that the parameter of interest has identical interpretation across the different models. Model-averaged predictions can be obtained in a similar manner.

A problem that arises in all three areas – parameter estimation, model comparison, and BMA – is that an analytical expression of the marginal likelihood can be obtained only for certain restricted examples. This is a pressing problem in Bayesian modeling, and in particular in mathematical psychology where models can be non-linear and equipped with a large number of parameters, especially when the models are implemented in a hierarchical framework. Such a framework incorporates both commonalities and differences between participants of one group by assuming that the model parameters of each participant are drawn from a group-level distribution (for advantages of the Bayesian hierarchical framework see Ahn, Krawitz, Kim, Bussemeyer, & Brown, 2011; Navarro, Griffiths, Steyvers, & Lee, 2006; Rouder & Lu, 2005; Rouder, Lu, Morey, Sun, & Speckman, 2008; Rouder et al., 2005; Scheibehenne & Pachur, 2015; Shiffrin et al., 2008; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010). For instance, consider a four-parameter Bayesian hierarchical model with four group-level distributions each characterized by two parameters and a group size of 30 participants; this then results in  $30 \times 4$  individual-level parameters and  $2 \times 4$  group-level parameters for a total of 128 parameters. In sum, even simple models quickly become complex once hierarchical aspects are introduced and this frustrates the derivation of the marginal likelihood.

To overcome this problem, several Monte Carlo sampling methods have been proposed to approximate the marginal likelihood. In this tutorial we focus on four

such methods: the bridge sampling estimator (Bennett, 1976; Chapter 5 of M.-H. Chen, Shao, & Ibrahim, 2002; Meng & Wong, 1996), and its three commonly used special cases – the naive Monte Carlo estimator, the importance sampling estimator, and the generalized harmonic mean estimator (for alternative methods see Gamerman & Lopes, 2006, Chapter 7; and for alternative approximation methods relevant to model comparison and BMA see Carlin & Chib, 1995; Green, 1995).<sup>1</sup> As we will illustrate throughout this tutorial, the bridge sampler is accurate, efficient, and relatively straightforward to implement (e.g., DiCiccio, Kass, Raftery, & Wasserman, 1997; Frühwirth-Schnatter, 2004; Meng & Wong, 1996).

The goal of this tutorial is to bring the bridge sampling estimator to the attention of mathematical psychologists. We aim to explain this estimator and facilitate its application by suggesting a step-by-step implementation scheme. To this end, we first show how bridge sampling and the three special cases can be used to approximate the marginal likelihood in a simple beta-binomial model. We begin with the naive Monte Carlo estimator and progressively work our way up – via the importance sampling estimator and the generalized harmonic mean estimator – to the most general case considered: the bridge sampling estimator. This order was chosen such that key concepts are introduced gradually and estimators are of increasing complexity and sophistication. The first three estimators are included in this tutorial with the sole purpose of facilitating the reader’s understanding of bridge sampling. In the second part of this tutorial, we outline how the bridge sampling estimator can be used to derive the marginal likelihood for the Expectancy Valence (EV; Bussemeyer & Stout, 2002) model – a popular, yet relatively complex reinforcement-learning model for the Iowa gambling task (Bechara, Damasio, Damasio, & Anderson, 1994). We apply bridge sampling to both an individual-level and a hierarchical implementation of the EV model.

Throughout the chapter, we use the software package R (R Core Team, 2019) to implement the bridge sampling estimator for the various models. The interested reader is invited to reproduce our results by downloading the code and all relevant materials from our Open Science Framework folder at <https://osf.io/f9cq4/>.

## 2.2 Four Sampling Methods to Approximate the Marginal Likelihood

In this section we outline four standard methods to approximate the marginal likelihood. For more detailed explanations and derivations, we recommend Ntzoufras (2009, Chapter 11) and Gamerman and Lopes (2006, Chapter 7); a comparative review of the different sampling methods is presented in DiCiccio et al. (1997). The marginal likelihood is the probability of the observed data  $y$  given a specific model of interest  $\mathcal{M}$ , and is defined as the integral of the likelihood over the prior:

$$\underbrace{p(y \mid \mathcal{M})}_{\text{marginal likelihood}} = \int \underbrace{p(y \mid \theta, \mathcal{M})}_{\text{likelihood}} \underbrace{p(\theta \mid \mathcal{M})}_{\text{prior}} d\theta, \quad (2.4)$$

---

<sup>1</sup>The appendix provides a derivation showing that the first three estimators are indeed special cases of the bridge sampler.

with  $\theta$  a vector containing the model parameters. Equation 2.4 illustrates that the marginal likelihood can be interpreted as a weighted average of the likelihood of the data given a specific value for  $\theta$  where the weight is the a priori plausibility of that specific value. Equation 2.4 can therefore be written as an expected value:

$$p(y \mid \mathcal{M}) = \mathbb{E}_{\text{prior}} [p(y \mid \theta, \mathcal{M})],$$

where the expectation is taken with respect to the prior distribution. This idea is central to the four sampling methods that we discuss in this tutorial.

### 2.2.1 Introduction of the Running Example: The Beta-Binomial Model

Our running example focuses on estimating the marginal likelihood for a binomial model assuming a uniform prior on the rate parameter  $\theta$  (i.e., the beta-binomial model). Consider a single participant who answered  $k = 2$  out of  $n = 10$  true/false questions correctly. Assume that the number of correct answers follows a binomial distribution, that is,  $k \sim \text{Binomial}(n, \theta)$  with  $\theta \in (0, 1)$ , where  $\theta$  represents the latent probability for answering any one question correctly. The probability mass function (PMF) of the binomial distribution is given by:

$$\text{Binomial}(k \mid n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \quad (2.5)$$

where  $k, n \in \mathbb{Z}_{\geq 0}$ , and  $k \leq n$ . The PMF of the binomial distribution serves as the likelihood function in our running example.

In the Bayesian framework, we also have to specify the prior distribution of the model parameters; the prior distribution expresses our knowledge about the parameters before the data have been observed. In our running example, we assume that all values of  $\theta$  are equally likely a priori. This prior belief is captured by a uniform distribution across the range of  $\theta$ , that is,  $\theta \sim \text{Uniform}(0, 1)$  which can equivalently be written in terms of a beta distribution  $\theta \sim \text{Beta}(1, 1)$ . This prior distribution is represented by the dotted line in Figure 2.1. It is evident that the density of the prior distribution equals 1 for all values of  $\theta$ . One advantage of expressing the prior distribution by a beta distribution is that its two parameters (i.e., in its general form the shape parameters  $\alpha$  and  $\beta$ ) can be thought of as counts of “prior successes” and “prior failures”, respectively. In its general form, the PDF of a Beta( $\alpha, \beta$ ) distribution ( $\alpha, \beta > 0$ ) is given by:

$$\text{Beta}(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1} (1 - \theta)^{\beta-1}}{B(\alpha, \beta)},$$

where  $B(\alpha, \beta)$  is the beta function that is defined as:  $B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1 - t)^{\beta-1} dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ , and  $\Gamma(n) = (n - 1)!$  for  $n \in \mathbb{N}$ .

#### 2.2.1.1 Analytical Derivation of the Marginal Likelihood

As we will see in this section, the beta-binomial model constitutes one of the rare examples where the marginal likelihood is analytic. Assuming a general  $k$  and  $n$ ,

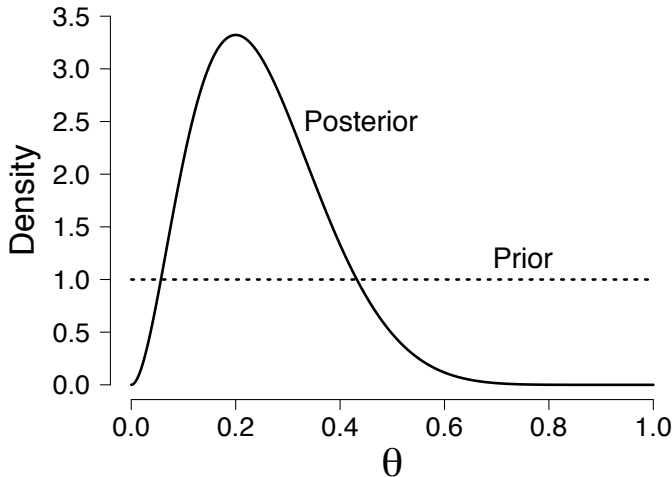


Figure 2.1: Prior and posterior distribution for the rate parameter  $\theta$  from the beta-binomial model. The Beta(1, 1) prior on the rate parameter  $\theta$  is represented by the dotted line; the Beta(3, 9) posterior distribution is represented by the solid line and was obtained after having observed 2 correct responses out of 10 trials. Available at <https://tinyurl.com/yc8bw98v> under CC license <https://creativecommons.org/licenses/by/2.0/>.

we obtain the marginal likelihood as:

$$\begin{aligned} p(k | n) &\stackrel{\text{Eq. 2.4}}{=} \int_0^1 p(k | n, \theta) p(\theta) d\theta = \int_0^1 \binom{n}{k} \theta^k (1 - \theta)^{n-k} 1 d\theta \\ &= \binom{n}{k} B(k + 1, n - k + 1) = \frac{1}{n + 1}, \end{aligned} \quad (2.6)$$

where we suppress the “model” in the conditioning part of the probability statements because we focus on a single model in this running example. Using  $k = 2$  and  $n = 10$  of our example, we obtain:  $p(k = 2 | n = 10) = 1/11 \approx 0.0909$ . This value will be estimated in the remainder of the running example using the naive Monte Carlo estimator, the importance sampling estimator, the generalized harmonic mean estimator, and finally the bridge sampling estimator.

As we will see below, the importance sampling estimator, generalized harmonic mean estimator, and bridge sampling estimator require samples from the posterior distribution. These samples can be obtained using computer software such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS (Plummer, 2003), or Stan (Stan Development Team, 2016), even when the marginal likelihood that functions here as a normalizing constant is not known (Equation 2.1). However, in our running example MCMC samples are not required because we can derive an analytical expression of the posterior distribution for  $\theta$  after having observed

the data. Using the analytic expression of the marginal likelihood (Equation 2.6) and Bayes' theorem, we obtain:

$$p(\theta | k, n) = \frac{p(k | n, \theta) p(\theta)}{p(k | n)} = \frac{\binom{n}{k} \theta^k (1 - \theta)^{n-k} 1}{\binom{n}{k} B(k + 1, n - k + 1)} = \frac{\theta^k (1 - \theta)^{n-k}}{B(k + 1, n - k + 1)},$$

which we recognize as the PDF of the  $\text{Beta}(k + 1, n - k + 1)$  distribution. Thus, if we assume a uniform prior on  $\theta$  and observe  $k = 2$  correct responses out of  $n = 10$  trials, we obtain a  $\text{Beta}(3, 9)$  distribution as posterior distribution. This distribution is represented by the solid line in Figure 2.1. In general, if  $k | n, \theta \sim \text{Binomial}(n, \theta)$  and  $\theta \sim \text{Beta}(1, 1)$ , then  $\theta | n, k \sim \text{Beta}(k + 1, n - k + 1)$ .

## 2.2.2 Method 1: The Naive Monte Carlo Estimator of the Marginal Likelihood

The simplest method to approximate the marginal likelihood is provided by the naive Monte Carlo estimator (Hammersley & Handscomb, 1964; Raftery & Banfield, 1991). This method uses the standard definition of the marginal likelihood (Equation 2.4), and relies on the central idea that the marginal likelihood can be written as an expected value with respect to the prior distribution, that is,  $p(y) = \mathbb{E}_{\text{prior}} [p(y | \theta)]$ . This expected value of the likelihood of the data with respect to the prior can be approximated by evaluating the likelihood in  $N$  samples from the prior distribution for  $\theta$  and averaging the resulting values. This yields the naive Monte Carlo estimator  $\hat{p}_1(y)$ :

$$\hat{p}_1(y) = \underbrace{\frac{1}{N} \sum_{i=1}^N p(y | \tilde{\theta}_i)}_{\text{average likelihood}}, \quad \underbrace{\tilde{\theta}_i \sim p(\theta)}_{\text{samples from the prior distribution}}. \quad (2.7)$$

### 2.2.2.1 Running Example

To obtain the naive Monte Carlo estimate of the marginal likelihood in our running example, we need  $N$  samples from the  $\text{Beta}(1, 1)$  prior distribution for  $\theta$ . For illustrative purposes, we limit the number of samples to 12 whereas in practice one should take  $N$  to be very large. We obtain the following samples:

$$\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\} = \{0.58, 0.76, 0.03, 0.93, 0.27, 0.97, 0.45, 0.46, 0.18, 0.64, 0.06, 0.15\},$$

where we use the tilde symbol to emphasize that we refer to a sampled value. All sampled values are represented by the gray dots in Figure 2.2.

Following Equation 2.7, the next step is to calculate the likelihood (Equation 2.5) for each  $\tilde{\theta}_i$ , and then to average all obtained likelihood values. This yields the naive Monte Carlo estimate of the marginal likelihood:

$$\hat{p}_1(k = 2 | n = 10) = \frac{1}{12} \sum_{i=1}^{12} p(k = 2 | n = 10, \tilde{\theta}_i) = \frac{1}{12} \sum_{i=1}^{12} \binom{n}{k} (\tilde{\theta}_i)^k (1 - \tilde{\theta}_i)^{n-k}$$

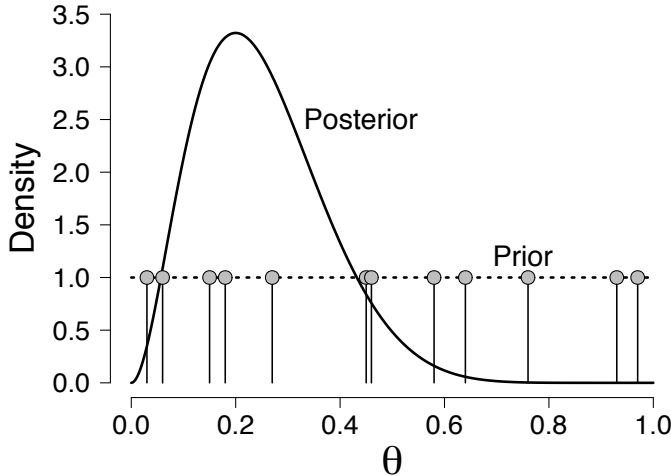


Figure 2.2: Illustration of the naive Monte Carlo estimator for the beta-binomial example. The dotted line represents the prior distribution and the solid line represents the posterior distribution that was obtained after having observed 2 correct responses out of 10 trials. The gray dots represent the 12 samples  $\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\}$  randomly drawn from the  $\text{Beta}(1, 1)$  prior distribution. Available at <https://tinyurl.com/y8uf6t8f> under CC license <https://creativecommons.org/licenses/by/2.0/>.

$$\begin{aligned}
 &= \frac{1}{12} \binom{10}{2} (0.58^2(1 - 0.58)^8 + \dots + 0.15^2(1 - 0.15)^8) \\
 &= 0.0945.
 \end{aligned}$$

### 2.2.3 Method 2: The Importance Sampling Estimator of the Marginal Likelihood

The naive Monte Carlo estimator introduced in the last section performs well if the prior and posterior distribution have a similar shape and strong overlap. However, the estimator is unstable if the posterior distribution is peaked relative to the prior (e.g., Gamerman & Lopes, 2006; Ntzoufras, 2009). In such a situation, most of the sampled values for  $\theta$  result in likelihood values close to zero and contribute only minimally to the estimate. This means that those few samples that result in high likelihood values dominate estimates of the marginal likelihood. Consequently, the variance of the estimator is increased (Newton & Raftery, 1994; Pajor, 2017).<sup>2</sup>

---

<sup>2</sup>The interested reader is referred to Pajor (2017) for a recent improvement on the calculation of the naive Monte Carlo estimator. The proposed improvement involves trimming the prior distribution in such a way that regions with low likelihood values are eliminated, thereby increasing the accuracy and efficiency of the estimator.

The importance sampling estimator, on the other hand, overcomes this shortcoming by boosting sampled values in regions of the parameter space where the integrand of Equation 2.4 is large. This is realized by using samples from a so-called importance density  $g_{IS}(\theta)$  instead of the prior distribution. The advantage of sampling from an importance density is that values for  $\theta$  that result in high likelihood values are sampled most frequently, whereas values for  $\theta$  with low likelihood values are sampled only rarely.

To derive the importance sampling estimator, Equation 2.4 is used as starting point which is then extended by the importance density  $g_{IS}(\theta)$ :

$$\begin{aligned} p(y) &= \int p(y | \theta) p(\theta) d\theta = \int p(y | \theta) p(\theta) \frac{g_{IS}(\theta)}{g_{IS}(\theta)} d\theta = \int \frac{p(y | \theta) p(\theta)}{g_{IS}(\theta)} g_{IS}(\theta) d\theta \\ &= \mathbb{E}_{g_{IS}(\theta)} \left( \frac{p(y | \theta) p(\theta)}{g_{IS}(\theta)} \right). \end{aligned}$$

This yields the importance sampling estimator  $\hat{p}_2(y)$ :

$$\hat{p}_2(y) = \underbrace{\frac{1}{N} \sum_{i=1}^N \frac{p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}}_{\text{average adjusted likelihood}}, \quad \underbrace{\tilde{\theta}_i \sim g_{IS}(\theta)}_{\text{samples from the importance density}}. \quad (2.8)$$

A suitable importance density should (1) be easy to evaluate; (2) have the same domain as the posterior distribution; (3) closely resemble the posterior distribution; and (4) have fatter tails than the posterior distribution (Neal, 2001; Vandekerckhove, Matzke, & Wagenmakers, 2015). The latter criterion ensures that values in the tails of the distribution cannot misleadingly dominate the estimate (Neal, 2001).<sup>3</sup>

### 2.2.3.1 Running Example

To obtain the importance sampling estimate of the marginal likelihood in our running example, we first need to choose an importance density  $g_{IS}(\theta)$ . An importance density that fulfills the four above mentioned desiderata is a mixture between a beta density that provides the best fit to the posterior distribution and a uniform density across the range of  $\theta$  (Vandekerckhove et al., 2015). The relative impact of the uniform density is quantified by a mixture weight  $\gamma$  that ranges between 0 and 1. The larger  $\gamma$ , the higher the influence of the uniform density resulting in a less peaked distribution with thick tails. If  $\gamma = 1$ , the beta

---

<sup>3</sup>To illustrate the need for an importance density with fatter tails than the posterior distribution, imagine you sample from the tail region of an importance density with thinner tails. In this case, the numerator in Equation 2.8 would be substantially larger than the denominator resulting in a very large ratio. Since this specific ratio is only one component of the sum displayed in Equation 2.8, this component would dominate the importance sampling estimate. Hence, thinner tails of the importance density run the risk of producing unstable estimates across repeated computations. In fact, the estimator may have infinite variance (e.g., Ionides, 2008; Owen & Zhou, 2000).

mixture density simplifies to the uniform distribution on  $[0, 1]$ ;<sup>4</sup> and if  $\gamma = 0$ , the beta mixture density simplifies to the beta density that provides the best fit to the posterior distribution.

In our specific example, we already know that the  $\text{Beta}(3, 9)$  density is the beta density that provides the best fit to the posterior distribution because this is the analytic expression of the posterior distribution. However, to demonstrate the general case, we show how we can find the beta distribution with the best fit to the posterior distribution using the method of moments. This particular method works as follows. First, we draw samples from our  $\text{Beta}(3, 9)$  posterior distribution and obtain:<sup>5</sup>

$$\{\theta_1^*, \theta_2^*, \dots, \theta_{12}^*\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\}.$$

Note that here we use  $\theta_i^*$  to refer to the  $i^{\text{th}}$  sample from the posterior distribution to distinguish it from the previously used  $\tilde{\theta}_i$  – the  $i^{\text{th}}$  sample from a distribution other than the posterior distribution, such as a prior distribution or an importance density. Second, we compute the mean and variance of these posterior samples. We obtain a mean of  $\bar{\theta}^* = 0.232$  and a variance of  $s_{\theta^*}^2 = 0.014$ .

Third, knowing that, if  $X \sim \text{Beta}(\alpha, \beta)$ , then  $\mathbb{E}(X) = \alpha/(\alpha + \beta)$  and  $V(X) = \alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$ , we obtain the following method of moment estimates for  $\alpha$  and  $\beta$ :

$$\hat{\alpha} = \bar{\theta}^* \left( \frac{\bar{\theta}^*(1 - \bar{\theta}^*)}{s_{\theta^*}^2} - 1 \right) = 0.232 \left( \frac{0.232(1 - 0.232)}{0.014} - 1 \right) = 2.721,$$

$$\hat{\beta} = (1 - \bar{\theta}^*) \left( \frac{\bar{\theta}^*(1 - \bar{\theta}^*)}{s_{\theta^*}^2} - 1 \right) = (1 - 0.232) \left( \frac{0.232(1 - 0.232)}{0.014} - 1 \right) = 9.006.$$

Using a mixture weight on the uniform component of  $\gamma = 0.30$  – a choice that was made to ensure that, visually, the tails of the importance density are clearly thicker than the tails of the posterior distribution – we obtain the following importance density:  $\gamma \times \text{Beta}(\theta; 1, 1) + (1 - \gamma) \times \text{Beta}(\theta; \hat{\alpha}, \hat{\beta}) = .3 + .7 \text{Beta}(\theta; 2.721, 9.006)$ . This importance density is represented by the dashed line in Figure 2.3. The figure also shows the posterior distribution (solid line). As is evident from the figure, the beta mixture importance density resembles the posterior distribution, but has fatter tails.

In general, it is advised to choose the mixture weight on the uniform component  $\gamma$  small enough to make the estimator efficient, yet large enough to produce fat tails to stabilize the estimator. A suitable mixture weight can be realized by gradually minimizing the mixture weight and investigating whether stability is still guaranteed (i.e., robustness analysis).

---

<sup>4</sup>In our running example, the importance sampling estimator then reduces to the naive Monte Carlo estimator.

<sup>5</sup>Note that, when the analytical expression of the posterior distribution is not known, posterior samples can be obtained using computer software such as WinBUGS, JAGS, or Stan, even when the marginal likelihood that functions here as a normalizing constant is not known (Equation 2.1).

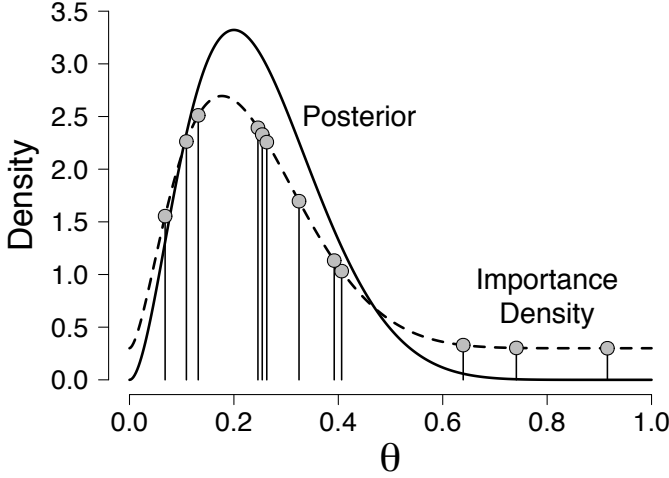


Figure 2.3: Illustration of the importance sampling estimator for the beta-binomial model. The dashed line represents our beta mixture importance density and the solid gray line represents the posterior distribution that was obtained after having observed 2 correct responses out of 10 trials. The gray dots represent the 12 samples  $\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\}$  randomly drawn from our beta mixture importance density. Available at <https://tinyurl.com/yc7ho7hr> under CC license <https://creativecommons.org/licenses/by/2.0/>.

Drawing  $N = 12$  samples for  $\theta$  from our beta mixture importance density results in:

$$\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\} = \{0.11, 0.07, 0.32, 0.25, 0.41, 0.39, 0.25, 0.13, 0.64, 0.26, 0.74, 0.92\}.$$

These samples are represented by the gray dots in Figure 2.3.

The final step is to compute the average adjusted likelihood for the 12 samples using Equation 2.8. This yields the importance sampling estimate of the marginal likelihood as:

$$\begin{aligned} \hat{p}_2(k=2 \mid n=10) &= \frac{1}{12} \sum_{i=1}^{12} \frac{p(k=2 \mid n=10, \tilde{\theta}_i) p(\tilde{\theta}_i)}{.3 + .7 \text{Beta}(\tilde{\theta}_i; 2.721, 9.006)} \\ &= \frac{1}{12} \left( \frac{\binom{10}{2} 0.11^2 (1-0.11)^8 \times 1}{.3 + .7 \text{Beta}(0.11; 2.721, 9.006)} + \dots + \frac{\binom{10}{2} 0.92^2 (1-0.92)^8 \times 1}{.3 + .7 \text{Beta}(0.92; 2.721, 9.006)} \right) \\ &= \frac{1}{12} \binom{10}{2} (0.0021 + \dots + 7.3 \times 10^{-9}) \\ &= 0.0827. \end{aligned}$$

### 2.2.4 Method 3: The Generalized Harmonic Mean Estimator of the Marginal Likelihood

Just as the importance sampling estimator, the generalized harmonic mean estimator focuses on regions of the parameter space where the integrand of Equation 2.4 is large by using an importance density  $g_{IS}(\theta)$  (Gelfand & Dey, 1994).<sup>6</sup> However, in contrast to the importance sampling estimator, the generalized harmonic mean estimator requires an importance density with thinner tails for an analogous reason as in importance sampling.

To derive the generalized harmonic mean estimator, also known as reciprocal importance sampling estimator (Frühwirth-Schnatter, 2004), we use the following identity:

$$\begin{aligned} \frac{1}{p(y)} &= \int \frac{1}{p(y)} g_{IS}(\theta) d\theta = \int \frac{p(\theta | y)}{p(y | \theta)p(\theta)} g_{IS}(\theta) d\theta = \int \frac{g_{IS}(\theta)}{p(y | \theta)p(\theta)} p(\theta | y) d\theta \\ &= \mathbb{E}_{\text{post}} \left( \frac{g_{IS}(\theta)}{p(y | \theta)p(\theta)} \right). \end{aligned}$$

Rewriting results in:

$$p(y) = \left( \mathbb{E}_{\text{post}} \left( \frac{g_{IS}(\theta)}{p(y | \theta)p(\theta)} \right) \right)^{-1},$$

which is used to define the generalized harmonic mean estimator  $\hat{p}_3(y)$  (Gelfand & Dey, 1994) as follows:

$$\hat{p}_3(y) = \left( \frac{1}{N} \sum_{j=1}^N \frac{\overbrace{g_{IS}(\theta_j^*)}^{\text{importance density}}}{\underbrace{p(y | \theta_j^*)}_{\text{likelihood}} \underbrace{p(\theta_j^*)}_{\text{prior}}} \right)^{-1}, \quad \underbrace{\theta_j^* \sim p(\theta | y)}_{\text{samples from the posterior distribution}}. \quad (2.9)$$

Note that the generalized harmonic mean estimator – in contrast to the importance sampling estimator – evaluates samples from the posterior distribution. In addition, note that the ratio in Equation 2.9 is the reciprocal of the ratio in Equation 2.8; this explains why the importance density for the generalized harmonic mean estimator should have thinner tails than the posterior distribution in order to avoid inflation of the ratios that are part of the summation displayed in Equation 2.9. Thus, in the case of the generalized harmonic mean estimator, a suitable importance density should (1) have thinner tails than the posterior distribution (DiCiccio et al., 1997; Newton & Raftery, 1994), and as in importance sampling, it should (2) be easy to evaluate; (3) have the same domain as the posterior distribution; and (4) closely resemble the posterior distribution.

---

<sup>6</sup>Note that the generalized harmonic mean estimator is a more stable version of the harmonic mean estimator (Newton & Raftery, 1994). A problem of the harmonic mean estimator is that it is dominated by the samples that have small likelihood values.

### 2.2.4.1 Running Example

To obtain the generalized harmonic mean estimate of the marginal likelihood in our running example, we need to choose a suitable importance density. In our running example, an importance density that fulfills the four above mentioned desiderata can be obtained by following four steps: First, we draw  $N = 12$  samples from the posterior distribution. Reusing the samples from the last section, we obtain:

$$\{\theta_1^*, \theta_2^*, \dots, \theta_{12}^*\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\}.$$

Second, we probit-transform all posterior samples (i.e.,  $\xi_j^* = \Phi^{-1}(\theta_j^*)$ , with  $j \in \{1, 2, \dots, 12\}$ ).<sup>7</sup> The result of this transformation is that the samples range across the entire real line instead of the  $(0, 1)$  interval only. We obtain:

$$\{\xi_1^*, \xi_2^*, \dots, \xi_{12}^*\} = \{-0.77, -0.99, -1.34, -0.39, -1.55, -0.61, -0.64, -0.23, -0.84, -0.18, -0.81, -1.17\}.$$

These probit-transformed samples are represented by the gray dots in Figure 2.4.

Third, we search for the normal distribution that provides the best fit to the probit-transformed posterior samples  $\xi_j^*$ . Using the method of moments, we obtain as estimates  $\hat{\mu} = -0.793$  and  $\hat{\sigma} = 0.423$ . Note that the choice of a normal importance density justifies step 2; the probit transformation (or an equivalent transformation) was required to match the range of the posterior distribution to the one of the normal distribution.

Finally, as importance density we choose a normal distribution with mean  $\hat{\mu} = -0.793$  and standard deviation  $\hat{\sigma} = 0.423/1.5$ . This additional division by 1.5 is to ensure thinner tails of the importance density than of the probit-transformed posterior distribution (for a discussion of alternative importance densities see Di-Ciccio et al., 1997). We decided to divide  $\hat{\sigma}$  by 1.5 for illustrative purposes only. Our importance density is displayed in Figure 2.4 (dashed line) together with the probit-transformed posterior distribution (solid line).

The generalized harmonic mean estimate can now be obtained using either the original posterior samples  $\theta_j^*$  or the probit-transformed samples  $\xi_j^*$ . Here we use the latter ones (see also Overstall & Forster, 2010). Incorporating our specific importance density and a correction for having used the probit-transformation, Equation 2.9 becomes:<sup>8</sup>

$$\hat{p}_3(y) = \left( \frac{1}{N} \sum_{j=1}^N \frac{\overbrace{\frac{1}{\hat{\sigma}} \phi\left(\frac{\xi_j^* - \hat{\mu}}{\hat{\sigma}}\right)}^{\text{importance density}}}{\underbrace{p(y | \Phi(\xi_j^*))}_{\text{likelihood}} \underbrace{\phi(\xi_j^*)}_{\text{prior}}} \right)^{-1}, \quad \underbrace{\xi_j^* = \Phi^{-1}(\theta_j^*) \text{ and } \theta_j^* \sim p(\theta | y)}_{\text{probit-transformed samples from the posterior distribution}}. \quad (2.10)$$

<sup>7</sup>Other transformation are conceivable (e.g., logit transformation).

<sup>8</sup>A detailed explanation is provided in the appendix. Note that using the original posterior samples  $\theta_j^*$  would involve transforming the importance density (e.g., the normal density on  $\xi$ ) to the  $(0, 1)$  interval.

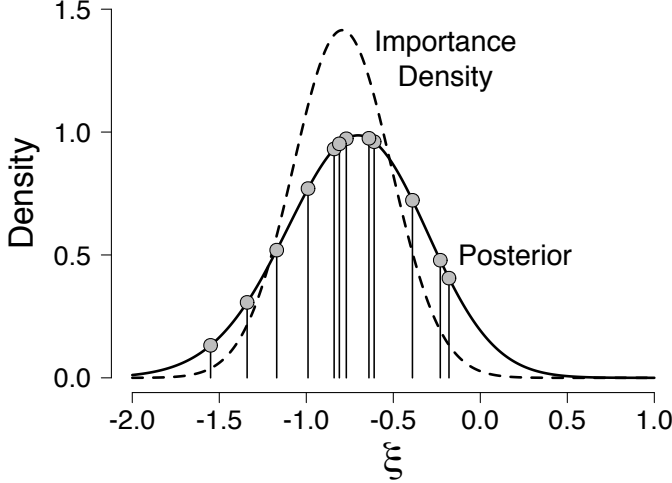


Figure 2.4: Illustration of the generalized harmonic mean estimator for the beta-binomial model. The solid line represents the probit-transformed Beta(3,9) posterior distribution that was obtained after having observed 2 correct responses out of 10 trials, and the dashed line represents the importance density  $\mathcal{N}(\xi; \mu = -0.793, \sigma = 0.423/1.5)$ . The gray dots represent the 12 probit-transformed samples  $\{\xi_1^*, \xi_2^*, \dots, \xi_{12}^*\}$  randomly drawn from the Beta(3,9) posterior distribution. Available at <https://tinyurl.com/yazgk8kj> under CC license <https://creativecommons.org/licenses/by/2.0/>.

For our beta-binomial model, we now obtain the generalized harmonic mean estimate of the marginal likelihood as:

$$\begin{aligned}
 \hat{p}_3(k=2 \mid n=10) &= \left( \frac{1}{12} \sum_{j=1}^{12} \frac{\frac{1}{0.423/1.5} \phi\left(\frac{\xi_j^* + 0.793}{0.423/1.5}\right)}{p(k=2 \mid n=10, \Phi(\xi_j^*)) \phi(\xi_j^*)} \right)^{-1} \\
 &= \left( \frac{1}{12} \left( \frac{\frac{1}{0.423/1.5} \phi\left(\frac{-0.77 + 0.793}{0.423/1.5}\right)}{\binom{10}{2} 0.22^2 (1-0.22)^8 \phi(-0.77)} + \dots + \frac{\frac{1}{0.423/1.5} \phi\left(\frac{-1.17 + 0.793}{0.423/1.5}\right)}{\binom{10}{2} 0.12^2 (1-0.12)^8 \phi(-1.17)} \right) \right)^{-1} \\
 &= \left( \frac{1}{12} \frac{1}{\binom{10}{2}} (716.89 + \dots + 555.50) \right)^{-1} \\
 &= 0.092.
 \end{aligned}$$

### 2.2.5 Method 4: The Bridge Sampling Estimator of the Marginal Likelihood

As became evident in the last two sections, both the importance sampling estimator and the generalized harmonic mean estimator impose strong constraints on the tail behavior of the importance density relative to the posterior distribution to guarantee a stable estimator. Such requirements can make it difficult to find a suitable importance density, especially when a high-dimensional posterior is considered. The bridge sampler, on the other hand, alleviates such requirements (e.g., Frühwirth-Schnatter, 2004).

Originally, bridge sampling was developed to directly estimate the Bayes factor, that is, the ratio of the marginal likelihoods of two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  (e.g., Jeffreys, 1961; Kass & Raftery, 1995). However, in this tutorial, we use a version of bridge sampling that allows us to approximate the marginal likelihood of a *single* model (for an earlier application see for example Overstall & Forster, 2010). This version is based on the following identity:

$$1 = \frac{\int p(y | \theta) p(\theta) h(\theta) g(\theta) d\theta}{\int p(y | \theta) p(\theta) h(\theta) g(\theta) d\theta}, \quad (2.11)$$

where  $g(\theta)$  is the so-called proposal distribution and  $h(\theta)$  the so-called bridge function. Multiplying both sides of Equation 2.11 by the marginal likelihood  $p(y)$  results in:

$$\begin{aligned} p(y) &= \frac{\int p(y | \theta) p(\theta) h(\theta) g(\theta) d\theta}{\int \frac{p(y | \theta) p(\theta)}{p(y)} h(\theta) g(\theta) d\theta} = \frac{\int p(y | \theta) p(\theta) h(\theta) \overbrace{g(\theta)}^{\text{proposal distribution}} d\theta}{\int h(\theta) g(\theta) \underbrace{p(\theta | y)}_{\text{posterior distribution}} d\theta} \\ &= \frac{\mathbb{E}_{g(\theta)} [p(y | \theta) p(\theta) h(\theta)]}{\mathbb{E}_{\text{post}} [h(\theta) g(\theta)]}. \end{aligned}$$

The marginal likelihood can now be approximated using:

$$\hat{p}(y) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(y | \tilde{\theta}_i) p(\tilde{\theta}_i) h(\tilde{\theta}_i)}{\frac{1}{N_1} \sum_{j=1}^{N_1} h(\theta_j^*) g(\theta_j^*)}, \quad \underbrace{\tilde{\theta}_i \sim g(\theta)}_{\text{samples from the proposal distribution}}, \quad \underbrace{\theta_j^* \sim p(\theta | y)}_{\text{samples from the posterior distribution}}. \quad (2.12)$$

Equation 2.12 illustrates that we need samples from both the proposal distribution and the posterior distribution to obtain the bridge sampling estimate for the marginal likelihood. However, before we can apply Equation 2.12 to our running example, we have to discuss how we can obtain a suitable proposal distribution and bridge function. Conceptually, the proposal distribution is similar to an importance density, should resemble the posterior distribution, and should

have sufficient overlap with the posterior distribution. According to Overstall and Forster (2010), a convenient proposal distribution is often a normal distribution with its first two moments chosen to match those of the posterior distribution. In our experience, this choice for the proposal distribution works well for a wide range of scenarios. However, this proposal distribution might produce unstable estimates in case of high-dimensional posterior distributions that clearly do not follow a multivariate normal distribution. In such a situation, it might be advisable to consider more sophisticated versions of bridge sampling (e.g., Frühwirth-Schnatter, 2004; Meng & Schilling, 2002; L. Wang & Meng, 2016).

### 2.2.5.1 Choosing the Optimal Bridge Function

In this tutorial we use the bridge function defined as (Meng & Wong, 1996):

$$h(\theta) = C \cdot \frac{1}{s_1 p(y | \theta) p(\theta) + s_2 p(y) g(\theta)} , \quad (2.13)$$

where  $s_1 = \frac{N_1}{N_2 + N_1}$ ,  $s_2 = \frac{N_2}{N_2 + N_1}$ , and  $C$  a constant; its particular value is not required because  $h(\theta)$  is part of both the numerator and the denominator of Equation 2.12, and therefore the constant  $C$  cancels. This particular bridge function is referred to as the “optimal bridge function” because Meng and Wong (1996, p. 837) proved that it minimizes the relative mean-squared error (Equation 2.16).

Equation 2.13 shows that the optimal bridge function depends on the marginal likelihood  $p(y)$  which is the very entity we want to approximate. We can resolve this issue by applying an iterative scheme that updates an initial guess of the marginal likelihood until the estimate of the marginal likelihood has converged according to a predefined tolerance level. To do so, we insert the expression for the optimal bridge function (Equation 2.13) in Equation 2.12 (Meng & Wong, 1996). The formula to approximate the marginal likelihood on iteration  $t + 1$  is then specified as follows:

$$\hat{p}(y)^{(t+1)} = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{s_1 p(y | \tilde{\theta}_i) p(\tilde{\theta}_i) + s_2 \hat{p}(y)^{(t)} g(\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{s_1 p(y | \theta_j^*) p(\theta_j^*) + s_2 \hat{p}(y)^{(t)} g(\theta_j^*)}} , \quad (2.14)$$

$$\underbrace{\tilde{\theta}_i \sim g(\theta)}_{\text{samples from the proposal distribution}} , \quad \underbrace{\theta_j^* \sim p(\theta | y)}_{\text{samples from the posterior distribution}} ,$$

where  $\hat{p}(y)^{(t)}$  denotes the estimate of the marginal likelihood on iteration  $t$  of the iterative scheme. Note that Equation 2.14 illustrates why bridge sampling is robust to the tail behavior of the proposal distribution relative to the posterior distribution; the difference to the importance sampling and generalized harmonic mean estimator is that, in the case of the bridge sampling estimator, samples from

the tail region cannot inflate individual summation terms and thus dominate the estimate. To illustrate this, we consider what happens to the bridge sampling estimator, the importance sampling estimator, and the generalized harmonic mean estimator in case (1) the proposal/importance distribution has fatter tails than the posterior distribution, and (2) the proposal/importance distribution has thinner tails than the posterior distribution (see also Frühwirth–Schnatter, 2004). Specifically, we look at a single term in the respective sums and consider the limit of that term as we move further and further out in the tails. This is insightful since a single term can have a lasting effect on the estimator (e.g., in case a single term in a sum is very large or even infinite).

In case (1) (i.e., the proposal/importance distribution has fatter tails than the posterior), the ratio in the importance sampling estimator (i.e., Equation 2.8) goes to zero as we move further out in the tails. Since samples in the tails may only be obtained occasionally and a zero term in the sum does not inflate the estimate this is not a reason for concern. In contrast, when we consider the ratio in the generalized harmonic mean estimator (i.e., Equation 2.9), we see that the ratio goes to infinity as we move further out in the tails. Even if this occurs only very rarely, this is an issue since the resulting value will dominate the estimate. Consequently, the resulting estimator may have a large variance since samples from the tail regions may be obtained only occasionally across repeated applications. For the bridge sampling estimator (i.e., Equation 2.14), we need to consider the ratio in the numerator and denominator. The ratio in the numerator will go to zero and the ratio in the denominator will go to  $\frac{1}{s_2 \hat{p}(y)^{(t)}}$ . Hence, both of these ratios are bounded and will not inflate the two sums, hence also not the resulting estimate.

In case (2) (i.e., the proposal/importance distribution has thinner tails than the posterior), the ratio in the importance sampling estimator (i.e., Equation 2.8) goes to infinity as we move further out in the tails, inflating the estimate. In contrast, when we consider the ratio in the generalized harmonic mean estimator (i.e., Equation 2.9), we see that the ratio goes to zero. As explained above, this is not a reason for concern. These considerations explain why in importance sampling, the importance distribution should have fatter tails than the posterior whereas for the generalized harmonic mean estimator, it should have thinner tails. For the bridge sampling estimator (i.e., Equation 2.14), the ratio in the numerator will go to  $1/s_1$  and the ratio in the denominator will go to zero. Again, both of these ratios are bounded making the bridge sampling estimator more robust to the tail behavior than the other two estimators. This of course assumes that not all terms in the denominator (for case (2)) and the numerator (for case (1)) will be zero, that is, the proposal and the posterior distribution have sufficient overlap. In the extreme scenario of no overlap the bridge sampling estimate is not defined because both sums of Equation 2.14 would be zero.

Extending the numerator of the right side of Equation 2.14 with  $\frac{1/g(\tilde{\theta}_i)}{1/g(\theta_i^*)}$ , and the denominator with  $\frac{1/g(\theta_j^*)}{1/g(\tilde{\theta}_j)}$ , and subsequently defining  $l_{1,j} := \frac{p(y|\theta_j^*)p(\theta_j^*)}{g(\theta_j^*)}$  and  $l_{2,i} := \frac{p(y|\tilde{\theta}_i)p(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}$ , we obtain the formula for the iterative scheme of the bridge

sampling estimator  $\hat{p}_4(y)^{(t+1)}$  at iteration  $t + 1$  (Meng & Wong, 1996, p. 837):

$$\begin{aligned}
 \hat{p}_4(y)^{(t+1)} &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{s_1 p(y | \tilde{\theta}_i) p(\tilde{\theta}_i) + s_2 \hat{p}_4(y)^{(t)} g(\tilde{\theta}_i)} \frac{1/g(\tilde{\theta}_i)}{1/g(\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{s_1 p(y | \theta_j^*) p(\theta_j^*) + s_2 \hat{p}_4(y)^{(t)} g(\theta_j^*)} \frac{1/g(\theta_j^*)}{1/g(\theta_j^*)}} \\
 &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}_4(y)^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}_4(y)^{(t)}}}, \quad \underbrace{\tilde{\theta}_i \sim g(\theta)}_{\text{samples from the proposal distribution}}, \quad \underbrace{\theta_j^* \sim p(\theta | y)}_{\text{samples from the posterior distribution}}.
 \end{aligned} \tag{2.15}$$

Equation 2.15 suggests that, in order to obtain the bridge sampling estimate of the marginal likelihood, a number of requirements need to be fulfilled. First, we need  $N_2$  samples from the proposal distribution  $g(\theta)$  and  $N_1$  samples from the posterior distribution  $p(\theta | y)$ . Second, for all  $N_2$  samples from the proposal distribution, we have to evaluate  $l_{2,i}$ . This involves obtaining the value of the unnormalized posterior (i.e., the product of the likelihood times the prior) and of the proposal distribution for all samples. Third, we evaluate  $l_{1,j}$  for all  $N_1$  samples from the posterior distribution. This is analogous to evaluating  $l_{2,i}$ . Fourth, we have to determine the constants  $s_1$  and  $s_2$  that only depend on  $N_1$  and  $N_2$ . Fifth, we need an initial guess of the marginal likelihood  $\hat{p}_4(y)$ . Since some of these five requirements can be obtained easier than others, we will point out possible challenges.

A first challenge is that using a suitable proposal distribution may involve transforming the posterior samples. Consequently, we have to determine how the transformation affects the definition of the bridge sampling estimator for the marginal likelihood (Equation 2.15).

A second challenge is how to use the  $N_1$  samples from the posterior distribution. One option is to use all  $N_1$  samples for both fitting the proposal distribution and for computing the bridge sampling estimate. However, Overstall and Forster (2010) showed that such a procedure may result in an underestimation of the marginal likelihood. To obtain more reliable estimates they propose to divide the posterior samples in two parts; the first part is used to obtain the best-fitting proposal distribution, and the second part is used to compute the bridge sampling estimate. Throughout this tutorial, we use two equally large parts. In the remainder we therefore state that we draw  $2N_1$  samples from the posterior distribution. The first  $N_1$  of the total of  $2N_1$  samples are used for fitting the proposal distribution and the remaining  $N_1$  samples are used in the iterative scheme (i.e., Equation 2.15).<sup>9</sup>

---

<sup>9</sup>In case the posterior samples are obtained via MCMC sampling using multiple chains, we

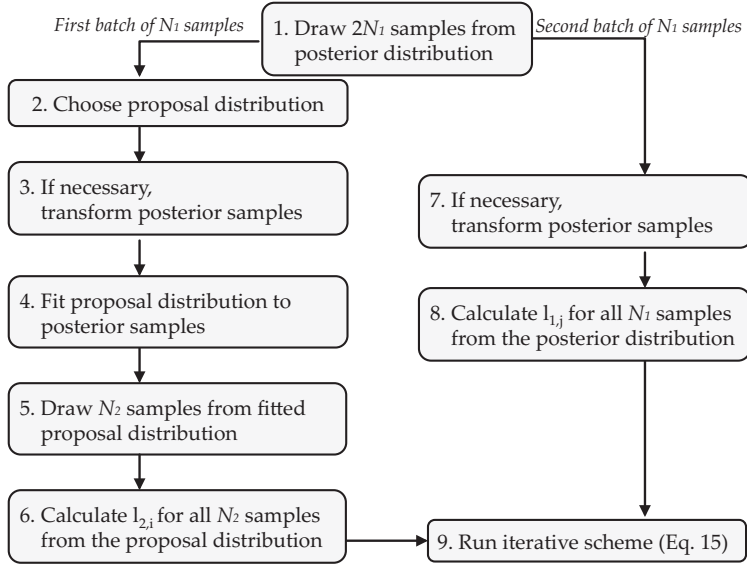


Figure 2.5: Schematic illustration of the steps involved in obtaining the bridge sampling estimate of the marginal likelihood. Available at <https://tinyurl.com/y7b2kze7> under CC license <https://creativecommons.org/licenses/by/2.0/>.

To summarize, the discussion of the requirements and challenges encountered in bridge sampling illustrated that the bridge sampling estimator imposes less strict requirements on the proposal distribution than the importance sampling and generalized harmonic mean estimator and allows for an almost automatic application due to the default choice of the bridge function.<sup>10</sup>

### 2.2.5.2 Running Example

To obtain the bridge sampling estimate of the marginal likelihood in the beta-binomial example, we follow the eight steps illustrated in Figure 2.5:

1. We draw  $2N_1 = 24$  samples from the  $\text{Beta}(3, 9)$  posterior distribution for  $\theta$ . We obtain the following sample of 24 values:

$$\{\theta_1^*, \theta_2^*, \dots, \theta_{24}^*\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12, 0.15, 0.21, 0.24, 0.18, 0.12, 0.22, 0.15, 0.22, 0.23, 0.26, 0.29, 0.28\}.$$

Note that the first 12 samples equal the ones used in the last section, whereas the last 12 samples were obtained from drawing again 12 values from the  $\text{Beta}(3, 9)$  posterior distribution for  $\theta$ .

use the first half of the iterations per chain for fitting the proposal distribution and the second half of the iterations per chain for the iterative scheme.

<sup>10</sup>For an explanation of where the name “bridge” comes from see <https://osf.io/9jzm3/>.

2. *We choose a proposal distribution.*

Here we opt for an approach that can be easily generalized to models with multiple parameters and select a normal distribution as the proposal distribution  $g(\theta)$ .<sup>11</sup>

3. *We transform the first batch of  $N_1$  posterior samples.*

Since we use a normal proposal distribution, we have to transform the posterior samples from the rate scale to the real line so that the range of the posterior distribution matches the range of the proposal distribution. This can be achieved by probit-transforming the posterior samples, that is,  $\xi_j^* = \Phi^{-1}(\theta_j^*)$  with  $j \in \{1, 2, \dots, 12\}$ . We obtain:

$$\{\xi_1^*, \xi_2^*, \dots, \xi_{12}^*\} = \{-0.77, -0.99, -1.34, -0.39, -1.55, -0.61, -0.64, -0.23, \\ -0.84, -0.18, -0.81, -1.17\}.$$

4. *We fit the proposal distribution to the first batch of  $N_1$  probit-transformed posterior samples.*

We use the method of moment estimates  $\hat{\mu} = -0.793$  and  $\hat{\sigma} = 0.423$  from the first batch of  $N_1$  probit-transformed posterior samples to obtain our proposal distribution  $g(\xi; \mu = -0.793, \sigma = 0.423) = \frac{1}{0.423} \phi\left(\frac{\xi + 0.793}{0.423}\right)$ .

5. *We draw  $N_2$  samples from the proposal distribution.*

We obtain:

$$\{\tilde{\xi}_1, \tilde{\xi}_2, \dots, \tilde{\xi}_{12}\} = \{-1.11, -0.63, -1.48, -0.59, -0.48, -0.69, -0.74, -0.51, \\ -0.82, -1.54, -0.76, -0.96\}.$$

6. *We calculate  $l_{2,i}$  for all  $N_2$  samples from the proposal distribution.*

This step involves assessing the value of the unnormalized posterior and the proposal distribution for all  $N_2$  samples from the proposal distribution. As in the running example for the generalized harmonic mean estimator, we obtain the unnormalized posterior as:  $p\left(k = 2 \mid n = 10, \Phi\left(\tilde{\xi}_i\right)\right) \phi\left(\tilde{\xi}_i\right)$ , where  $\phi\left(\tilde{\xi}_i\right)$  comes from using the change-of-variable method (see running example for the generalized harmonic mean estimator and the appendix for details). Thus, as in the case of the generalized harmonic mean estimator, the uniform prior on  $\theta$  translates to a standard normal prior on  $\xi$ . The values of the proposal distribution can easily be obtained (for example using the R software).

7. *We transform the second batch of  $N_1$  posterior samples.*

As in step 2, we use the probit transformation and obtain:

$$\{\xi_{13}^*, \xi_{14}^*, \dots, \xi_{24}^*\} = \{-1.04, -0.81, -0.71, -0.92, -1.17, -0.77, -1.04, -0.77, \\ -0.74, -0.64, -0.55, -0.58\}.$$

---

<sup>11</sup>There exist several candidates for the proposal distribution. Alternative proposal distributions are, for example, the importance density that we used for the importance sampling estimator or for the generalized harmonic mean estimator, or the analytically derived Beta(3, 9) posterior distribution.

8. We calculate  $l_{1,j}$  for the second batch of  $N_1$  probit-transformed samples from the posterior distribution.

This is analogous to step 6.

9. We run the iterative scheme (Equation 2.15) until our predefined tolerance criterion is reached.

As tolerance criterion we choose  $|\hat{p}_4(k=2 | n=10)^{(t+1)} - \hat{p}_4(k=2 | n=10)^{(t)}| / \hat{p}_4(k=2 | n=10)^{(t+1)} \leq 10^{-10}$ . This requires an initial guess for the marginal likelihood  $\hat{p}_4(k=2 | n=10)^{(0)}$  which we set to 0.<sup>12</sup>

The simplicity of the beta-binomial model allows us to calculate the bridge sampling estimate by hand. To determine  $\hat{p}_4(y)^{(t+1)}$  according to Equation 2.15, we need to calculate the constants  $s_1$  and  $s_2$ . Since  $N_1 = N_2 = 12$ , we obtain:  $s_1 = s_2 = N_2 / (N_2 + N_1) = 0.5$ . In addition, we need to calculate  $l_{2,i}$  ( $i \in \{1, 2, \dots, 12\}$ ) for all samples from the proposal distribution, and  $l_{1,j}$  ( $j \in \{1, 2, \dots, 12\}$ ) for the second batch of the probit-transformed samples from the posterior distribution. Here we show how to calculate  $l_{2,1}$  and  $l_{1,1}$  using the first sample from the proposal distribution and the first sample of the second batch of the posterior samples, respectively:

$$l_{2,1} = \frac{p(k | n, \Phi(\tilde{\xi}_1))\phi(\tilde{\xi}_1)}{g(\tilde{\xi}_1)} = \left( \frac{\binom{10}{2} 0.13^2 (1 - 0.13)^8 \cdot 0.22}{\frac{1}{0.423} \phi\left(\frac{-1.11 + 0.793}{0.423}\right)} \right) = 0.077,$$

$$l_{1,1} = \frac{p(k | n, \Phi(\xi_{13}^*))\phi(\xi_{13}^*)}{g(\xi_{13}^*)} = \left( \frac{\binom{10}{2} 0.15^2 (1 - 0.15)^8 \cdot 0.23}{\frac{1}{0.423} \phi\left(\frac{-1.04 + 0.793}{0.423}\right)} \right) = 0.080.$$

For  $\hat{p}_4(k=2 | n=10)^{(t+1)}$ , we then get:

$$\begin{aligned} \hat{p}_4(k=2 | n=10)^{(t+1)} &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}_4(k=2 | n=10)^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}_4(k=2 | n=10)^{(t)}}} \\ &= \frac{\frac{1}{12} \left( \frac{0.077}{0.5 \cdot 0.077 + 0.5 \cdot \hat{p}_4(k=2 | n=10)^{(t)}} + \dots + \frac{0.084}{0.5 \cdot 0.084 + 0.5 \cdot \hat{p}_4(k=2 | n=10)^{(t)}} \right)}{\frac{1}{12} \left( \frac{1}{0.5 \cdot 0.080 + 0.5 \cdot \hat{p}_4(k=2 | n=10)^{(t)}} + \dots + \frac{1}{0.5 \cdot 0.103 + 0.5 \cdot \hat{p}_4(k=2 | n=10)^{(t)}} \right)}. \end{aligned}$$

Using  $\hat{p}(y)^{(0)} = 0$ , we obtain as updated estimate of the marginal likelihood  $\hat{p}_4(k=2 | n=10)^{(1)} = 0.0908$ . This iterative procedure has to be repeated until

<sup>12</sup>A better initial guess can be obtained from, for example, the importance sampling estimator or the generalized harmonic mean estimator explained in the previous sections. In our experience, however, usually the exact choice of the initial value does not seem to influence the convergence of the bridge sampler much.

our predefined tolerance criterion is reached. For our running example, this criterion is reached after five iterations. We now obtain the bridge sampling estimate of the marginal likelihood as  $\hat{p}_4(k = 2 \mid n = 10)^{(5)} = 0.0902$ .

### 2.2.6 Interim Summary

So far we used the beta-binomial model to illustrate the computation of four different estimators of the marginal likelihood. These four estimators were discussed in order of increasing sophistication, such that the first three estimators provided the proper context for understanding the fourth, most general estimator – the bridge sampler. This estimator is the focus in the remainder of this tutorial. The goal of the next sections is to demonstrate that bridge sampling is particularly suitable to estimate the marginal likelihood of popular models in mathematical psychology. Importantly, bridge sampling may be used to obtain accurate estimates of the marginal likelihood of hierarchical models (for a detailed comparison of bridge sampling versus its special cases see Frühwirth–Schnatter, 2004; Sinharay & Stern, 2005).

### 2.2.7 Assessing the Accuracy of the Bridge Sampling Estimate

In this section we show how to quantify the accuracy of the bridge sampling estimate. A straightforward approach would be to apply the bridge sampling procedure multiple times and investigate the variability of the marginal likelihood estimate. In practice, however, this solution is often impractical due to the substantial computational burden of obtaining the posterior samples and evaluating the relevant quantities in the bridge sampling procedure.

Frühwirth–Schnatter (2004) proposed an alternative approach that approximates the estimator’s expected relative mean-squared error:

$$RE^2 = \frac{\mathbb{E} \left[ (\hat{p}_4(y) - p(y))^2 \right]}{p(y)^2}. \quad (2.16)$$

The derivation of this approximate relative mean-squared error by Frühwirth–Schnatter takes into account that the samples from the proposal distribution  $g(\theta)$  are independent, whereas the MCMC samples from the posterior distribution  $p(\theta \mid y)$  may be autocorrelated. The approximate relative mean-squared error is given by:

$$\widehat{RE}^2 = \frac{1}{N_2} \frac{V_{g(\theta)}(f_1(\theta))}{\mathbb{E}_{g(\theta)}^2(f_1(\theta))} + \frac{\rho_{f_2}(0)}{N_1} \frac{V_{\text{post}}(f_2(\theta))}{\mathbb{E}_{\text{post}}^2(f_2(\theta))}, \quad (2.17)$$

where  $f_1(\theta) = \frac{p(\theta|y)}{s_1 p(\theta|y) + s_2 g(\theta)}$ ,  $f_2(\theta) = \frac{g(\theta)}{s_1 p(\theta|y) + s_2 g(\theta)}$ ,  $V_{g(\theta)}(f_1(\theta)) = \int (f_1(\theta) - \mathbb{E}[f_1(\theta)])^2 g(\theta) d\theta$  denotes the variance of  $f_1(\theta)$  with respect to the proposal distribution  $g(\theta)$  (the variance  $V_{\text{post}}(f_2(\theta))$  is defined analogously), and  $\rho_{f_2}(0)$  corresponds to the normalized spectral density of the autocorrelated process  $f_2(\theta)$  at the frequency 0.

In practice, we approximate the unknown variances and expected values by the corresponding sample variances and means. Hence, for evaluating the variance and

expected value with respect to  $g(\theta)$ , we use the  $N_2$  samples for  $\tilde{\theta}_i$  from the proposal distribution. To evaluate the variance and expected value with respect to the posterior distribution, we use the second batch of  $N_1$  samples  $\theta_j^*$  from the posterior distribution which we also use in the iterative scheme for computing the marginal likelihood. Because the posterior samples are obtained via an MCMC procedure and are hence autocorrelated, the second term in Equation 2.17 is adjusted by the normalized spectral density (for details see Frühwirth–Schnatter, 2004).<sup>13</sup> To evaluate the normalized posterior density which appears in the numerator of  $f_1(\theta)$  and the denominator of both  $f_1(\theta)$  and  $f_2(\theta)$ , we use the bridge sampling estimate as normalizing constant.

Note that, under the assumption that the bridge sampling estimator  $\hat{p}_4(y)$  is an unbiased estimator of the marginal likelihood  $p(y)$ , the square root of the relative mean-squared error (Equation 2.16) can be interpreted as the coefficient of variation (i.e., the ratio of the standard deviation and the mean; C. E. Brown, 1998). In the remainder of this chapter, we report the coefficient of variation to quantify the accuracy of the bridge sampling estimate.

## 2.3 Case Study: Bridge Sampling for Reinforcement Learning Models

In this section, we illustrate the computation of the marginal likelihood using bridge sampling in the context of a published data set (Busemeyer & Stout, 2002) featuring the Expectancy Valence (EV) model – a popular reinforcement learning (RL) model for the Iowa gambling task (IGT; Bechara et al., 1994). We first introduce the task and the model, and then use bridge sampling to estimate the marginal likelihood of the EV model implemented in both an individual-level and a hierarchical Bayesian framework. For the individual-level framework, we compare estimates obtained from bridge sampling to importance sampling estimates published in Steingroever, Wetzels, and Wagenmakers (2016). For the hierarchical framework, we compare our results to estimates from the Savage-Dickey density ratio test (Dickey, 1971; Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Grasman, & Wagenmakers, 2010).

### 2.3.1 The Iowa Gambling Task

In this section we describe the IGT (see also Steingroever, Pachur, Šmíra, & Lee, 2018; Steingroever, Wetzels, Horstmann, Neumann, & Wagenmakers, 2013; Steingroever, Wetzels, & Wagenmakers, 2013a, 2013b, 2014; Steingroever et al., 2016). Originally, Bechara et al. (1994) developed the IGT to distinguish decision-making strategies of patients with lesions to the ventromedial prefrontal cortex from the ones of healthy controls (see also Bechara, Damasio, Damasio, & Lee, 1999; Bechara, Damasio, Tranel, & Anderson, 1998; Bechara, Tranel, & Damasio, 2000). During the last decades, the scope of application of the IGT has

<sup>13</sup>We estimate the spectral density at frequency zero by fitting an autoregressive model using the `spectrum0.ar()` function as implemented in the `coda` R package (Plummer, Best, Cowles, & Vines, 2006).

Table 2.1: Summary of the payoff scheme of the traditional IGT as developed by Bechara et al. (1994).

	Deck A	Deck B	Deck C	Deck D
	Bad deck with fre- quent losses	Bad deck with infre- quent losses	Good deck with fre- quent losses	Good deck with infre- quent losses
Reward/trial	100	100	50	50
Number of losses/10 cards	5	1	5	1
Loss/10 cards	-1250	-1250	-250	-250
Net outcome/10 cards	-250	-250	250	250

increased tremendously covering clinical populations with, for example, pathological gambling (Cavedini, Riboldi, Keller, D’Annuncci, & Bellodi, 2002), obsessive-compulsive disorder (Cavedini, Riboldi, D’Annuncci, et al., 2002), psychopathic tendencies (Blair, Colledge, & Mitchell, 2001), and schizophrenia (Bark, Dieckmann, Bogerts, & Northoff, 2005; Martino, Bucay, Butman, & Allegrì, 2007).

The IGT is a card game that requires participants to choose, over several rounds, cards from four different decks in order to maximize their long-term net outcome (Bechara et al., 1994; Bechara, Damasio, Tranel, & Damasio, 1997). The four decks differ in their payoffs, and two of them result in negative long-term outcomes (i.e., the bad decks), whereas the remaining two decks result in positive long-term outcomes (i.e., the good decks). After each choice, participants receive feedback on the rewards and losses (if any) associated with that card, as well as their running tally of net outcomes over all trials so far. Unbeknownst to the participants, the task (typically) contains 100 trials.

A crucial aspect of the IGT is whether and to what extent participants eventually learn to prefer the good decks because only choosing from the good decks maximizes their long-term net outcome. The good decks are typically labeled as decks C and D, whereas the bad decks are labeled as decks A and B. Table 2.1 presents a summary of the traditional payoff scheme as developed by Bechara et al. (1994). This table illustrates that decks A and B yield high constant rewards, but even higher unpredictable losses: hence, the long-term net outcome is negative. Decks C and D, on the other hand, yield low constant rewards, but even lower unpredictable losses: hence, the long-term net outcome is positive. In addition to the different payoff magnitudes, the decks also differ in the frequency of losses: decks A and C yield frequent losses, while decks B and D yield infrequent losses.

### 2.3.2 The Expectancy Valence Model

In this section, we describe the EV model (see also Steingroever et al., 2018; Steingroever, Wetzels, & Wagenmakers, 2013a; Steingroever et al., 2014, 2016). Originally proposed by Busemeyer and Stout (2002), the EV model is arguably the most popular model for the IGT (for references see Steingroever, Wetzels, & Wagenmakers, 2013a, and for alternative IGT models see Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Dai, Kerestes, Upton, Busemeyer, & Stout, 2015; Stein-

groever et al., 2014; Worthy & Maddox, 2014; Worthy, Pang, & Byrne, 2013). The model formalizes participants' performance on the IGT through the interaction of three model parameters that represent distinct psychological processes. The first model assumption is that after choosing a card from deck  $k$ ,  $k \in \{1, 2, 3, 4\}$ , on trial  $t$ , participants compute a weighted mean of the experienced reward  $W(t)$  and loss  $L(t)$  to obtain the utility of deck  $k$  on trial  $t$ ,  $v_k(t)$ :

$$v_k(t) = (1 - w)W(t) + wL(t).$$

The weight that participants assign to losses relative to rewards is the attention weight parameter  $w$ . A small value of  $w$ , that is,  $w < .5$ , is characteristic for decision makers who put more weight on the immediate rewards and can thus be described as reward-seeking, whereas a large value of  $w$ , that is,  $w > .5$ , is characteristic for decision makers who put more weight on the immediate losses and can thus be described as loss-averse (Ahn et al., 2008; Bussemeyer & Stout, 2002).

The EV model further assumes that decision makers use the utility of deck  $k$  on trial  $t$ ,  $v_k(t)$ , to update only the expected utility of deck  $k$ ,  $Ev_k(t)$ ; the expected utilities of the unchosen decks are left unchanged. This updating process is described by the Delta learning rule, also known as the Rescorla-Wagner rule (Rescorla & Wagner, 1972):

$$Ev_k(t) = Ev_k(t - 1) + a(v_k(t) - Ev_k(t - 1)).$$

If the experienced utility  $v_k(t)$  is higher than expected, the expected utility of deck  $k$  is adjusted upward. If the experienced utility  $v_k(t)$  is lower than expected, the expected utility of deck  $k$  is adjusted downward. This updating process is influenced by the second model parameter – the updating parameter  $a$ . This parameter quantifies the memory for rewards and losses. A value of  $a$  close to zero indicates slow forgetting and weak recency effects, whereas a value of  $a$  close to one indicates rapid forgetting and strong recency effects. For all models, we initialized the expectancies of all decks to zero,  $Ev_k(0) = 0$  ( $k \in \{1, 2, 3, 4\}$ ). This setting reflects neutral prior knowledge about the payoffs of the decks.

In the next step, the model assumes that the expected utilities of each deck guide participants' choices on the next trial  $t + 1$ . This assumption is formalized by the softmax choice rule, also known as the ratio-of-strength choice rule (Luce, 1959):

$$Pr[S_k(t + 1)] = \frac{e^{\theta(t) \cdot Ev_k(t)}}{\sum_{j=1}^4 e^{\theta(t) \cdot Ev_j(t)}}.$$

The EV model uses this rule to compute the probability of choosing each deck on each trial. This rule contains a sensitivity parameter  $\theta$  that indexes the extent to which trial-by-trial choices match the expected deck utilities. Values of  $\theta$  close to zero indicate random choice behavior (i.e., strong exploration), whereas large values of  $\theta$  indicate choice behavior that is strongly determined by the expected utilities (i.e., strong exploitation). The EV model uses a trial-dependent sensitivity parameter  $\theta(t)$ , which also depends on the final model parameter, response consistency  $c' \in [-5, 5]$ :

$$\theta(t) = (t/10)^{c'}.$$

If  $c'$  is positive, successive choices become less random and more determined by the expected deck utilities; if  $c'$  is negative, successive choices become more random and less determined by the expected deck utilities, a pattern that is clearly non-optimal. We restricted the consistency parameter of the EV model to the range  $[-2, 2]$  instead of the proposed range  $[-5, 5]$  (Busemeyer & Stout, 2002). This modification improved the estimation of the EV model and prevented the choice rule from producing numbers that exceed machine precision (see also Steingroever et al., 2014).

In sum, the EV model has three parameters: (1) the attention weight parameter  $w \in [0, 1]$ , which quantifies the weight of losses over rewards; (2) the updating parameter  $a \in [0, 1]$ , which determines the memory for past expectancies; and (3) the response consistency parameter  $c' \in [-2, 2]$ , which determines the balance between exploitation and exploration.

### 2.3.3 Data

We applied bridge sampling to a data set published by Busemeyer and Stout (2002). The data set consists of 30 healthy participants each contributing  $T = 100$  IGT card selections (see Busemeyer and Stout for more details on the data sets).<sup>14</sup>

### 2.3.4 Application of Bridge Sampling to an Individual-Level Implementation of the EV Model

In this section we describe how we use bridge sampling to estimate the marginal likelihood of an individual-level implementation of the EV model. This implementation estimates model parameters for each participant separately. Accordingly, we also obtain a marginal likelihood of the EV model for every participant.

#### 2.3.4.1 Schematic Execution of the Bridge Sampler

To obtain the bridge sampling estimate of the marginal likelihood for each participant, we follow the steps outlined in Figure 2.5.

For each participant  $s$ ,  $s \in \{1, 2, \dots, 30\}$ , we proceed as follows:

1. *For each parameter, we draw  $2N_1$  samples from the posterior distribution.* Since Steingroever et al. (2016) already fit an individual-level implementation of the EV model separately to the data of each participant in Busemeyer and Stout (2002), we reuse their posterior samples (see Steingroever et al., 2016, for details on the prior distributions and model implementation). Note that they parameterized the model not in terms of  $c' \in [-2, 2]$ , but in terms of  $c = (c' + 2)/4$ ,  $c \in [0, 1]$ , and in the remainder of this chapter, we also use this reparameterization.

For each participant, we choose  $2N_1$  to match the number of samples obtained from Steingroever et al. (2016) which was at least 5,000; however, whenever this number of samples was insufficient to ensure convergence of

---

<sup>14</sup>Note that we excluded three participants due to incomplete choice data.

the Hamiltonian Monte Carlo (HMC) chains, Steingroever et al. (2016) repeated the fitting routine with 5,000 additional samples. Steingroever et al. (2016) confirmed convergence of the HMC chains by reporting that all  $\hat{R}$  statistics were below 1.05.

2. *We choose a proposal distribution.*

We generalize our approach from the running example and use a multivariate normal distribution as a proposal distribution.

3. *We transform the first batch of  $N_1$  posterior samples.*

Since we use a multivariate normal distribution as a proposal distribution, we have to transform all posterior samples to the real line using the probit transformation, that is,  $\omega_{s,j}^* = \Phi^{-1}(w_{s,j}^*)$ ,  $\alpha_{s,j}^* = \Phi^{-1}(a_{s,j}^*)$ ,  $\gamma_{s,j}^* = \Phi^{-1}(c_{s,j}^*)$ ,  $j = \{1, 2, \dots, N_1\}$ .

4. *We fit the proposal distribution to the first batch of  $N_1$  probit-transformed posterior samples.*

We use method of moment estimates for the mean vector and the covariance matrix obtained from the first batch of  $N_1$  probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. *We draw  $N_2$  samples from the proposal distribution.*

We use the R software to randomly draw  $N_2$  samples from the proposal distribution obtained in step 4. We obtain  $(\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$  with  $i \in \{1, 2, \dots, N_2\}$ .

6. *We calculate  $l_{2,i}$  for all  $N_2$  samples from the proposal distribution.*

This step involves assessing the value of the unnormalized posterior and the proposal distribution for all  $N_2$  samples from the proposal distribution. Before we can assess the value of the unnormalized posterior (i.e., the product of the likelihood and the prior), we have to derive how our transformation in step 3 affects the unnormalized posterior.

First, we derive how our transformation affects the likelihood. To evaluate the likelihood, we need to transform the probit-transformed samples back to the original parameter scales. That is, we evaluate the likelihood for  $(\Phi(\tilde{\omega}_{s,i}), \Phi(\tilde{\alpha}_{s,i}), \Phi(\tilde{\gamma}_{s,i}))$ . Before formalizing the likelihood of the observed choices of participant  $s$ , we define the following variables: We define  $Ch_s(t)$  as a vector containing the sequence of choices made by participant  $s$  up to and including trial  $t$ , and  $X_s(t)$  as a vector containing the corresponding sequence of net outcomes. We now obtain the following expression for the likelihood of the observed choices of participant  $s$ :

$$p(Ch_s(T) \mid \Phi(\tilde{\omega}_{s,i}), \Phi(\tilde{\alpha}_{s,i}), \Phi(\tilde{\gamma}_{s,i}), X_s(T-1)) = \prod_{t=0}^{T-1} \prod_{k=1}^4 Pr[S_k(t+1)] \cdot \delta_k(t+1). \quad (2.18)$$

Here  $T$  is the total number of trials,  $Pr[S_k(t+1)]$  is the probability of choosing deck  $k$  on trial  $t+1$ , and  $\delta_k(t+1)$  is a dummy variable which is 1 if deck  $k$  is chosen on trial  $t+1$  and 0 otherwise.

Second, we have to derive how our transformation affects the priors on each EV model parameter to yield priors on the probit-transformed model parameters. Since Steingroever et al. (2016) used independent uniform priors on  $[0, 1]$  we obtain standard normal priors on the probit-transformed model parameters (see beta-binomial example and Appendix D for an explanation).

7. *We transform the second batch of  $N_1$  posterior samples.*

This is analogous to step 2.

8. *We calculate  $l_{1,j}$  for the second batch of  $N_1$  probit-transformed samples from the posterior distribution.*

This is analogous to step 6.

9. *We run the iterative scheme (Equation 2.15) until our predefined tolerance criterion is reached.*

We use a tolerance criterion and initialization analogous to the running example. Once convergence is reached, we receive an estimate of the marginal likelihood for each participant, and derive the coefficient of variation for each participant using Equation 2.17. The largest coefficient of variation is 2.07% suggesting that the bridge sampler has low variance.<sup>15</sup>

### 2.3.4.2 Assessing the Accuracy of Our Implementation

To assess the accuracy of our implementation, we compared the marginal likelihood estimates obtained with our bridge sampler to the estimates obtained with importance sampling (Steingroever et al., 2016). Figure 2.6 shows the log marginal likelihoods for the 30 participants of Busemeyer and Stout (2002) obtained with bridge sampling (x-axis) and importance sampling reported by Steingroever et al. (2016; y-axis). The two sets of estimates correspond almost perfectly. These results indicate a successful implementation of the bridge sampler. Thus, this section emphasizes that both the importance sampler and bridge sampler can be used to estimate the marginal likelihood for the data of individual participants. However, when we want to estimate the marginal likelihood of a Bayesian hierarchical model, it may be difficult to find a suitable importance density. The bridge sampler, on the other hand, can be applied more easily and more efficiently.

### 2.3.5 Application of Bridge Sampling to a Hierarchical Implementation of the EV Model

In this section we illustrate how bridge sampling can be used to estimate the marginal likelihood of a hierarchical EV model. This hierarchical implementation assumes that the parameters  $w$ ,  $a$ , and  $c$  from each participant are drawn from three separate group-level distributions. This model specification hence incorporates both the differences and the similarities between participants. We illustrate this application using again the Busemeyer and Stout (2002) data set, and assume that these participants constitute one group.

---

<sup>15</sup>Note that this measure relates to the marginal likelihoods, not to the log marginal likelihoods.

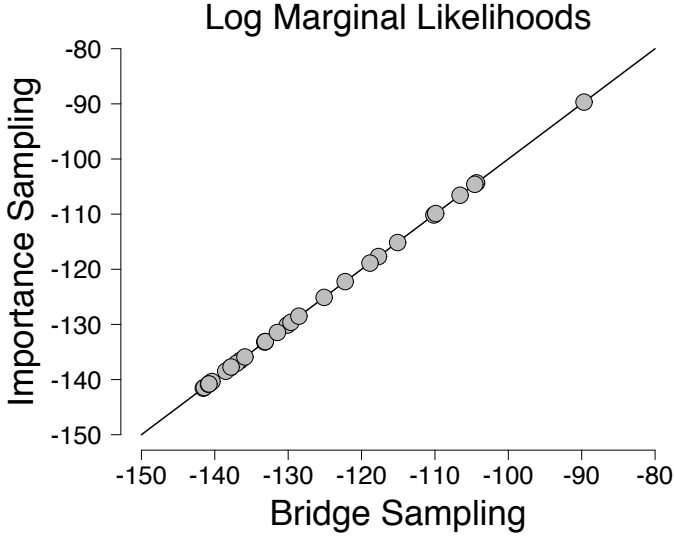


Figure 2.6: Comparison of the log marginal likelihoods obtained with bridge sampling (x-axis) and importance sampling reported by Steingroever et al. (2016; y-axis). The main diagonal indicates perfect correspondence between the two methods. Available at <https://tinyurl.com/yac3o8qs> under CC license <https://creativecommons.org/licenses/by/2.0/>.

### 2.3.5.1 Schematic Execution of the Bridge Sampler

To compute the marginal likelihood, we again follow the steps outlined in Figure 2.5, with a few minor modifications.

1. *For each parameter, that is, all individual-level and group-level parameters, we draw  $2N_1 = 60,000$  samples from the posterior distribution.*

To obtain the posterior samples, we fit a hierarchical Bayesian implementation of the EV model to the Busemeyer and Stout (2002) data set using the software JAGS (Plummer, 2003).<sup>16</sup> We assume that, for each participant  $s$ ,  $s \in \{1, 2, \dots, 30\}$ , each probit-transformed individual-level parameter (i.e.,  $\omega_s = \Phi^{-1}(w_s)$ ,  $\alpha_s = \Phi^{-1}(a_s)$ ,  $\gamma_s = \Phi^{-1}(c_s)$ ) is drawn from a group-level normal distribution characterized by a group-level mean and standard deviation parameter. For all group-level mean parameters  $\mu_\omega, \mu_\alpha, \mu_\gamma$  we assume a standard normal distribution, and for all group-level standard deviation parameters  $\sigma_\omega, \sigma_\alpha, \sigma_\gamma$  a uniform distribution ranging from 0 to 1.5. For a detailed explanation of the hierarchical implementation of the EV model, see Wetzels, Vandekerckhove, et al. (2010).

To reach convergence and reduce autocorrelation, we collect two MCMC chains, each with 120,000 samples from the posterior distributions after

---

<sup>16</sup>We used a model file that is an adapted version of the model file used by Ahn et al. (2011).

having excluded the first 30,000 samples as burn-in. Out of these 120,000 samples per chain, we retained every fourth value yielding 30,000 samples per chain. This setting resulted in all  $\hat{R}$  statistics below 1.05 suggesting that all chains have successfully converged from their starting values to their stationary distributions.

2. *We choose a proposal distribution.*

We use a multivariate normal distribution as a proposal distribution.

3. *We transform the first batch of  $N_1$  posterior samples.*

As before, we ensure that the range of the posterior distribution matches the range of the proposal distribution by using the probit transformation, that is,  $\omega_{s,j}^* = \Phi^{-1}(w_{s,j}^*)$ ,  $\alpha_{s,j}^* = \Phi^{-1}(a_{s,j}^*)$ ,  $\gamma_{s,j}^* = \Phi^{-1}(c_{s,j}^*)$ ,  $\tau_{\omega,j}^* = \Phi^{-1}((\sigma_{\omega,j}^*)/1.5)$ ,  $\tau_{\alpha,j}^* = \Phi^{-1}((\sigma_{\alpha,j}^*)/1.5)$ , and  $\tau_{\gamma,j}^* = \Phi^{-1}((\sigma_{\gamma,j}^*)/1.5)$ ,  $j = \{1, 2, \dots, N_1\}$ . The group-level mean parameters do not have to be transformed because they already range across the entire real line.

4. *We fit the proposal distribution to the first batch of the  $N_1$  probit-transformed posterior samples.*

We use method of moment estimates for the mean vector and the covariance matrix obtained from the first batch of  $N_1$  probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. *We draw  $N_2$  samples from the proposal distribution.*

We use the R software to randomly draw  $N_2$  samples from the proposal distribution obtained in step 4. We obtain  $(\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$  and  $(\tilde{\mu}_{\omega,i}, \tilde{\tau}_{\omega,i}, \tilde{\mu}_{\alpha,i}, \tilde{\tau}_{\alpha,i}, \tilde{\mu}_{\gamma,i}, \tilde{\tau}_{\gamma,i})$  with  $i \in \{1, 2, \dots, N_2\}$  and  $s \in \{1, 2, \dots, 30\}$ .

6. *We calculate  $l_{2,i}$  for all  $N_2$  samples from the proposal distribution.*

This step involves assessing the value of the unnormalized posterior and the proposal distribution for all  $N_2$  samples from the proposal distribution. The unnormalized posterior is defined as:

$$\left( \prod_{s=1}^{30} p(Ch_s(T) \mid \Phi(\tilde{\mathbf{\kappa}}_{s,i}), X_s(T-1)) p(\tilde{\mathbf{\kappa}}_{s,i} \mid \tilde{\boldsymbol{\zeta}}_i) \right) p(\tilde{\boldsymbol{\zeta}}_i), \quad \text{where } Ch_s(T)$$

refers to all choices of subject  $s$ ,  $X_s(T-1)$  to the net outcomes that subject  $s$  experienced on trials  $1, 2, \dots, T-1$ ,  $\tilde{\mathbf{\kappa}}_{s,i} = (\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$  to the  $i^{th}$  sample from the proposal distribution for the individual-level parameters of subject  $s$ , and  $\tilde{\boldsymbol{\zeta}}_i$  to the  $i^{th}$  sample from the proposal distribution for all group-level parameters (e.g.,  $\tilde{\boldsymbol{\zeta}}_i = (\tilde{\mu}_{\omega,i}, \tilde{\tau}_{\omega,i}, \tilde{\mu}_{\alpha,i}, \tilde{\tau}_{\alpha,i}, \tilde{\mu}_{\gamma,i}, \tilde{\tau}_{\gamma,i})$ ).

The likelihood function for a given participant is the same as in the individual case. However, for each participant we now have to add besides the prior on the individual-level parameters also the prior on the group-level parameters. The product of the likelihood and the priors gives the unnormalized posterior density (see Appendix E for more details).

7. We follow steps 7 – 9, as outlined for the bridge sampler of the individual-level implementation of the EV model.

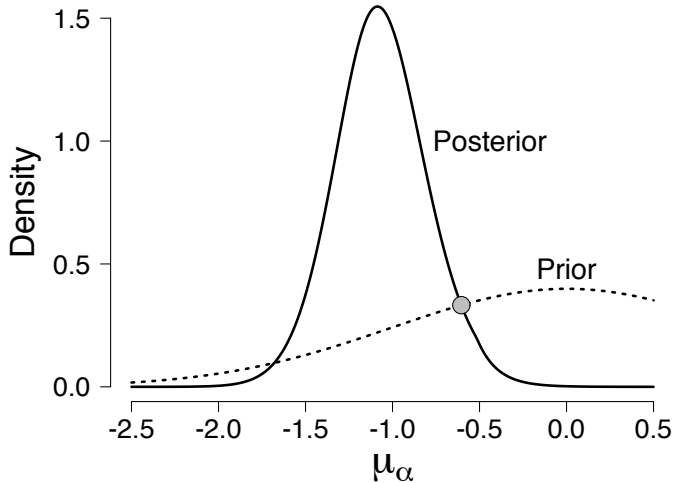


Figure 2.7: Prior and posterior distribution of the group-level mean  $\mu_\alpha$  in the Bussemeyer and Stout (2002) data set. The figure shows the posterior distribution (solid line) and the prior distribution (dotted line). The gray dot indicates the intersection of the prior and the posterior distributions, for which the Savage-Dickey Bayes factor equals 1. Available at <https://tinyurl.com/y7cyxclq> under CC license <https://creativecommons.org/licenses/by/2.0/>.

### 2.3.5.2 Assessing the Accuracy of Our Implementation

To investigate the accuracy of our implementation, we compare Bayes factors obtained with bridge sampling to Bayes factors obtained from the Savage-Dickey density ratio test (Dickey, 1971; Dickey & Lientz, 1970; for a tutorial, see Wagenmakers et al., 2010). The Savage-Dickey density ratio is a simple method for computing Bayes factors for nested models. We artificially create three nested models by taking the full EV model  $\mathcal{M}_f$  in which all parameters are free to vary, and then restricting one of the three group-level mean parameters, that is,  $\mu_\omega$ ,  $\mu_\alpha$ , or  $\mu_\gamma$ , to a predefined value. For these values we choose the intersection point of the prior and posterior distribution of each group-level mean parameter. To obtain these intersection points, we fit the full EV model and then use a nonparametric logspline density estimator (C. J. Stone, Hansen, Kooperberg, & Truong, 1997). The obtained values are presented in Table 2.2. Since we compare the full model to each restricted model, we obtain three Bayes factors.

According to the Savage-Dickey density ratio test, the Bayes factor for the full model versus a specific restricted model  $\mathcal{M}_r$  can be obtained by dividing the height of the prior density at the predefined parameter value  $\theta_0$  by the height of

Table 2.2: Bayes factors comparing the full EV model to the restricted EV models, log marginal likelihoods, and coefficient of variation (with respect to the marginal likelihood) expressed as a percentage.

Model	Bayes Factor	Log Marginal Likelihood	CV[%]
full model	–	–3800.434	10.13
restricted at $\mu_\omega = -0.334$	1.202	–3800.618	16.44
restricted at $\mu_\alpha = -0.604$	1.052	–3800.484	9.71
restricted at $\mu_\gamma = 0.92$	1.068	–3800.500	12.03

the posterior at the same location:

$$\text{BF}_{\mathcal{M}_f, \mathcal{M}_r} = \frac{p(y \mid \mathcal{M}_f)}{p(y \mid \mathcal{M}_r)} = \frac{p(\theta = \theta_0 \mid \mathcal{M}_f)}{p(\theta = \theta_0 \mid y, \mathcal{M}_f)}. \quad (2.19)$$

Since we choose  $\theta_0$  to be the intersection point of the prior and posterior distribution,  $\text{BF}_{\mathcal{M}_f, \mathcal{M}_r}$  equals 1. This Savage-Dickey Bayes factor of 1 indicates that the marginal likelihood under the full model equals the marginal likelihood under the restricted model. Figure 2.7 illustrates the Savage-Dickey Bayes factor comparing the full model to the model assuming  $\mu_\alpha$  fixed to  $-0.604$ .

The computation of the three bridge sampling Bayes factors, on the other hand, works as follows: First, we follow the steps outlined above to obtain the bridge sampling estimate of the full EV model. Second, we obtain the bridge sampling estimate of the marginal likelihood for the three restricted models. This requires adapting the steps outlined above to each of the three restricted models. Lastly, we use the first equality in Equation 2.19 to obtain the three Bayes factors.

The Bayes factors derived from bridge sampling are reported in Table 2.2. It is evident that Bayes factors derived from bridge sampling closely approximate the Savage-Dickey Bayes factors of 1. These results suggest a successful implementation of the bridge sampler. This is also reflected by the close match between the log marginal likelihoods of the four models presented in the third column of Table 2.2.

Finally, we confirm that the bridge sampler has low variance; the coefficient of variation with respect to the marginal likelihood of the full model and the three restricted models ranges between 9.71 and 16.44%.

## 2.4 Discussion

In this tutorial, we explained how bridge sampling can be used to estimate the marginal likelihood of popular models in mathematical psychology. As a running example, we used the beta-binomial model to illustrate step-by-step the bridge sampling estimator. To facilitate the understanding of the bridge sampler, we first discussed three of its special cases – the naive Monte Carlo estimator, the importance sampling estimator, and the generalized harmonic mean estimator.

Consequently, we introduced key concepts that became gradually more complicated and sophisticated. In the second part of this tutorial, we showed how bridge sampling can be used to estimate the marginal likelihood of both an individual-level and a hierarchical implementation of the Expectancy Valence (EV; Busemeyer & Stout, 2002) model – a popular reinforcement-learning model for the Iowa gambling task (IGT; Bechara et al., 1994). The running example and the application of bridge sampling to the EV model demonstrated the positive aspects of the bridge sampling estimator, that is, its accuracy, reliability, practicality, and ease-of-implementation (DiCiccio et al., 1997; Frühwirth-Schnatter, 2004; Meng & Wong, 1996).

The bridge sampling estimator is superior to the naive Monte Carlo estimator, the importance sampling estimator, and the generalized harmonic mean estimator for several reasons. First, Meng and Wong (1996) showed that, among the four estimators discussed in this chapter, the bridge sampler presented in this chapter minimizes the mean-squared error because it uses the optimal bridge function. Second, in bridge sampling, choosing a suitable proposal distribution is much easier than choosing a suitable importance density for the importance sampling estimator or the generalized harmonic mean estimator because bridge sampling is more robust to the tail behavior of the proposal distribution relative to the posterior distribution. This advantage facilitates the application of the bridge sampler to higher-dimensional and complex models. This characteristic of the bridge sampler combined with the popularity of higher-dimensional and complex models in mathematical psychology suggests that bridge sampling can advance model comparison exercises in many areas of mathematical psychology (e.g., reinforcement-learning models, response time models, multinomial processing tree models, etc.). Third, bridge sampling is relatively straightforward to implement. In particular, our step-by-step procedure can be easily applied to other models with only minor changes of the code (i.e., the unnormalized posterior and potentially the proposal function have to be adapted). In our opinion, this is one of the most attractive features of bridge sampling: It is an accurate yet very generic method. Exploiting this generic characteristic, we have implemented the bridge sampling procedure in the **bridgesampling** R package (Gronau, Singmann, & Wagenmakers, 2020) in order to maximize its accessibility.

Despite the numerous advantages of the bridge sampler, the take-home message of this tutorial is not that the bridge sampler should be used blindly. There exist a large variety of methods to approximate the marginal likelihood that differ in their efficiency.<sup>17</sup> The most appropriate method optimizes the trade-off between accuracy and implementation effort. This trade-off depends on a number of aspects such as the complexity of the model, the number of models under consideration, the statistical experience of the researcher, and the time available. This suggests that the choice of the method should be reconsidered each time a marginal likelihood needs to be obtained. Obviously, when the marginal likelihood can be determined analytically, bridge sampling is not needed at all. If the goal is to compare (at least)

---

<sup>17</sup>In general, a large number of approaches for model selection exist which are based on MCMC posterior sampling and some of them are not based on approximating the models' marginal likelihoods (e.g., Ando, 2007; Spiegelhalter, Best, Carlin, & van der Linde, 2002). A comparison of these methods is beyond the scope of this tutorial.

two nested models, the Savage-Dickey density ratio test (Dickey, 1971; Dickey & Lientz, 1970) might be a better alternative. Note, however, that this requires an approximation of the marginal posterior density of one or more parameters which can be unstable in case the test value falls in the tail of the distribution. If only an individual-level implementation of a model is used, importance sampling may be easier to implement and may require less computational effort. This presupposes that one can find a proposal distribution with fatter tails than the posterior which may not always be trivial (even in an individual-level case). If the goal is to obtain the marginal likelihood of a large number of relatively simple models, the product space or reversible jump method (RJMCMC) might be more appropriate (Carlin & Chib, 1995; Green, 1995; Lodewyckx et al., 2011). In contrast to bridge sampling, implementations of these methods tend to be problem-specific rather than generic (but see Lunn, Best, & Whittaker, 2009). If a researcher with a limited programming background and/or little time resources wants to conduct a model comparison exercise, rough approximations of the Bayes factor, such as the Bayesian information criterion, might be more suitable (Schwarz, 1978). It should be kept in mind, however, that this approximation assumes a certain prior structure that may not respect the knowledge or information that researchers have at their disposal. On the other hand, a researcher with an extensive background in programming and mathematical statistics might consider using path sampling – a generalization of bridge sampling (Gelman & Meng, 1998).

To conclude, in this tutorial we showed that bridge sampling offers a reliable and easy-to-implement approach to estimating a model’s marginal likelihood. Bridge sampling can be profitably applied to a wide range of problems in mathematical psychology involving parameter estimation, model comparison, and Bayesian model averaging.

R scripts for reproducing the analyses presented in this chapter are available at <https://osf.io/f9cq4/>.

## 2.A The Bridge Sampling Estimator as a General Case of Methods 1 – 3

In this section we show that the naive Monte Carlo, the importance sampling, and the generalized harmonic mean estimators are special cases of the bridge sampling estimator under specific choices of the bridge function  $h(\theta)$  and the proposal distribution  $g(\theta)$ .<sup>18</sup> An overview is provided in Table 2.3.

To prove that the bridge sampling estimator reduces to the naive Monte Carlo estimator, consider bridge sampling, choose the prior distribution as the proposal distribution (i.e.,  $g(\theta) = p(\theta)$ ), and specify the bridge function as  $h(\theta) = 1/g(\theta)$ . Inserting these specifications into Equation 2.12 yields:

$$\begin{aligned} \hat{p}_4(y | h(\theta) = \frac{1}{g(\theta)}, g(\theta) = p(\theta)) \\ &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{1}{p(\tilde{\theta}_i)} p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{p(\theta_j^*)} p(\theta_j^*)}, \quad \tilde{\theta}_i \sim p(\theta), \quad \theta_j^* \sim p(\theta | y) \\ &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(y | \tilde{\theta}_i)}{\frac{1}{N_1} N_1} = \frac{1}{N_2} \sum_{i=1}^{N_2} p(y | \tilde{\theta}_i), \quad \tilde{\theta}_i \sim p(\theta), \end{aligned}$$

which is equivalent to the naive Monte Carlo estimator shown in Equation 2.7.

To prove that the bridge sampling estimator reduces to the importance sampling estimator, consider bridge sampling, choose the importance density as the proposal distribution (i.e.,  $g(\theta) = g_{IS}(\theta)$ ), and specify the bridge function as  $h(\theta) = 1/g(\theta)$ . Inserting these specifications into Equation 2.12 yields:

$$\begin{aligned} \hat{p}_4(y | h(\theta) = \frac{1}{g(\theta)}, g(\theta) = g_{IS}(\theta)) \\ &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{1}{g_{IS}(\tilde{\theta}_i)} p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{g_{IS}(\theta_j^*)} g_{IS}(\theta_j^*)}, \quad \tilde{\theta}_i \sim g_{IS}(\theta), \quad \theta_j^* \sim p(\theta | y) \\ &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}}{\frac{1}{N_1} N_1} = \frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}, \quad \tilde{\theta}_i \sim g_{IS}(\theta), \end{aligned}$$

which is equivalent to the importance sampling estimator shown in Equation 2.8.

<sup>18</sup>Note that bridge sampling is also a general case of the Chib and Jeliazkov (2001) method of estimating the marginal likelihood using the Metropolis-Hastings acceptance probability (Meng & Schilling, 2002; Mira & Nicholls, 2004).

Table 2.3: Summary of the bridge sampling estimator for the marginal likelihood, and its special cases: the naive Monte Carlo, importance sampling, and generalized harmonic mean estimator.

Method	Estimator	Samples	Bridge Function $h(\theta)$
Bridge sampling	$\frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(y   \tilde{\theta}_i) p(\tilde{\theta}_i) h(\tilde{\theta}_i)}{\frac{1}{N_1} \sum_{j=1}^{N_1} h(\theta_j^*) g(\theta_j^*)}$	$\tilde{\theta}_i \sim g(\theta)$ $\theta_j^* \sim p(\theta   y)$	$h(\theta) = \frac{C}{\frac{N_1}{N_2 + N_1} p(y   \theta) p(\theta) + \frac{N_2}{N_2 + N_1} p(y) g(\theta)}$
Naive Monte Carlo	$\frac{1}{N} \sum_{i=1}^N p(y   \tilde{\theta}_i)$	$\tilde{\theta}_i \sim p(\theta)$	$h(\theta) = \frac{1}{g(\theta)}$ and $g(\theta) = p(\theta)$
Importance sampling	$\frac{1}{N} \sum_{i=1}^N \frac{p(y   \tilde{\theta}_i) p(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}$	$\tilde{\theta}_i \sim g_{IS}(\theta)$	$h(\theta) = \frac{1}{g(\theta)}$ and $g(\theta) = g_{IS}(\theta)$
Generalized harmonic mean	$\left( \frac{1}{N} \sum_{i=1}^N \frac{g_{IS}(\theta_i^*)}{p(y   \theta_i^*) p(\theta_i^*)} \right)^{-1}$	$\theta_i^* \sim p(\theta   y)$	$h(\theta) = \frac{1}{p(y   \theta) p(\theta)}$ and $g(\theta) = g_{IS}(\theta)$

*Note.*  $p(\theta)$  is the prior distribution,  $g_{IS}(\theta)$  is the importance density,  $p(\theta | y)$  is the posterior distribution,  $g(\theta)$  is the proposal distribution,  $h(\theta)$  is the bridge function, and  $C$  is a constant. The last column shows the bridge function needed to obtain the special cases.

To prove that the bridge sampling estimator reduces to the generalized harmonic mean estimator, consider bridge sampling, choose the importance density as the proposal distribution (i.e.,  $g(\theta) = g_{IS}(\theta)$ ), and specify the bridge function as  $h(\theta) = 1/(p(y | \theta) p(\theta))$ . Inserting these specifications into Equation 2.12 yields:

$$\begin{aligned}
 \hat{p}_4(y | h(\theta) = \frac{1}{p(y | \theta) p(\theta)}, g(\theta) = g_{IS}(\theta)) \\
 &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{1}{p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)} p(y | \tilde{\theta}_i) p(\tilde{\theta}_i)}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{p(y | \theta_j^*) p(\theta_j^*)} g_{IS}(\theta_j^*)}, \quad \tilde{\theta}_i \sim g_{IS}(\theta), \quad \theta_j^* \sim p(\theta | y) \\
 &= \frac{\frac{1}{N_2} N_2}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g_{IS}(\theta_j^*)}{p(y | \theta_j^*) p(\theta_j^*)}} = \left( \frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g_{IS}(\theta_j^*)}{p(y | \theta_j^*) p(\theta_j^*)} \right)^{-1}, \quad \theta_j^* \sim p(\theta | y),
 \end{aligned}$$

which is equivalent to the generalized harmonic mean estimator shown in Equation 2.9.

## 2.B Bridge Sampling Implementation: Avoiding Numerical Issues

In order to avoid numerical issues, we can rewrite Equation 2.15 in the following way:

$$\begin{aligned}
 \hat{p}_4(y)^{(t+1)} &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}_4(y)^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}_4(y)^{(t)}}} \\
 &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}))}{s_1 \exp(\log(l_{2,i})) + s_2 \hat{p}_4(y)^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 \exp(\log(l_{1,j})) + s_2 \hat{p}_4(y)^{(t)}}} \\
 &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i})) \exp(-l^*)}{s_1 \exp(\log(l_{2,i})) \exp(-l^*) + s_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{\exp(-l^*)}{s_1 \exp(\log(l_{1,j})) \exp(-l^*) + s_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\exp(-l^*)} \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - l^*)}{s_1 \exp(\log(l_{2,i}) - l^*) + s_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 \exp(\log(l_{1,j}) - l^*) + s_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}} \\
 &= \exp(l^*) \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - l^*)}{s_1 \exp(\log(l_{2,i}) - l^*) + s_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 \exp(\log(l_{1,j}) - l^*) + s_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}.
 \end{aligned}$$

$l^*$  is a constant which we can choose in a way that keeps the terms in the sums manageable. We used  $l^* = \text{median}(\log(l_{1,j}))$ . Let

$$\hat{r}^{(t)} = \hat{p}_4(y)^{(t)} \exp(-l^*),$$

so that

$$\hat{p}_4(y)^{(t)} = \hat{r}^{(t)} \exp(l^*).$$

Then we obtain

$$\begin{aligned}
 \hat{p}_4(y)^{(t+1)} &= \exp(l^*) \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - l^*)}{s_1 \exp(\log(l_{2,i}) - l^*) + s_2 \hat{r}^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 \exp(\log(l_{1,j}) - l^*) + s_2 \hat{r}^{(t)}}} \\
 \hat{p}_4(y)^{(t+1)} \exp(-l^*) &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - l^*)}{s_1 \exp(\log(l_{2,i}) - l^*) + s_2 \hat{r}^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 \exp(\log(l_{1,j}) - l^*) + s_2 \hat{r}^{(t)}}} \\
 \hat{r}^{(t+1)} &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - l^*)}{s_1 \exp(\log(l_{2,i}) - l^*) + s_2 \hat{r}^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 \exp(\log(l_{1,j}) - l^*) + s_2 \hat{r}^{(t)}}}.
 \end{aligned}$$

Hence, we can run the iterative scheme with respect to  $\hat{r}$  which is more convenient because it keeps the terms in the sums manageable and multiply the result by  $\exp(l^*)$  to obtain the estimate of the marginal likelihood or, equivalently, we can take the logarithm of the result and add  $l^*$  to obtain an estimate of the logarithm of the marginal likelihood.

## 2.C Correcting for the Probit Transformation

In this section we describe how the probit transformation affects our expression of the generalized harmonic mean estimator (Equation 2.9) to yield Equation 2.10.

Recall that we derived the generalized harmonic mean estimator using the following equality:

$$\frac{1}{p(y)} = \int \frac{g_{IS}(\theta)}{p(y | \theta)p(\theta)} p(\theta | y) d\theta. \quad (2.20)$$

For practical reasons, in the running example, we used a normal distribution on  $\xi$  as importance density. This  $\xi$  was defined as the probit transform of  $\theta$  (i.e.,  $\xi = \Phi^{-1}(\theta)$ ). In particular, the normal importance density was given by  $\frac{1}{\hat{\sigma}} \phi\left(\frac{\xi - \hat{\mu}}{\hat{\sigma}}\right)$ . Note that this importance density is a function of  $\xi$ , whereas the general importance density  $g_{IS}$  in Equation 2.20 is specified in terms of  $\theta$ . Therefore, to include our specific importance density into Equation 2.20, we need to write it in terms of  $\theta$ . This yields  $\frac{1}{\hat{\sigma}} \phi\left(\frac{\Phi^{-1}(\theta) - \hat{\mu}}{\hat{\sigma}}\right) \frac{1}{\phi(\Phi^{-1}(\theta))}$ , where the latter factor comes from applying the change-of-variable method. Replacing  $g_{IS}(\theta)$  in Equation 2.20 by this expression, results in:

$$\begin{aligned} \frac{1}{p(y)} &= \int \frac{\frac{1}{\hat{\sigma}} \phi\left(\frac{\Phi^{-1}(\theta) - \hat{\mu}}{\hat{\sigma}}\right) \frac{1}{\phi(\Phi^{-1}(\theta))}}{p(y | \theta)p(\theta)} p(\theta | y) d\theta \\ &= \mathbb{E}_{\text{post}} \left( \frac{\frac{1}{\hat{\sigma}} \phi\left(\frac{\Phi^{-1}(\theta) - \hat{\mu}}{\hat{\sigma}}\right) \frac{1}{\phi(\Phi^{-1}(\theta))}}{p(y | \theta) p(\theta)} \right). \end{aligned} \quad (2.21)$$

Rewriting results in:

$$p(y) = \left( \mathbb{E}_{\text{post}} \left( \frac{\frac{1}{\hat{\sigma}} \phi\left(\frac{\Phi^{-1}(\theta) - \hat{\mu}}{\hat{\sigma}}\right) \frac{1}{\phi(\Phi^{-1}(\theta))}}{p(y | \theta) p(\theta)} \right) \right)^{-1},$$

which can be approximated as:

$$\begin{aligned} \hat{p}_3(y) &= \left( \frac{1}{N} \sum_{j=1}^N \frac{\overbrace{\frac{1}{\hat{\sigma}} \phi\left(\frac{\Phi^{-1}(\theta_j^*) - \hat{\mu}}{\hat{\sigma}}\right) \frac{1}{\phi(\Phi^{-1}(\theta_j^*))}}^{\text{importance density}}}{\underbrace{p(y | \theta_j^*)}_{\text{likelihood}} \underbrace{p(\theta_j^*)}_{\text{prior}}} \right)^{-1}, \quad \underbrace{\theta_j^* \sim p(\theta | y)}_{\text{samples from the posterior distribution}} \\ &= \left( \frac{1}{N} \sum_{j=1}^N \frac{\overbrace{\frac{1}{\hat{\sigma}} \phi\left(\frac{\xi_j^* - \hat{\mu}}{\hat{\sigma}}\right)}^{\text{importance density}}}{\underbrace{p(y | \Phi(\xi_j^*))}_{\text{likelihood}} \underbrace{p(\Phi(\xi_j^*)) \phi(\xi_j^*)}_{\text{prior}}} \right)^{-1}, \quad \underbrace{\xi_j^* = \Phi^{-1}(\theta_j^*) \text{ and } \theta_j^* \sim p(\theta | y)}_{\text{probit-transformed samples from the posterior distribution}} \end{aligned} \quad (2.22)$$

which shows that the generalized harmonic estimate can be obtained using the samples from the posterior distribution, or the probit-transformed ones. In the online-provided code, we use the latter approach (see also Overstall & Forster, 2010). Note that in our running example,  $\forall \xi_j^* : p(\Phi(\xi_j^*)) = 1$ .

## 2.D Details on the Application of Bridge Sampling to the Individual-Level EV Model

In this section, we provide more details on how we obtained the unnormalized posterior distribution for a specific participant  $s$ ,  $s \in \{1, 2, \dots, 30\}$ . Since we focus on one specific participant, we drop the subscript  $s$  in the remainder of this section. As explained in Appendix B, we run the iterative scheme with respect to  $\hat{r}$  to avoid numerical issues. Consequently, we have to compute  $\log(l_{1,j})$  and  $\log(l_{2,i})$ . Using  $\tilde{\boldsymbol{\kappa}}_i = (\tilde{\omega}_i, \tilde{\alpha}_i, \tilde{\gamma}_i)$  for the  $i^{th}$  sample from the proposal distribution, we get for  $\log(l_{2,i})$  ( $\log(l_{1,j})$  works analogously):

$$\log(l_{2,i}) = \log \left( \frac{p(Ch(T) \mid \Phi(\tilde{\boldsymbol{\kappa}}_i), X(T-1)) p(\Phi(\tilde{\boldsymbol{\kappa}}_i)) \phi(\tilde{\boldsymbol{\kappa}}_i)}{g(\tilde{\boldsymbol{\kappa}}_i)} \right).$$

Therefore, instead of computing the unnormalized posterior distribution directly, we compute the logarithm of the unnormalized posterior distribution:

$$\begin{aligned} \log(p(Ch(T) \mid \Phi(\tilde{\boldsymbol{\kappa}}_i), X(T-1)) p(\Phi(\tilde{\boldsymbol{\kappa}}_i)) \phi(\tilde{\boldsymbol{\kappa}}_i)) &= \log(p(Ch(T) \mid \Phi(\tilde{\boldsymbol{\kappa}}_i), X(T-1))) + \\ &\quad \log(\phi(\tilde{\omega}_i)) + \log(\phi(\tilde{\alpha}_i)) + \log(\phi(\tilde{\gamma}_i)), \end{aligned}$$

because we assumed independent priors on each model parameter  $w$ ,  $a$ ,  $c$ .  $\log(p(\Phi(\tilde{\boldsymbol{\kappa}}_i))) = 0$  because  $p$  refers to the uniform prior on  $[0, 1]$ .

## 2.E Details on the Application of Bridge Sampling to the Hierarchical EV Model

Analogous to the last section, we explain here how we obtained the logarithm of the unnormalized posterior for the hierarchical implementation of the EV model. Using  $\tilde{\boldsymbol{\kappa}}_{s,i} = (\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$  for the  $i^{th}$  sample from the proposal distribution for the individual-level parameters of subject  $s$ , and  $\tilde{\boldsymbol{\zeta}}_i$  for the  $i^{th}$  sample from the proposal distribution for all group-level parameters (i.e.,  $\tilde{\boldsymbol{\zeta}}_i = (\tilde{\mu}_{\omega,i}, \tilde{\tau}_{\omega,i}, \tilde{\mu}_{\alpha,i}, \tilde{\tau}_{\alpha,i}, \tilde{\mu}_{\gamma,i}, \tilde{\tau}_{\gamma,i})$ ), we get:

$$\begin{aligned} \log \left( \left( \prod_{s=1}^{30} p(Ch_s(T) \mid \Phi(\tilde{\boldsymbol{\kappa}}_{s,i}), X_s(T-1)) p(\tilde{\boldsymbol{\kappa}}_{s,i} \mid \tilde{\boldsymbol{\zeta}}_i) \right) p(\tilde{\boldsymbol{\zeta}}_i) \right) \\ = \sum_{s=1}^N [\log(p(Ch_s(T) \mid \Phi(\tilde{\boldsymbol{\kappa}}_{s,i}), X_s(T-1)))] + \end{aligned}$$

$$\begin{aligned}
& \log \left( \frac{1}{1.5\Phi(\tilde{\tau}_{\omega,i})} \phi \left( \frac{\tilde{\omega}_{s,i} - \tilde{\mu}_{\omega,i}}{1.5\Phi(\tilde{\tau}_{\omega,i})} \right) \right) + \log \left( \frac{1}{1.5\Phi(\tilde{\tau}_{\alpha,i})} \phi \left( \frac{\tilde{\alpha}_{s,i} - \tilde{\mu}_{\alpha,i}}{1.5\Phi(\tilde{\tau}_{\alpha,i})} \right) \right) + \\
& \log \left( \frac{1}{1.5\Phi(\tilde{\tau}_{\gamma,i})} \phi \left( \frac{\tilde{\gamma}_{s,i} - \tilde{\mu}_{\gamma,i}}{1.5\Phi(\tilde{\tau}_{\gamma,i})} \right) \right) \Bigg] + \\
& \log(\phi(\tilde{\mu}_{\omega,i})) + \log(\phi(\tilde{\mu}_{\alpha,i})) + \log(\phi(\tilde{\mu}_{\gamma,i})) + \\
& \log(\phi(\tilde{\tau}_{\omega,i})) + \log(\phi(\tilde{\tau}_{\alpha,i})) + \log(\phi(\tilde{\tau}_{\gamma,i})) .
\end{aligned}$$



# A Simple Method for Comparing Complex Models: Bayesian Model Comparison for Hierarchical Multinomial Processing Tree Models using Warp-III Bridge Sampling

---

## Abstract

Multinomial processing trees (MPTs) are a popular class of cognitive models for categorical data. Typically, researchers compare several MPTs, each equipped with many parameters, especially when the models are implemented in a hierarchical framework. A Bayesian solution is to compute posterior model probabilities and Bayes factors. Both quantities, however, rely on the marginal likelihood, a high-dimensional integral that cannot be evaluated analytically. In this chapter, we show how Warp-III bridge sampling can be used to compute the marginal likelihood for hierarchical MPTs. We illustrate the procedure with two published data sets and demonstrate how Warp-III facilitates Bayesian model averaging.

## 3.1 Introduction

Multinomial processing trees (MPTs; e.g., Riefer & Batchelder, 1988) are substantively motivated stochastic models for the analysis of categorical data. MPTs

---

This chapter is published as Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2019). A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using Warp-III bridge sampling. *Psychometrika*, 84, 261–284. doi: <https://doi.org/10.1007/s11336-018-9648-3>. Also available as *PsyArXiv preprint*: <https://psyarxiv.com/yxhfm/>

allow researchers to test theories about cognitive architecture by formalizing qualitatively different cognitive processes that underlie performance in an experimental paradigm. MPTs are popular in various areas of psychology and have been applied, for instance, in research on memory, perception, logical reasoning, and attitudes (for reviews, see Batchelder & Riefer, 1999; Erdfelder et al., 2009; Hütter & Klauer, 2016). MPTs are related to tree-based item response theory models as presented, for instance, in Böckenholt (2012a), Böckenholt (2012b), Culpepper (2014), and De Boeck and Partchev (2012).<sup>1</sup>

Traditionally, parameter estimation in MPTs has relied on maximum-likelihood methods for aggregated data (Hu & Batchelder, 1994; Singmann & Kellen, 2013). Recently, however, MPT modelers have become increasingly interested in using Bayesian hierarchical methods to examine individual differences in model parameters (Klauer, 2010; Matzke et al., 2015; J. B. Smith & Batchelder, 2010). Bayesian hierarchical modeling allows researchers to simultaneously account for the differences and similarities between participants and typically provides more accurate statistical inference than the analysis of aggregated data, especially in situations with moderate between-subject variability and scarce participant-level data (e.g., Gelman & Hill, 2007).

In typical applications, MPT modelers are interested in comparing a limited set of models. The models can be nested, which is the case when testing parameter constraints (e.g., Batchelder & Riefer, 1990; Singmann, Kellen, & Klauer, 2013), or non-nested, which is the case when comparing structurally different models (e.g., Fazio, Brashier, Payne, & Marsh, 2015; Kellen, Singmann, & Klauer, 2014). A wide range of model comparison and assessment methods exist both in the frequentist and Bayesian framework, each with its own goals and operating characteristics, such as Pearson’s  $\chi^2$  test, the likelihood ratio test, information criteria such as AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC (Spiegelhalter et al., 2002), and WAIC (Watanabe, 2010), leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2017), and posterior predictive checks (Gelman, 2013; Meng, 1994; Robins, van der Vaart, & Ventura, 2000). Furthermore, a range of powerful methods exist for analyzing multinomial data in particular (e.g., Bishop, Fienberg, & Holland, 1975; Maydeu-Olivares & Joe, 2005). The goal of this chapter is to enrich the model comparison toolkit of MPT modelers by illustrating – with examples from the literature – a computationally feasible approach to model comparison in hierarchical MPTs based on Bayes factors and posterior model probabilities.<sup>2</sup> Furthermore, the proposed approach also enables Bayesian model averaging which we advocate as a principled way of testing parameter constraints while fully taking into account model uncertainty.

Suppose one is interested in comparing a discrete set of  $M$  models denoted as  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$  with corresponding prior model probabilities  $p(\mathcal{M}_1), p(\mathcal{M}_2), \dots, p(\mathcal{M}_M)$ , which satisfy the constraints  $p(\mathcal{M}_i) \geq 0 \quad \forall i \in \{1, 2, \dots, M\}$  and  $\sum_{i=1}^M p(\mathcal{M}_i) = 1$ . The posterior model probability of  $\mathcal{M}_i$  is

---

<sup>1</sup>The interested reader is referred to Plieninger and Heck (2018) for a comparison of these model classes.

<sup>2</sup>Note that posterior model probabilities can also be obtained using information criteria (e.g., Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004).

then obtained using Bayes' rule:

$$\underbrace{p(\mathcal{M}_i \mid \text{data})}_{\text{posterior model probability}} = \underbrace{\frac{p(\text{data} \mid \mathcal{M}_i)}{\sum_{j=1}^M p(\text{data} \mid \mathcal{M}_j) p(\mathcal{M}_j)}}_{\text{updating factor}} \times \underbrace{p(\mathcal{M}_i)}_{\text{prior model probability}}, \quad (3.1)$$

where  $p(\text{data} \mid \mathcal{M}_i)$  is the *marginal likelihood* of model  $\mathcal{M}_i$ .

If model comparison involves assessing the tenability of parameter constraints in a set of nested models, posterior model probabilities can be used to quantify the model-averaged evidence that a parameter is free to vary or should be constrained across different groups or experimental conditions (e.g., Hoeting et al., 1999; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017). If the model comparison involves only two models,  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , it is convenient to consider the odds of one model over the other one. Bayes' rule yields:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \text{data})}{p(\mathcal{M}_2 \mid \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\text{data} \mid \mathcal{M}_1)}{p(\text{data} \mid \mathcal{M}_2)}}_{\text{Bayes factor BF}_{12}} \times \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{prior odds}}. \quad (3.2)$$

Equation 3.2 shows that the change in odds brought about by the data is given by the ratio of the marginal likelihoods of the models, a quantity known as the *Bayes factor* (Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Ly et al., 2016a).

Equation 3.1 and Equation 3.2 illustrate that the computation of posterior model probabilities and Bayes factors requires the computation of the marginal likelihood of the models. The marginal likelihood is obtained by integrating out the model parameters with respect to the parameters' prior distribution:

$$p(\text{data} \mid \mathcal{M}_i) = \int_{\Theta} p(\text{data} \mid \theta, \mathcal{M}_i) p(\theta \mid \mathcal{M}_i) d\theta. \quad (3.3)$$

The marginal likelihood includes a natural penalty for overdue model complexity and implements a form of the principle of parsimony also known as *Occam's razor* (e.g., Jefferys & Berger, 1992; Myung & Pitt, 1997; Vandekerckhove et al., 2015).<sup>3</sup> Although conceptually straightforward, in practice it is challenging to compute Bayes factors and posterior model probabilities for hierarchical MPTs because the marginal likelihood features a high-dimensional integral that cannot be solved analytically.

In this chapter, we show how Warp-III bridge sampling (Meng & Schilling, 2002; Meng & Wong, 1996, henceforth referred to as Warp-III) can be used to estimate the marginal likelihood for hierarchical MPTs. Warp-III may be used for nested and, crucially, also non-nested model comparisons, for which simpler methods, such as the Savage-Dickey density ratio (Dickey & Lientz, 1970), cannot be applied. Importantly, Warp-III is not specific to hierarchical MPTs; it may

<sup>3</sup>For details on the predictive interpretation of the marginal likelihood see the Supplemental Materials available at <https://osf.io/rycg6/>.

be used to compute the marginal likelihood for a wide range of complex cognitive models. In fact, Warp-III improves upon simpler bridge sampling techniques (e.g., DiCiccio et al., 1997; Gronau, Sarafoglou, et al., 2017) by respecting potential skewness in the posterior distribution – a typical consequence of estimating parameters of cognitive models from scarce data (e.g., Ly et al., in press; Matzke et al., 2015). Due to its accuracy and relatively straightforward implementation, we believe that Warp-III is a promising and timely addition to the Bayesian toolkit of cognitive modelers in general, and MPT modelers in particular.

The chapter is organized as follows. We first introduce the latent-trait approach to hierarchical MPTs. We then demonstrate how Warp-III can be used to estimate the marginal likelihood for latent-trait MPTs. Lastly, we apply the method to two model comparison problems from published studies. The first example focuses on Bayesian model averaging for nested models; the second example focuses on the computation of the Bayes factor for non-nested models.

## 3.2 Multinomial Processing Trees

Data for MPTs consist of categorical responses<sup>4</sup> from several participants to a set of items. MPTs are based on the assumption that these responses follow a multinomial distribution. MPTs reparametrize the category probabilities of the multinomial distribution in terms of the model parameters that represent the probabilities of latent cognitive processes (Riefer & Batchelder, 1988).

Consider the pair-clustering MPT depicted in Figure 3.1. The model was developed for the measurement of the storage and retrieval processes that determine the recall of semantically related word pairs (Batchelder & Riefer, 1980). A typical pair-clustering study involves a free recall memory experiment, where participants are presented with a list of study words in a word-by-word fashion. The study list consists of two types of items: semantically related word pairs such as *knife-fork*, and words without a category partner (i.e., singletons), such as *dog*. After the study phase, participants are required to recall as many of the study words as they can. Typically, semantically related word pairs are recalled consecutively as a “pair-cluster”.

The model represents the interplay between the hypothesized latent cognitive processes in a rooted tree structure. The pair-clustering MPT features  $K = 2$  independent category systems. Each category system corresponds to a separate multinomial distribution: one for word pairs ( $k = 1$ ) and one for singletons ( $k = 2$ ). The category probabilities in each system are modeled using a separate subtree with a finite number of branches.

Each branch of a subtree corresponds to a specific sequence of processing stages and terminates in one of  $L_k$  possible response categories denoted as  $C_{kl}$ , where  $l = 1, \dots, L_k$  indexes the  $l$ th of  $L_k$  possible responses in subtree  $k$ . In the pair-clustering MPT, the recall of word pairs is scored into  $L_1 = 4$  categories: (1) both words of the pair are recalled consecutively ( $C_{11}$ ); (2) both words are recalled but not consecutively ( $C_{12}$ ); (3) only one word is recalled ( $C_{13}$ ); (4) no word is recalled

---

<sup>4</sup>Hu (2001), Heck and Erdfelder (2016), and Heck, Erdfelder, and Kieslich (2018) proposed extensions that also incorporate response times.

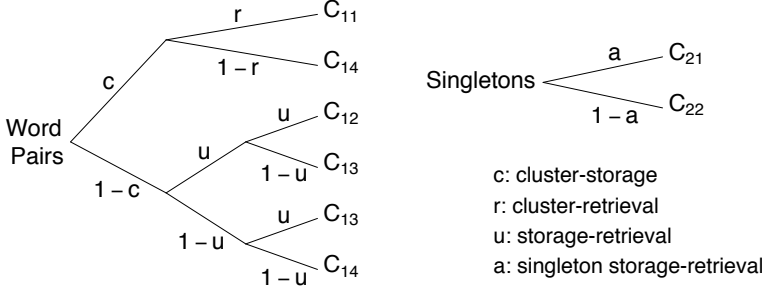


Figure 3.1: The pair-clustering MPT. Available at <https://tinyurl.com/yb7bma4e> under CC license <https://creativecommons.org/licenses/by/2.0/>.

( $C_{14}$ ). The recall of singletons is scored into  $L_2 = 2$  response categories: (1) the word is recalled ( $C_{21}$ ); (2) the word is not recalled ( $C_{22}$ ).

The response category probabilities are expressed as a function of the MPT parameters,  $\theta_p \in (0, 1) \quad \forall p \in \{1, 2, \dots, P\}$ , which can be collected in a vector  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_P)$ . The pair-clustering MPT features four parameters:  $\boldsymbol{\theta} = (c, r, u, a)$ . The *cluster-storage* parameter  $c$  corresponds to the probability that a word pair is stored as a cluster in memory. The *cluster-retrieval* parameter  $r$  corresponds to the conditional probability that a clustered word pair is retrieved from memory during the test phase. The model assumes that stored and retrieved word clusters are always recalled consecutively. The *storage-retrieval* parameter  $u$  corresponds to the conditional probability that a member of a word pair is stored and retrieved, given that the word pair was not clustered. The model makes the simplifying assumption that words from unclustered pairs are never recalled consecutively. The *singleton storage-retrieval* parameter  $a$  corresponds to the probability that a singleton is stored and retrieved. In many applications, researchers impose the constraint that  $a = u$ .

The response category probabilities are obtained as follows. First, we obtain the probability of each branch that terminates in a given response category. Let  $B_{klm}$  denote the  $m$ th of  $M_{kl}$  branches that terminate in response category  $C_{kl}$ . The probability of branch  $B_{klm}$  is obtained by traversing the tree from root to leaf and multiplying the encountered parameters:

$$\Pr(B_{klm} \mid \boldsymbol{\theta}) = \prod_{p=1}^P \theta_p^{v_{klmp}} (1 - \theta_p)^{w_{klmp}}, \quad (3.4)$$

where  $v_{klmp} \geq 0$  and  $w_{klmp} \geq 0$  are the number of nodes on branch  $B_{klm}$  that are related to parameter  $\theta_p$ ,  $p = 1, \dots, P$ , and  $1 - \theta_p$ , respectively. Second, we sum the probabilities of the  $M_{kl}$  branches that terminate in  $C_{kl}$ :

$$\Pr(C_{kl} \mid \boldsymbol{\theta}) = \sum_{m=1}^{M_{kl}} \Pr(B_{klm} \mid \boldsymbol{\theta}). \quad (3.5)$$

For instance, the probability of response category  $C_{14}$  is given by  $\Pr(C_{14} \mid \boldsymbol{\theta}) = c(1-r) + (1-c)(1-u)^2$ .

The probability of the observed response frequencies across category systems denoted by  $\mathbf{n} = (n_{11}, \dots, n_{1L_1}, \dots, n_{K1}, \dots, n_{KL_K})$ , where  $n_{kl}$  is the observed response frequency for category  $l = 1, \dots, L_k$  in category system (subtree)  $k = 1, \dots, K$ , is given by a product-multinomial distribution:

$$\Pr(\mathbf{N} = \mathbf{n} \mid \boldsymbol{\theta}) = \prod_{k=1}^K \left\{ \frac{J_k!}{n_{k1}! \times n_{k2}! \times \dots \times n_{kL_k}!} \prod_{l=1}^{L_k} [\Pr(C_{kl} \mid \boldsymbol{\theta})]^{n_{kl}} \right\}, \quad (3.6)$$

where  $J_k$  denotes the number of items in category system  $k$  (see also Klauer, 2010; Matzke et al., 2015).

### 3.2.1 Bayesian Hierarchical MPTs: The Latent-Trait Approach

Bayesian hierarchical approaches explicitly model heterogeneity in participants by introducing a group-level distribution from which the participant-level parameters are drawn (e.g., Gelman & Hill, 2007; Gill, 2002; Lee, 2011; Lee & Wagenmakers, 2013; Rouder & Lu, 2005).<sup>5</sup> Here we focus on Klauer’s (2010) latent-trait approach that relies on a multivariate normal group-level distribution to describe the between-subject variability and the correlations between the participant-level parameters.

To model participant heterogeneity, observed responses are aggregated over items, but not over participants, resulting in a vector of category frequencies for each participant  $i$ :  $\mathbf{n}_i$ ,  $i = 1, 2, \dots, I$ , where  $I$  is the total number of participants. Each participant obtains a participant-specific parameter vector  $\boldsymbol{\theta}_i$  of length  $P$ .

The latent-trait approach assumes that the probit-transformed participant-level parameter vectors  $\boldsymbol{\theta}'_i = \Phi^{-1}(\boldsymbol{\theta}_i)$  follow a  $P$ -dimensional multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ :  $\boldsymbol{\theta}'_i \sim \mathcal{N}_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . The probit-transformation  $\Phi^{-1}(\boldsymbol{\theta}_i)$  is defined component-wise, where  $\Phi^{-1}(\cdot)$  corresponds to the inverse of the cumulative distribution function of the normal distribution. Priors are assigned to  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ . We follow earlier implementations of the latent-trait approach and assign independent standard normal distributions to the  $P$  components of  $\boldsymbol{\mu}$  (Heck, Arnold, & Arnold, 2018; Matzke et al., 2015). This choice corresponds to uniform priors on the probability scale for the grand means. For the covariance matrix  $\boldsymbol{\Sigma}$ , a convenient prior choice would be an inverse Wishart prior with degrees of freedom  $\nu = P + 1$  and identity scale matrix. This setting leads to uniform priors on the correlation parameters; however, this choice is constraining on the standard deviation parameters. Although changing the degrees of freedom  $\nu$  affords more flexibility for modeling the standard deviations, it comes at the cost of constraining the prior on the correlation parameters (Gelman & Hill, 2007).

This dilemma can be circumvented by using a scaled inverse Wishart prior as introduced by Gelman and Hill (2007) and proposed in the context of hierarchical

---

<sup>5</sup>Bayesian hierarchical models can be also used to account for heterogeneity in items instead of participants.

MPT modeling by Klauer (2010). Compared to a regular inverse Wishart prior, the scaled version has the advantage that it allows one to model the standard deviations more flexibly while retaining the desirable uniform prior on the correlation parameters. The scaled inverse Wishart prior is based on the following decomposition of the covariance matrix  $\Sigma$ :

$$\Sigma = \text{Diag}(\boldsymbol{\xi}) \mathbf{Q} \text{Diag}(\boldsymbol{\xi}), \quad (3.7)$$

where  $\boldsymbol{\xi}$  is a vector of  $P$  scaling parameters and  $\mathbf{Q}$  corresponds to the  $P \times P$  unscaled covariance matrix. The scaled inverse Wishart prior is obtained by placing a regular inverse Wishart prior on the unscaled covariance matrix  $\mathbf{Q}$  and a suitable prior on the vector of scaling parameters  $\boldsymbol{\xi}$ .

We follow Klauer (2010) and assign  $\mathbf{Q}$  an inverse Wishart prior with degrees of freedom  $\nu = P + 1$  and scale matrix  $\mathbf{I}_P$  (i.e.,  $P \times P$  identity matrix). For the  $P$  components of  $\boldsymbol{\xi}$ , we follow Heck, Arnold, and Arnold (2018) and use independent uniform priors that range from zero to ten. These choices correspond to relatively diffuse priors for the standard deviations of the random effects on the probit scale and uniform priors for the correlations between the random effects.

Note that these prior distributions have been proposed in a context of parameter estimation, where the exact choice of the prior is irrelevant as long as sufficiently informative data are available. In contrast, in the context of model comparison, the priors have an important and lasting effect: As shown in Equation 3.3, the marginal likelihood is obtained by taking a weighted average of the probability of the data across all possible parameter settings where the weights correspond to the parameters' prior density. We argue that the standard normal and uniform priors for the grand means and the correlations, respectively, provide a reasonable default setting also from the perspective of model comparison. The choice of the prior for  $\boldsymbol{\xi}$  is less straightforward. We report the results corresponding to the default setting of the recently developed MPT software package **TreeBUGS** (Heck, Arnold, & Arnold, 2018), but we probed the robustness of our conclusions with a sensitivity analysis using  $\xi_p \sim \text{Uniform}(0, \xi_{\max}) \forall p \in \{1, 2, \dots, P\}$ , with  $\xi_{\max} = 2$  instead of  $\xi_{\max} = 10$ , a prior that was chosen based on the implied group-level distributions on the probability scale. As the conclusions were unaffected by the choice of the upper bound, the results of the sensitivity analysis are mentioned only briefly and are presented in more detail in the Supplemental Materials available at <https://osf.io/rycg6/>.

Under these prior settings, the probit-transformed participant-level MPT parameter vectors can be written as:

$$\boldsymbol{\theta}'_i = \boldsymbol{\mu} + \boldsymbol{\xi} \odot \boldsymbol{\omega}_i, \quad (3.8)$$

where  $\boldsymbol{\omega}_i$  is the  $P$ -dimensional vector with the unscaled random effects for participant  $i$ , and  $\odot$  denotes the Hadamard product (i.e., entry-wise multiplication, e.g., Liu & Trenkler, 2008). The unscaled random effects are drawn from a  $P$ -dimensional zero-centered multivariate normal distribution with covariance matrix  $\mathbf{Q}$ :  $\boldsymbol{\omega}_i \sim \mathcal{N}_P(\mathbf{0}, \mathbf{Q})$ .

Note that the model is overparameterized:  $\boldsymbol{\xi}$  and  $\mathbf{Q}$  cannot be interpreted separately. Similarly, the unscaled random effects  $\boldsymbol{\omega}_i$  cannot be interpreted on

their own but need to be combined with the scaling parameter vector  $\xi$  to form the random effects of interest. The scaling parameters  $\xi$ , the unscaled covariance matrix  $\mathbf{Q}$ , and the unscaled random effects  $\omega_i$  are not of interest in themselves and are simply an artifact of using a flexible scaled inverse Wishart prior on  $\Sigma$ : the parameters of interest are  $\theta'_i$ ,  $\mu$ , and  $\Sigma$ . Therefore, the scaled inverse Wishart prior can be regarded as a form of parameter expansion (e.g., Gelman & Hill, 2007) which has been reported to speed up convergence when fitting the model using Markov chain Monte Carlo sampling (MCMC; e.g., Gamerman & Lopes, 2006).

The reader is referred to Klauer (2010) and Matzke et al. (2015) for a more detailed description of the latent-trait approach. Parameter estimation may proceed using MCMC sampling implemented in standard Bayesian statistical software such as JAGS (Plummer, 2003) or Stan (Stan Development Team, 2016).

### 3.2.2 Computing the Marginal Likelihood

The marginal likelihood for latent-trait MPTs is given by:<sup>6</sup>

$$\begin{aligned}
 \Pr(\mathbf{N} = \mathbf{n}) &= \int \dots \int \prod_{i=1}^I \left[ \underbrace{\Pr(\mathbf{N}_i = \mathbf{n}_i \mid \mu, \xi, \omega_i)}_{\text{individual-level}} \underbrace{p(\omega_i \mid \mathbf{Q})}_{\text{group-level}} \right] \\
 &\quad \times \underbrace{p(\mathbf{Q})p(\mu)p(\xi)}_{\text{priors}} d\mathbf{Q} d\mu d\xi d\omega_1 \dots d\omega_I \\
 &= \int \dots \int \prod_{i=1}^I \left[ \underbrace{\prod_{k=1}^K \left\{ \frac{J_k!}{n_{ik1}! \times n_{ik2}! \times \dots \times n_{ikL_k}!} \prod_{l=1}^{L_k} [\Pr(C_{kl} \mid \mu, \xi, \omega_i)]^{n_{ikl}} \right\}}_{\Pr(\mathbf{N}_i = \mathbf{n}_i \mid \mu, \xi, \omega_i)} \right] \\
 &\quad \times \underbrace{(2\pi)^{-\frac{P}{2}} |\mathbf{Q}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \omega_i^\top \mathbf{Q}^{-1} \omega_i \right\}}_{p(\omega_i \mid \mathbf{Q})} \\
 &\quad \times \underbrace{\frac{1}{2^{\frac{\nu P}{2}} \Gamma_P(\frac{\nu}{2})} |\mathbf{Q}|^{-\frac{\nu+P+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{Q}^{-1}) \right\}}_{p(\mathbf{Q})} \\
 &\quad \times \underbrace{(2\pi)^{-\frac{P}{2}} \exp \left\{ -\frac{1}{2} \mu^\top \mu \right\}}_{p(\mu)} \underbrace{(\xi_{\max})^{-P}}_{p(\xi)} d\mathbf{Q} d\mu d\xi d\omega_1 \dots d\omega_I,
 \end{aligned} \tag{3.9}$$

where  $\Gamma_P(a) = \pi^{P(P-1)/4} \prod_{j=1}^P \Gamma(a + \frac{1-j}{2})$  and  $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$  are the multivariate and regular gamma function, respectively. In this parametrization, we do not need to explicitly integrate out the participant-level parameter vectors  $\theta_i$  since they are functions of  $\mu$ ,  $\xi$ , and  $\omega_i$  (see Equation 3.8).

---

<sup>6</sup>We omit conditioning on the model for enhanced legibility.

We exploit the fact that the covariance matrix  $\mathbf{Q}$  in Equation 3.9 can be integrated out in closed form (see also Overstall & Forster, 2010); a detailed derivation is provided in the Supplemental Materials. The marginal likelihood is then given by:

$$\begin{aligned} \Pr(\mathbf{N} = \mathbf{n}) = & \int \dots \int \prod_{i=1}^I \left[ \prod_{k=1}^K \left\{ \frac{J_k!}{n_{ik1}! \times n_{ik2}! \times \dots \times n_{ikL_k}!} \prod_{l=1}^{L_k} [\Pr(C_{kl} \mid \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\omega}_i)]^{n_{ikl}} \right\} \right] \\ & \times \frac{\Gamma_P(\frac{\nu+I}{2})}{\Gamma_P(\frac{\nu}{2})} \frac{\pi^{-\frac{IP}{2}}}{|\boldsymbol{\Omega}^\top \boldsymbol{\Omega} + \mathbf{I}_P|^{\frac{\nu+I}{2}}} \times (2\pi)^{-\frac{P}{2}} \exp \left\{ -\frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\mu} \right\} \\ & \times (\xi_{\max})^{-P} d\boldsymbol{\mu} d\boldsymbol{\xi} d\boldsymbol{\omega}_1 \dots d\boldsymbol{\omega}_I, \end{aligned} \quad (3.10)$$

where  $\boldsymbol{\Omega}$  is an  $I \times P$  matrix of the  $P$ -dimensional random-effects vectors  $\boldsymbol{\omega}_i$  of the  $I$  participants. Even after integrating out  $\mathbf{Q}$  the expression for the marginal likelihood is still a high-dimensional integral (i.e.,  $P(I+2)$  dimensions); the challenge is to find a method which yields accurate estimates of this integral.

### 3.3 Warp-III Bridge Sampling for MPTs

We propose to use Warp-III bridge sampling (Meng & Schilling, 2002; Meng & Wong, 1996; Overstall, 2010), an advanced version of bridge sampling, to evaluate the high-dimensional integral in Equation 3.10. Bridge sampling is a general method for estimating normalizing constants<sup>7</sup>, a problem that is not only encountered in Bayesian inference, but also in likelihood-based approaches (Gelman & Meng, 1998). We first outline the basic principles of bridge sampling, and then present the details of the advanced Warp-III method. The reader is referred to the recent tutorial by Gronau, Sarafoglou, et al. (2017) for a detailed explanation of the general bridge sampling approach.

Let  $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_I)$  be the vector of quantities that must be integrated out to obtain the marginal likelihood, so that

$$\Pr(\mathbf{N} = \mathbf{n}) = \int \Pr(\mathbf{N} = \mathbf{n} \mid \boldsymbol{\zeta}) p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}. \quad (3.11)$$

<sup>7</sup>Bridge sampling in its original form has been proposed to estimate a *ratio* of normalizing constants. This approach, however, becomes challenging and inefficient in case the two models have different parameter spaces (e.g., non-nested comparisons), and potentially very little overlap between the posterior distributions. For these cases, it may be easier and more efficient to compute each normalizing constant separately (e.g., DiCiccio et al., 1997; Overstall & Forster, 2010). This ensures that the two relevant distributions (i.e., proposal and posterior) for each of the separate bridge sampling applications are close to each other yielding an efficient estimator. Therefore, we recommend computing each normalizing constant separately to enable application of the method to a wide range of model comparison scenarios.

General bridge sampling is based on the following identity:

$$1 = \frac{\int \overbrace{h(\zeta)}^{\text{bridge function}} \underbrace{p(\zeta \mid \mathbf{N} = \mathbf{n})}_{\text{posterior distribution}} \overbrace{g(\zeta)}^{\text{proposal distribution}} d\zeta}{\int h(\zeta) \underbrace{p(\zeta \mid \mathbf{N} = \mathbf{n})}_{\text{posterior distribution}} g(\zeta) d\zeta}, \quad (3.12)$$

where  $p(\zeta \mid \mathbf{N} = \mathbf{n})$  is the posterior distribution of  $\zeta$ ,  $g(\zeta)$  is the probability density function of a proposal distribution, and  $h(\zeta)$  is a function such that  $0 < |\int h(\zeta) p(\zeta \mid \mathbf{N} = \mathbf{n}) g(\zeta) d\zeta| < \infty$ . It follows from Equation 3.12 that

$$\begin{aligned} \Pr(\mathbf{N} = \mathbf{n}) &= \frac{\int h(\zeta) \Pr(\mathbf{N} = \mathbf{n} \mid \zeta) p(\zeta) g(\zeta) d\zeta}{\int h(\zeta) g(\zeta) p(\zeta \mid \mathbf{N} = \mathbf{n}) d\zeta} \\ &= \frac{\mathbb{E}_{g(\zeta)} [h(\zeta) \Pr(\mathbf{N} = \mathbf{n} \mid \zeta) p(\zeta)]}{\mathbb{E}_{p(\zeta \mid \mathbf{N} = \mathbf{n})} [h(\zeta) g(\zeta)]}. \end{aligned} \quad (3.13)$$

The bridge sampling estimate of the marginal likelihood is then obtained by sampling from  $g(\zeta)$  and  $p(\zeta \mid \mathbf{N} = \mathbf{n})$  and then using Monte Carlo approximations to estimate the expected values.

The optimal choice of  $h(\zeta)$ , one that minimizes the relative mean-squared error of the estimator, is given by:

$$h_o(\zeta) \propto [s_1 \Pr(\mathbf{N} = \mathbf{n} \mid \zeta) p(\zeta) + s_2 \Pr(\mathbf{N} = \mathbf{n}) g(\zeta)]^{-1}, \quad (3.14)$$

where  $s_i = \frac{D_i}{D_1 + D_2}$ ,  $i \in \{1, 2\}$ ,  $D_1$  and  $D_2$  denote the number of draws from  $p(\zeta \mid \mathbf{N} = \mathbf{n})$  and  $g(\zeta)$ , respectively, used to approximate the expected values (Meng & Wong, 1996). We set  $D_1 = D_2$ . Note that  $h_o$  is only optimal if the draws from the posterior distribution are independent which is not the case with MCMC procedures. To account for this fact, we replace  $D_1$  in defining the weights  $s_1$  and  $s_2$  by the effective sample size obtained using the `coda` R package (Plummer et al., 2006).<sup>8</sup> As  $h_o(\zeta)$  depends on  $\Pr(\mathbf{N} = \mathbf{n})$ , the very quantity we want to estimate, we follow Meng and Wong (1996) and use an iterative scheme to update an initial guess of the marginal likelihood until convergence:<sup>9</sup>

$$\hat{\Pr}(\mathbf{N} = \mathbf{n})^{(t+1)} = \frac{\frac{1}{D_2} \sum_{r=1}^{D_2} \frac{l_{2,r}}{s_1 l_{2,r} + s_2} \hat{\Pr}(\mathbf{N} = \mathbf{n})^{(t)}}{\frac{1}{D_1} \sum_{j=1}^{D_1} \frac{1}{s_1 l_{1,j} + s_2} \hat{\Pr}(\mathbf{N} = \mathbf{n})^{(t)}}, \quad (3.15)$$

---

<sup>8</sup>Specifically, we used the median effective sample size across all posterior components.

<sup>9</sup>In our experience, the exact value of the initial guess typically does not have a lasting influence on the resulting estimate. Nevertheless, good initial values may lead to faster convergence. For implementation details, see Gronau, Sarafoglou, et al. (2017), especially Appendix B.

where  $l_{1,j} = \frac{\Pr(\mathbf{N}=\mathbf{n}|\zeta_j^*)p(\zeta_j^*)}{g(\zeta_j^*)}$ ,  $l_{2,r} = \frac{\Pr(\mathbf{N}=\mathbf{n}|\tilde{\zeta}_r)p(\tilde{\zeta}_r)}{g(\tilde{\zeta}_r)}$ ,  $\{\zeta_1^*, \dots, \zeta_{D_1}^*\}$  are  $D_1$  draws from  $p(\zeta | \mathbf{N} = \mathbf{n})$ , and  $\{\tilde{\zeta}_1, \dots, \tilde{\zeta}_{D_2}\}$  are  $D_2$  draws from  $g(\zeta)$ .

A remaining question is how to choose  $g(\zeta)$ . The precision of the bridge sampling estimator is governed by the number of samples from  $g(\zeta)$  and the overlap between  $g(\zeta)$  and  $p(\zeta | \mathbf{N} = \mathbf{n})$  (Meng & Wong, 1996). Therefore,  $g(\zeta)$  should closely resemble the posterior distribution. For instance, we may choose a multivariate normal distribution for  $g$  with mean vector and covariance matrix that match the corresponding quantities of the posterior samples. Although the multivariate normal approach works well in many applications (e.g., Gronau, Sarafoglou, et al., 2017; Overstall & Forster, 2010), it can be inefficient when the posterior distribution is skewed.

Warp-III improves upon the multivariate normal bridge sampling approach by matching, not only the first two, but also the third moment (i.e., skewness) of  $g$  and the posterior distribution. Consequently, in case there is no skewness, Warp-III results in estimates with the same precision as the ones from the simpler multivariate normal approach. However, crucially, in the presence of skewness, Warp-III is able to match  $g$  and the posterior distribution more closely which results in a higher precision of the marginal likelihood estimates compared to the simpler approach. How much of an improvement Warp-III is over the simpler multivariate normal approach may depend on the particular example at hand.

In Warp-III,  $g$  is fixed to a multivariate standard normal distribution. The posterior distribution is then manipulated – “warped” – so that its mean vector, covariance matrix, and skew match  $g$ . Crucially, the warped posterior distribution retains the normalizing constant of the posterior distribution. Figure 3.2 illustrates the rationale of the Warp-III transformation for the univariate case. The histogram in the upper-left panel shows hypothetical “unbounded” posterior samples that can range across the entire real line; the solid line shows the standard normal proposal distribution  $g$ . The overlap between the two distributions is clearly suboptimal. Bridge sampling applied to these two distributions can be thought of as “Warp-0” because the posterior distribution is not modified. The upper-right panel illustrates “Warp-I”: Subtracting the mean of the posterior samples from all posterior samples matches the first moment of the distributions. The lower-right panel illustrates “Warp-II”: Dividing the zero-centered posterior samples by their standard deviation matches the first two moments of the distributions. This approach is practically equivalent to the multivariate normal bridge sampling approach described above. Lastly, the lower-left panel illustrates Warp-III: Randomly assigning a minus sign to the standardized posterior samples matches also the third moment of the distributions.

Warp-III assumes that all components of the parameter vector can range across the entire real line. In the context of latent-trait MPTs, this assumption is not fulfilled since  $\xi_p \in (0, \xi_{\max}) \forall p \in \{1, \dots, P\}$ . We therefore transform  $\xi$  so that  $\xi_{\text{trans}} = \Phi^{-1}\left(\frac{\xi}{\xi_{\max}}\right)$  with Jacobian  $(\xi_{\max})^P \mathcal{N}_P(\xi_{\text{trans}}; \mathbf{0}, \mathbf{I}_P)$ , where  $\mathcal{N}_P(\mathbf{x}; \mathbf{y}, \mathbf{Z})$  denotes the probability density function of a  $P$ -dimensional normal distribution with mean vector  $\mathbf{y}$  and covariance matrix  $\mathbf{Z}$  which is evaluated for the vector

### 3. BAYESIAN MODEL COMPARISON FOR HIERARCHICAL MULTINOMIAL PROCESSING TREE MODELS USING WARP-III BRIDGE SAMPLING

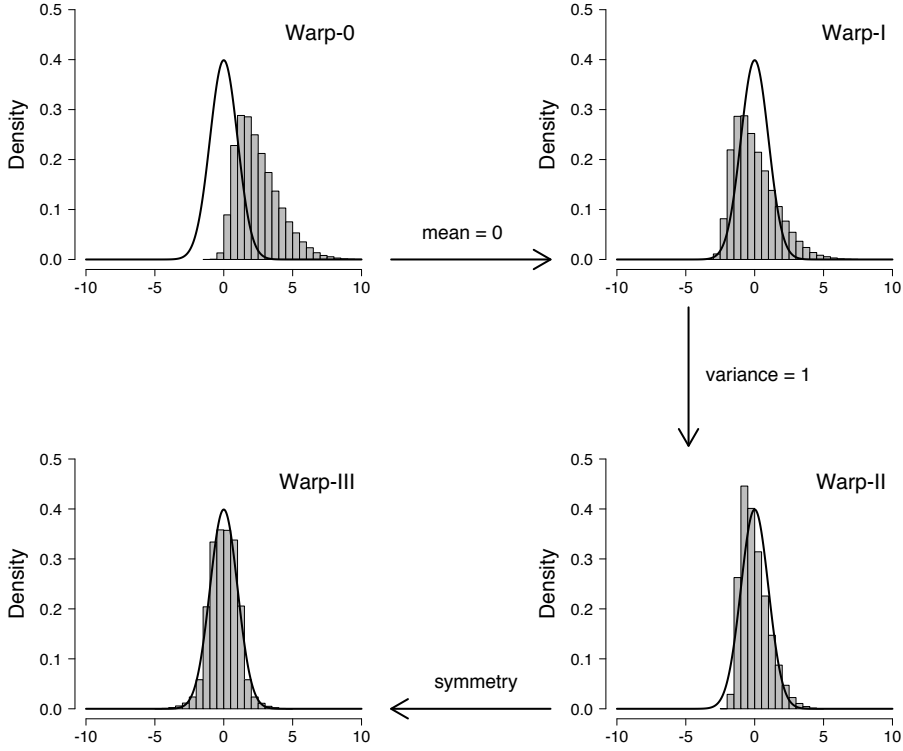


Figure 3.2: Matching the proposal and posterior distribution with warping. Histograms show the posterior distribution; density lines show the standard normal proposal distribution. Available at <https://tinyurl.com/y7owvsz3> under CC license <https://creativecommons.org/licenses/by/2.0/>.

$\mathbf{x}$ .<sup>10</sup> Let  $\boldsymbol{\psi} = (\boldsymbol{\mu}, \boldsymbol{\xi}_{\text{trans}}, \boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_I)$  denote the resulting parameter vector where all components are on the real line.

Warp-III is then based on applying the following stochastic transformation to  $\boldsymbol{\psi}$ :

$$\boldsymbol{\eta} = \underbrace{b}_{\text{symmetry}} \times \underbrace{\mathbf{R}^{-1}}_{\text{covariance}} \times \underbrace{(\boldsymbol{\psi} - \mathbf{v})}_{\text{mean } \mathbf{0}}, \quad (3.16)$$

where  $b \sim \text{Bernoulli}(0.5)$  on  $\{-1, 1\}$  and  $\mathbf{v}$  corresponds to the expected value of  $\boldsymbol{\psi}$  (i.e., the mean vector). The matrix  $\mathbf{R}$  is obtained via the Cholesky decomposition of the covariance matrix of  $\boldsymbol{\psi}$ , denoted as  $\mathbf{S}$ , thus,  $\mathbf{S} = \mathbf{R}\mathbf{R}^\top$ . In practice,  $\mathbf{v}$  and  $\mathbf{S}$  are unknown and must be approximated using the posterior samples. Note

<sup>10</sup>As before, the probit-transformation is defined component-wise

that Equation 3.16 simply generalizes the intuition illustrated in Figure 3.2 for the univariate case to the general case with multiple parameters.

Due to the Bernoulli random variable  $b$ , the warped posterior density has the form of a mixture density (see also Overstall, 2010, p. 70):

$$\begin{aligned} p_{\boldsymbol{\eta}}(\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n}) &= \frac{|\mathbf{R}|}{2} \left[ \frac{\tilde{p}_{\psi}(\mathbf{v} - \mathbf{R}\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n})}{\Pr(\mathbf{N} = \mathbf{n})} + \frac{\tilde{p}_{\psi}(\mathbf{v} + \mathbf{R}\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n})}{\Pr(\mathbf{N} = \mathbf{n})} \right] \\ &= \frac{\tilde{p}_{\boldsymbol{\eta}}(\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n})}{\Pr(\mathbf{N} = \mathbf{n})}, \end{aligned} \quad (3.17)$$

where  $\tilde{p}_{\boldsymbol{\eta}}(\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n}) = \frac{|\mathbf{R}|}{2} [\tilde{p}_{\psi}(\mathbf{v} - \mathbf{R}\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n}) + \tilde{p}_{\psi}(\mathbf{v} + \mathbf{R}\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n})]$  denotes the un-normalized warped posterior distribution and  $\tilde{p}_{\psi}(\cdot \mid \mathbf{N} = \mathbf{n})$  denotes the un-normalized posterior distribution that has been transformed to the real line (but not warped). This proves that the warped posterior distribution retains the normalizing constant of the original posterior distribution.

The Warp-III estimator of the marginal likelihood is then derived by using the warped posterior distribution  $p_{\boldsymbol{\eta}}(\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n})$  instead of  $p(\boldsymbol{\zeta} \mid \mathbf{N} = \mathbf{n})$  in Equation 3.12. Equation 3.13 shows that this results in a ratio of two expected values, where the numerator is an expected value with respect to the multivariate standard normal proposal distribution  $g(\boldsymbol{\eta})$  and the denominator is an expected value with respect to the warped posterior distribution  $p_{\boldsymbol{\eta}}(\boldsymbol{\eta} \mid \mathbf{N} = \mathbf{n})$ . Hence, we could obtain an estimate of the marginal likelihood by first warping the posterior samples using Equation 3.16, then sampling from the proposal distribution, and applying the iterative updating scheme in Equation 3.15.

However, in line with the literature (e.g., Sinharay & Stern, 2005), we rewrite the expected value in the denominator of Equation 3.13 in terms of the unbounded posterior samples that are transformed to the real line but are not warped; a derivation is provided in the Supplemental Materials. The estimate of the marginal likelihood is then obtained by applying the iterative scheme in Equation 3.15 using:

$$l_{1,j} = \frac{\frac{|\mathbf{R}|}{2} [\tilde{p}_{\psi}(2\mathbf{v} - \boldsymbol{\psi}_j^* \mid \mathbf{N} = \mathbf{n}) + \tilde{p}_{\psi}(\boldsymbol{\psi}_j^* \mid \mathbf{N} = \mathbf{n})]}{g(\mathbf{R}^{-1}(\boldsymbol{\psi}_j^* - \mathbf{v}))}, \quad (3.18)$$

and

$$l_{2,r} = \frac{\frac{|\mathbf{R}|}{2} [\tilde{p}_{\psi}(\mathbf{v} - \mathbf{R}\tilde{\boldsymbol{\eta}}_r \mid \mathbf{N} = \mathbf{n}) + \tilde{p}_{\psi}(\mathbf{v} + \mathbf{R}\tilde{\boldsymbol{\eta}}_r \mid \mathbf{N} = \mathbf{n})]}{g(\tilde{\boldsymbol{\eta}}_r)}, \quad (3.19)$$

where  $\{\boldsymbol{\psi}_1^*, \dots, \boldsymbol{\psi}_{D_1}^*\}$  are  $D_1$  draws from  $p_{\psi}(\boldsymbol{\psi} \mid \mathbf{N} = \mathbf{n})$ , and  $\{\tilde{\boldsymbol{\eta}}_1, \dots, \tilde{\boldsymbol{\eta}}_{D_2}\}$  are  $D_2$  draws from the proposal distribution  $g(\boldsymbol{\eta})$ . Furthermore,  $\tilde{p}_{\psi}(\boldsymbol{\psi} \mid \mathbf{N} = \mathbf{n})$  denotes the un-normalized posterior density of the unbounded posterior samples;

it is therefore written in terms of  $\boldsymbol{\xi}_{\text{trans}}$  and is adjusted by the Jacobian term:<sup>11</sup>

$$\begin{aligned} \tilde{p}_{\psi}(\boldsymbol{\psi} \mid \mathbf{N} = \mathbf{n}) = & \prod_{i=1}^I \left[ \prod_{k=1}^K \left\{ \frac{J_k!}{n_{ik1}! \times n_{ik2}! \times \dots \times n_{ikL_k}!} \prod_{l=1}^{L_k} [\Pr(C_{kl} \mid \boldsymbol{\mu}, \boldsymbol{\xi}_{\text{trans}}, \boldsymbol{\omega}_i)]^{n_{ikl}} \right\} \right] \\ & \times \frac{\Gamma_P(\frac{\nu+I}{2})}{\Gamma_P(\frac{\nu}{2})} \frac{\pi^{-\frac{IP}{2}}}{|\boldsymbol{\Omega}^{\top} \boldsymbol{\Omega} + \mathbf{I}_P|^{\frac{\nu+I}{2}}} \times (2\pi)^{-\frac{P}{2}} \exp \left\{ -\frac{1}{2} \boldsymbol{\mu}^{\top} \boldsymbol{\mu} \right\} \\ & \times (2\pi)^{-\frac{P}{2}} \exp \left\{ -\frac{1}{2} \boldsymbol{\xi}_{\text{trans}}^{\top} \boldsymbol{\xi}_{\text{trans}} \right\}. \end{aligned} \quad (3.20)$$

Note that rewriting the expected value in terms of  $\tilde{p}_{\psi}(\boldsymbol{\psi} \mid \mathbf{N} = \mathbf{n})$  is only a technical nicety. This approach is identical to applying the Warp-III transformation to the posterior samples and then using the iterative scheme with the warped posterior density and a multivariate standard normal proposal distribution.

### 3.4 Empirical Examples

#### 3.4.1 Example 1: Nested Model Comparison

We re-analyzed the pair-clustering data set reported in Riefer, Knapp, Batchelder, Bamber, and Manifold (2002) using the hierarchical latent-trait approach.<sup>12</sup> Experiment 4 examined the memory of patients with brain damage due to prolonged alcoholism in comparison to a control group of alcoholic patients without indications of brain damage. The participants attempted to memorize the same list of 20 categorically related word pairs in a series of six study-test trials.<sup>13</sup> For demonstration purposes, we focused on the free recall performance of the 21 control participants. Specifically, we investigated whether the model parameters change from the first to the second trial indicating a change in the storage and retrieval processes as a function of practice using posterior model probabilities and Bayesian model averaging.

##### 3.4.1.1 Model Specification

To model differences in parameters, we augmented Equation 3.8 with a parameter vector that captures the difference in parameters between the two trials:  $\boldsymbol{\delta} = (\delta_c, \delta_r, \delta_u)$ . The probit-transformed parameter vectors of participant  $i$  for the first

---

<sup>11</sup>Note that  $\xi_{\text{max}}$  drops out of the expression because it cancels with the first term of the Jacobian. Implicitly, however, it still influences the marginal likelihood because it appears in the transformation equation  $\boldsymbol{\xi}_{\text{trans}} = \Phi^{-1}\left(\frac{\boldsymbol{\xi}}{\xi_{\text{max}}}\right)$ . It is also needed for evaluating  $\Pr(C_{kl} \mid \boldsymbol{\mu}, \boldsymbol{\xi}_{\text{trans}}, \boldsymbol{\omega}_i)$  since in order to obtain the MPT parameters on the probit scale (i.e., Equation 3.8) we need to transform  $\boldsymbol{\xi}_{\text{trans}}$  back to  $\boldsymbol{\xi}$  via the inverse transformation  $\boldsymbol{\xi} = \xi_{\text{max}} \Phi(\boldsymbol{\xi}_{\text{trans}})$ .

<sup>12</sup>Data were obtained from <https://bayesmodels.com/>; see also Lee and Wagenmakers (2013).

<sup>13</sup>Riefer et al. (2002) did not administer singletons.

Table 3.1: Overview of the eight nested models for the analysis of the first two trials of the pair-clustering data set reported in Riefer et al. (2002).

Free Parameters	Model							
	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$
$c$	✓		✓	✓		✓		
$r$	✓	✓		✓			✓	
$u$	✓	✓	✓		✓			

Note.  $\mathcal{M}_1$  allows all three parameters to vary between trials,  $\mathcal{M}_8$  posits that none of the parameters vary between trials. Models  $\mathcal{M}_2$  to  $\mathcal{M}_7$  are between these extremes.

trial  $(\theta'_{1,i})$  and the second trial  $(\theta'_{2,i})$  are then obtained as follows:

$$\begin{aligned}
 \theta'_{1,i} &= \underbrace{\mu - \frac{\delta}{2}}_{\substack{\text{group mean} \\ \text{for first trial}}} + \xi \odot \omega_i, \\
 \theta'_{2,i} &= \underbrace{\mu + \frac{\delta}{2}}_{\substack{\text{group mean} \\ \text{for second trial}}} + \xi \odot \omega_i.
 \end{aligned} \tag{3.21}$$

For an alternative approach to modeling within-subject differences in model parameters, the reader is referred to Rouder et al. (2008).

Table 3.1 shows the  $2^3 = 8$  nested models that implement the eight sets of possible parameter constraints.  $\mathcal{M}_1$  allows all three parameters to vary between trials so that  $\delta = (\delta_c, \delta_r, \delta_u)$ . In contrast,  $\mathcal{M}_8$  posits that none of the parameters vary between trials so that  $\delta = (0, 0, 0)$ . Models  $\mathcal{M}_2$  to  $\mathcal{M}_7$  are between these extremes and allow either one or two parameters to vary between trials.

We used independent zero-centered normal priors for the components of  $\delta$ . We explored a narrow ( $\sigma_\delta^{\text{narrow}} \approx 0.52$ ), medium ( $\sigma_\delta^{\text{medium}} \approx 0.84$ ), and a wide ( $\sigma_\delta^{\text{wide}} \approx 1.28$ ) zero-centered normal prior to assess the sensitivity of the results to the width of the test-relevant prior distribution. As shown in the Supplemental Materials, the standard deviations  $\sigma_\delta$  were chosen to correspond to small, medium, and large effects on the probability scale centered around 0.5. Priors for the remaining parameters followed the specification described earlier.

We estimated the posterior distribution of the model parameters using JAGS by adapting the script provided by Matzke et al. (2015). The JAGS code is available in the Supplemental Materials. We ran three MCMC chains with over-dispersed start values, discarded the first 4,000 posterior samples as burn in, and retained only every 20th sample to reduce autocorrelation. Results reported below are based on a total of 90,000 posterior samples. Convergence of the MCMC chains was assessed by visual inspection and the  $\hat{R}$  statistic ( $\hat{R} < 1.05$  for all parameters; Gelman & Rubin, 1992).

Figure 3.3 shows the resulting posterior distributions of the probit group-level means from the full model  $\mathcal{M}_1$ ; the parameters were transformed back to the probability scale. The posteriors were computed using the medium prior setting

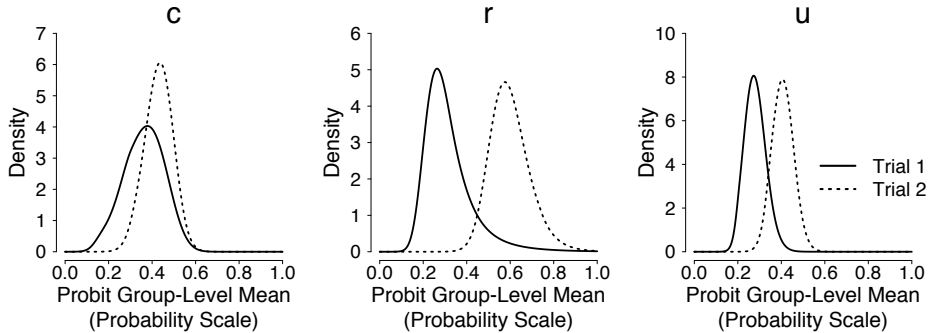


Figure 3.3: Posterior distributions of the probit group-level means (plotted on the probability scale) from the full model  $\mathcal{M}_1$  for the analysis of the first two trials of the pair-clustering data reported in Riefer et al. (2002). The solid lines correspond to the posteriors for the first trial, the dotted lines to the posteriors for the second trial. Available at <https://tinyurl.com/y9a3314t> under CC license <https://creativecommons.org/licenses/by/2.0/>.

( $\sigma_{\delta}^{\text{medium}}$ ) – results obtained with the narrow and wide prior were highly similar and are not displayed. The plot of the posterior distributions based on the alternative prior choice for the elements of  $\xi$  (i.e., uniform priors with upper bound  $\xi_{\max} = 2$  instead of  $\xi_{\max} = 10$ ) was visually almost indistinguishable from the one presented here and has hence been relegated to the Supplemental Materials. The cluster-storage  $c$  parameter did not change substantially, whereas the storage-retrieval  $u$ , and especially the cluster-retrieval  $r$  parameter seemed to increase from the first trial to the second.

#### 3.4.1.2 Computing Marginal Likelihoods with Warp-III

Equation 3.20 was adjusted to include the relevant prior distributions for the elements of  $\delta$ . For each model, we split the 90,000 posterior samples in two equal parts (first and second half of the iterations per chain) and used the first part for estimating  $\mathbf{R}$  and  $\mathbf{v}$ , and the second part for the iterative updating scheme in Equation 3.15 (Overstall & Forster, 2010). Hence,  $D_1 = D_2 = 45,000$ . To assess the accuracy of the resulting estimates, we repeated this procedure 50 times.<sup>14</sup> We implemented the procedure in R (R Core Team, 2019). For efficiency, we parallelized the computations, and coded the computationally intensive elements in efficient C++ code which was called from within R using Rcpp (Eddelbuettel et al., 2011). Using a standard personal computer and four CPU cores, computing the marginal likelihood for each repetition took less than one minute per model. The code is available in the Supplemental Materials.

<sup>14</sup>We assessed the accuracy of the estimates conditional on the posterior samples, that is, for each repetition, we used the same posterior samples but generated new samples from the proposal distribution. Whenever feasible, it may be advantageous to also generate new posterior samples in each repetition.

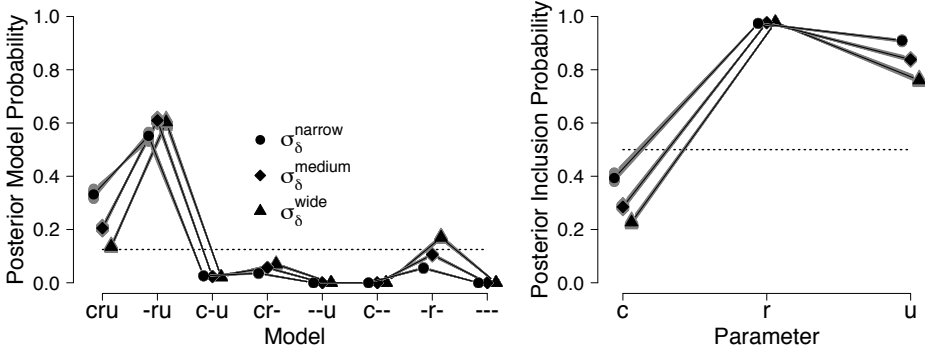


Figure 3.4: Posterior model probabilities (left panel) and posterior inclusion probabilities (right panel) for the analysis of the first two trials of the pair-clustering data reported in Riefer et al. (2002) obtained with Warp-III bridge sampling. In the left panel, the  $x$ -axis indicates which parameters were allowed to vary from the first to the second trial (e.g.,  $c - u$  corresponds to  $\mathcal{M}_3$  where  $r$  was fixed between trials). Gray symbols show the results of the 50 repetitions and black symbols display the posterior model probabilities and posterior inclusion probabilities that are based on the median of the 50 estimated log marginal likelihoods. Circles show results obtained with the narrow prior, diamonds with the medium prior, and triangles with the wide prior. The dotted lines show the prior model probabilities and prior inclusion probabilities. Available at <https://tinyurl.com/yaxbj9o6> under CC license <https://creativecommons.org/licenses/by/2.0/>.

### 3.4.1.3 Posterior Model Probabilities

To formally quantify evidence for the differences in parameters, we computed the posterior model probabilities of the eight models using the marginal likelihoods obtained with Warp-III. We assumed that all models were equally likely a priori. The left panel of Figure 3.4 shows the posterior model probabilities for the narrow, medium, and wide prior settings. The plot of the posterior model probabilities based on the alternative prior choice for the elements of  $\xi$  (i.e., uniform priors with upper bound  $\xi_{\max} = 2$  instead of  $\xi_{\max} = 10$ ) was visually almost indistinguishable from the one presented here and has hence been relegated to the Supplemental Materials. Formal model comparison confirmed the results of the visual inspection of the posterior distributions shown in Figure 3.3:  $\mathcal{M}_2$ , the model that allows for a difference in  $r$  and  $u$ , received the most support from the data. As expected, the width of the test-relevant prior  $\delta$  influenced the value of the marginal likelihood, but it did not change the conclusions qualitatively. Warp-III provided accurate estimates of the posterior model probabilities as indicated by the small variability across the 50 repetitions (i.e., gray symbols). For this nested example, the posterior model probabilities can be also obtained using the Savage-Dickey density ratio representation of the Bayes factor (Dickey & Lientz, 1970; Wagenmakers et

al., 2010). As shown in the Supplemental Materials, the Savage-Dickey procedure resulted in posterior model probabilities that were highly similar to the ones obtained with Warp-III.

#### 3.4.1.4 Bayesian Model Averaging

Bayesian model averaging does not require researchers to commit to a single “best” model; it allows researchers to acknowledge uncertainty about the choice of the correct model (e.g., Hoeting et al., 1999; Rouder et al., 2017). This is achieved by considering the posterior inclusion probabilities of the parameters. Posterior inclusion probabilities quantify the model-averaged evidence for a change in a given parameter; they can be obtained by summing the posterior model probabilities of the models that allow the parameter to differ between the trials. For instance, the posterior inclusion probability of the  $c$  parameter is obtained by summing the posterior model probabilities of  $\mathcal{M}_1$ ,  $\mathcal{M}_3$ ,  $\mathcal{M}_4$ , and  $\mathcal{M}_6$ . Posterior inclusion probabilities are then compared to the prior inclusion probabilities, in this case 0.5, which are obtained in an analogous manner but based on the prior model probabilities.<sup>15</sup> The right panel of Figure 3.4 shows the posterior inclusion probabilities for the three prior settings. The plot of the posterior inclusion probabilities based on the alternative prior choice for the elements of  $\xi$  (i.e., uniform priors with upper bound  $\xi_{\max} = 2$  instead of  $\xi_{\max} = 10$ ) was visually almost indistinguishable from the one presented here and has hence been relegated to the Supplemental Materials. The posterior inclusion probabilities of the  $r$  and  $u$  parameter are higher than the prior inclusion probabilities, indicating evidence for a difference in these parameters between trials. In contrast, the posterior inclusion probability of  $c$  is lower than the corresponding prior inclusion probability, indicating evidence for invariance between the trials. As before, the width of the  $\delta$  prior does not change the conclusions qualitatively.

#### 3.4.1.5 Substantive Contribution

The data from Riefer et al. (2002) have been analyzed in a number of articles. The original article analyzed the aggregated data (an approach known to suffer from limitations in case there is heterogeneity across participants, e.g., Klauer, 2006) and considered the  $p$ -values of  $G^2$  statistics to investigate whether parameters differ across trials. J. B. Smith and Batchelder (2010) reanalyzed a subset of the data using the hierarchical beta-MPT model (which specifies group-level beta distributions and thus differs from the latent-trait approach that we used).<sup>16</sup> To investigate whether parameters differ across trials, Smith and Batchelder (a) considered the posterior distribution of the difference between trials for the group-level mean parameters and (b) ran a classical paired sample  $t$ -test on the individual-level parameter estimates. These approaches, however, do not allow one to quantify evidence for an invariance (i.e., a simpler model where some parameters do not differ

---

<sup>15</sup>The change from prior inclusion *odds* to posterior inclusion *odds* can also be quantified by means of an inclusion Bayes factor (not reported).

<sup>16</sup>Note that this data set has been also analyzed in Lee and Wagenmakers (2013, chapter 14). In this case the hierarchical latent-trait approach was used, however, no explicit model comparison or hypothesis testing was conducted.

across trials) on a continuous scale in a systematic way and, crucially, they do not allow one to disentangle “absence of evidence” (i.e., the data are uninformative) and “evidence of absence” (i.e., the data support a simpler model).<sup>17</sup> These shortcomings can be addressed by computing Bayes factors and posterior model and posterior inclusion probabilities. “Absence of evidence” can be inferred from Bayes factors close to one and posterior model and posterior inclusion probabilities close to the corresponding prior probabilities. In contrast, “evidence of absence” can be inferred from large Bayes factors in favor of the simpler model, and in situations when the posterior model probability of the simpler model is the highest or when the posterior inclusion probability is smaller than the prior inclusion probability.

Our Bayesian re-analysis suggests that there is strong evidence that the probability of retrieving word pairs that have been stored as a cluster (i.e.,  $r$ ) changed from the first to the second trial. Furthermore, there is evidence that the probability of storing and retrieving words that have not been stored as a cluster (i.e.,  $u$ ) differed between the two trials. Crucially, our approach also allowed us to conclude that there is some evidence that the probability of storing a word pair as a cluster (i.e.,  $c$ ) did *not* change from the first to the second trial (although this evidence is not that pronounced since the posterior inclusion probability for a difference in  $c$  is – depending on the prior choice – relatively close to the prior inclusion probability of .5). Another key improvement of our analysis over the above mentioned analyses is the use of Bayesian model averaging. In this example,  $\mathcal{M}_2$  received the highest posterior probability; however,  $\mathcal{M}_1$  also received substantive posterior probability. Therefore, selecting a single best model (i.e.,  $\mathcal{M}_2$ ) and basing final inference solely on this model might be suboptimal at best and misleading at worst. In contrast, when using the model-averaged posterior inclusion probabilities for drawing conclusions about which parameters differ between trials, one takes into account all models under consideration according to their plausibilities in light of the observed data.

Finally, note that one might argue that this data set is relatively small and is thus uninformative. However, one strength of the Bayesian approach is that it allows one to quantify whether the data are informative or not. For this example, the Bayesian results suggest that the data are in fact informative which is indicated by posterior model/inclusion probabilities that are quite different from the corresponding prior probabilities.

### 3.4.2 Example 2: Non-Nested Model Comparison

We re-analyzed data from Experiment 2 reported by Fazio et al. (2015) who investigated the influence of knowledge on the illusory truth effect. The illusory truth effect refers to the phenomenon that, in the absence of knowledge about the truth status of a statement, repeated statements are easier to process and are judged more truthful than new statements. Fazio et al., however, provided evidence that participants tend to rely on the ease of processing (i.e., fluency) even when they have knowledge about the statement.

<sup>17</sup>Note also that it is well-known that the two-step procedure (b) used by J. B. Smith and Batchelder can yield biased conclusions (Boehm, Hawkins, Brown, van Rijn, & Wagenmakers, 2016).

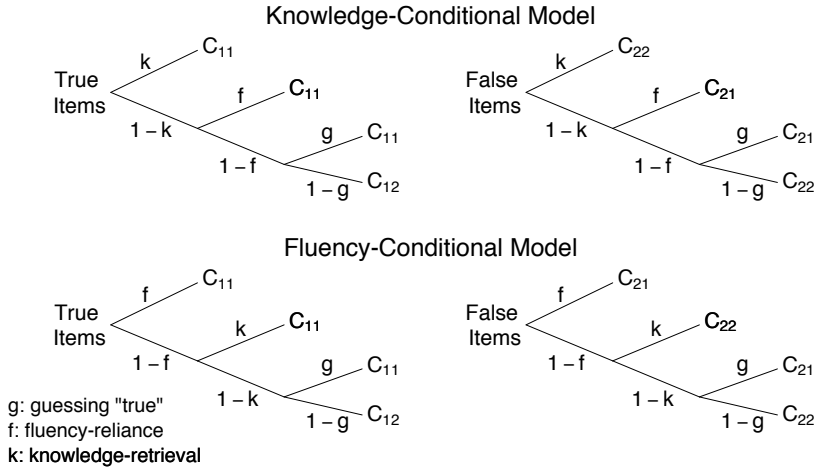


Figure 3.5: The knowledge-conditional (top panel) and fluency-conditional (bottom panel) MPTs. Available at <https://tinyurl.com/ya8sovfr> under CC license <https://creativecommons.org/licenses/by/2.0/>

We re-analyzed data from 39 participants who indicated the truthfulness (i.e., “true”/“false”) of 176 statements, half of which were true and half of which were false. Half of the statements were likely to be known according to general knowledge norms (“known” statements) and half of them were likely to be unknown (“unknown” statements). An example of a true known statement is “The Pacific Ocean is the largest ocean on Earth”. An example of a false unknown statement is “Billy the Kid’s last name is Garrett”. To manipulate fluency, half of the statements were presented twice, once in the exposure phase and once in the truth-rating phase, whereas the other half was only presented in the truth-rating phase. Hence, the experiment had a  $2$  (truth status: true vs. false)  $\times$   $2$  (assumed knowledge: known vs. unknown)  $\times$   $2$  (repetition: repeated vs. not repeated) balanced within-subject design, and each cell of the design featured 22 statements.

### 3.4.2.1 Model Specification

Fazio et al. (2015) constructed two MPTs to study the illusory truth effect. The knowledge-conditional model depicted in the top panel of Figure 3.5 assumes that participants rely on knowledge when assessing truthfulness and only rely on fluency when they are unable to retrieve knowledge about the statement. Parameter  $k$  represents the probability of retrieving knowledge about the statement from memory. If knowledge is retrieved, participants are assumed to give the correct response (i.e., “true” for true statements and “false” for false statements). If no knowledge is retrieved with probability  $1 - k$ , participants rely on fluency with probability  $f$  and respond “true”. If participants do not rely on fluency with probability  $1 - f$ , they guess “true” with probability  $g$  and “false” with probability  $1 - g$ . Responses to true statements are scored into the categories  $C_{11}$

(correct “true” response) and  $C_{12}$  (incorrect “false” response). Responses to false statements are scored into the categories  $C_{21}$  (incorrect “true” response) and  $C_{22}$  (correct “false” response). In contrast, the fluency-conditional model depicted in the bottom panel reflects the notion that participants mainly rely on fluency and only use knowledge in the absence of fluency. The models feature the same set of parameters, but they assume a different conditional probability structure.

For each model, we replicated the two subtrees four times (i.e., a total of eight subtrees per model) to accommodate the design of the experiment: the first replicate corresponded to known true and false statements that were not repeated, the second to known true and false statements that were repeated, the third to unknown true and false statements that were not repeated, and the fourth to unknown true and false statements that were repeated. Following Fazio et al. (2015), we used separate knowledge parameters for known ( $k_k$ ) and unknown ( $k_u$ ) statements, and separate fluency parameters for repeated statements ( $f_r$ ) and statements shown only once ( $f_n$ ). The guessing parameter  $g$  was constrained to be equal across the four replicates. We implemented the models within the hierarchical latent-trait approach, using the prior specifications described earlier.

We estimated the posterior distribution of the model parameters using JAGS, ran three MCMC chains with over-dispersed start values, discarded the first 4,000 posterior samples as burn in, and retained only every 50th sample. Results reported below are based on a total of 180,000 posterior samples. The posterior distributions of the group-level mean parameters are displayed in the Supplemental Materials.

### 3.4.2.2 Computing Bayes Factors with Warp-III

For each model, we split the 180,000 posterior samples in two equal parts (first and second half of the iterations per chain) and used the first part for estimating  $\mathbf{R}$  and  $\mathbf{v}$ , and the second part for the iterative updating scheme in Equation 3.15 ( $D_1 = D_2 = 90,000$ ). Using a standard personal computer and four CPU cores, computing the marginal likelihood took approximately three minutes per model.

The resulting marginal likelihoods were used to compute the Bayes factor in favor of the fluency-conditional model over the knowledge-conditional model. To assess the accuracy of the resulting Bayes factor, we repeated this procedure 50 times. Estimates of the Bayes factor ranged from  $1.3 \times 10^{42}$  to  $3.6 \times 10^{43}$  in favor of the fluency-conditional model. Estimates of the Bayes factor based on the alternative prior choice for the elements of  $\boldsymbol{\xi}$  (i.e., uniform priors with upper bound  $\xi_{\max} = 2$  instead of  $\xi_{\max} = 10$ ) ranged from  $1.7 \times 10^{41}$  to  $1.7 \times 10^{43}$  in favor of the fluency-conditional model. In line with the conclusion drawn by Fazio et al. (2015) based on the  $G^2$  statistic, this result provides overwhelming evidence in favor of the fluency-conditional model.<sup>18</sup>

<sup>18</sup>Although the Bayes factor indicates overwhelming evidence in favor of the fluency-conditional model, it should be kept in mind that the Bayes factor quantifies the evidence of two models relative to each other. In practice, researchers should also check that the model that is favored by the Bayes factor provides an adequate fit to the observed data (e.g., Steingrover et al., 2014).

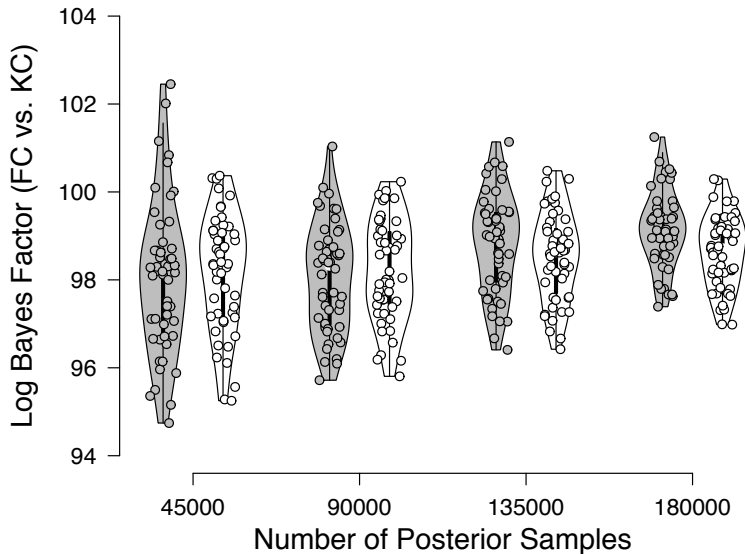


Figure 3.6: Log Bayes factor estimates in favor of the fluency-conditional (FC) model over the knowledge-conditional (KC) model as a function of the number of posterior samples. The Warp-III estimates are displayed in white, the estimates based on the simpler multivariate normal approach are displayed in gray. Available at <https://tinyurl.com/ydbfev7w> under CC license <https://creativecommons.org/licenses/by/2.0/>.

Figure 3.6 displays the Warp-III Bayes factor estimates (on the log scale) in white as a function of the number of posterior samples used in the bridge sampling procedure.<sup>19</sup> As a comparison, the estimates based on the simpler multivariate normal bridge sampling approach are displayed in gray. As the number of posterior samples increases, the Bayes factor estimates become more precise. For this particular example, it is apparent that the Warp-III estimates are less variable than the estimates based on the simpler multivariate normal approach.

### 3.4.2.3 Substantive Contribution

The authors of the original article analyzed the aggregated data (again, an approach known to be suboptimal in case there is heterogeneity across participants)

---

<sup>19</sup>Posterior sample sizes smaller than 180,000 were obtained by considering only a subset of the 180,000 posterior samples for each model (i.e., no new posterior samples were obtained). Note that the same posterior sample sizes were used for the Warp-III and the simpler multivariate normal approach, but the results of the two methods are displayed with an offset to avoid overlapping symbols. Plots for each model's marginal likelihood estimates are presented in the Supplemental Materials.

and considered the  $G^2$  statistics with corresponding  $p$ -values. Based on the fact that the knowledge-conditional model had a larger, significant  $G^2$  statistic compared to the fluency-conditional model that had a lower, non-significant  $G^2$  statistic, the authors concluded that the knowledge-conditional model fit the data poorly and the fluency-conditional model fit the data well. Therefore, the authors favored the fluency-conditional model based on two binary accept-reject decisions. This makes it difficult to gauge the degree of support that the data provide in favor of the fluency-conditional model. The Bayes factor may be 10, or 100, or 1,000 – these are very different levels of evidence. In fact, our analysis shows that the Bayes factor is about  $1.3 \times 10^{42}$  to  $3.6 \times 10^{43}$  in favor of the fluency-conditional model, which represents an overwhelming amount of evidence.

It could be argued that, since the compared models have the same number of parameters, comparing  $G^2$  statistics may result in choosing the same model as based on considering AIC or BIC. AIC is asymptotically equivalent to cross-validation (M. Stone, 1977) which is known to be inconsistent in the sense that, when the number of observations goes to infinity, the data-generating model will not be chosen with certainty (Shao, 1993). In contrast, when using Bayes factors, model-selection consistency is generally fulfilled (Bayarri, Berger, Forte, & García-Donato, 2012). Although the BIC is a rough approximation of the Bayes factor, we believe that it is better to compute proper Bayes factors which are transparent with respect to the prior assumptions.

Finally, one might argue again that this data set is relatively small and is thus uninformative. However, the resulting Bayes factor is very different from 1, indicating that the data are in fact highly informative with respect to adjudicating between the fluency-conditional and the knowledge-conditional model.

### 3.5 Discussion

Bayesian hierarchical techniques for MPT modeling are increasingly popular. Current hierarchical MPT approaches, however, do not incorporate Bayesian model comparison methods based on Bayes factors and posterior model probabilities, possibly because of the computational challenges associated with the evaluation of the marginal likelihood. In this chapter, we addressed this challenge and showed how Warp-III bridge sampling can be used to obtain accurate and stable estimates of the marginal likelihood of hierarchical MPTs. We applied the method to model comparison problems from two published studies and illustrated how the marginal likelihood can be used for Bayesian model averaging and for the computation of the Bayes factor.

Our examples highlighted that Bayesian model comparison based on posterior model/inclusion probabilities and Bayes factors allows researchers to disentangle between “absence of evidence” and “evidence of absence”. Note that it is crucial in all stages of cognitive model development, validation, and application that one is able to quantify evidence in favor of invariances (i.e., “evidence of absence”) in a coherent and systematic way. For model development and validation, it is important to show that certain experimental manipulations selectively influence only a subset of the model parameters whereas the remaining parameters are unaffected

(i.e., selective influence studies). Once a cognitive model has been established as a valid measurement tool, it can be used, for instance, to investigate which subprocesses are targeted by new experimental manipulations or which subprocesses differ or do not differ in clinical subpopulations (cognitive psychometrics; e.g., Riefer et al., 2002). In these applications it is important to be able to quantify evidence for a difference but, crucially, also for an invariance since one might wish to make statements of the form “there is evidence that retrieval processes are not affected”.

There are often a number of different candidate models for the analysis of observed data. In Example 1, we demonstrated how Bayesian model averaging can be used to draw conclusions that fully take into model uncertainty. In our opinion, Bayesian model averaging is an extremely powerful approach and, to the best of our knowledge, it is currently not used in the context of hierarchical MPTs and cognitive modeling more generally. We believe that attending researchers to this approach and providing the computational tools to facilitate its application (i.e., Warp-III) is one of the key contributions of this work.

Our examples illustrated that Warp-III is relatively straightforward to implement once posterior samples from the models have been obtained with MCMC sampling. Another advantage of Warp-III bridge sampling is its relative speed. In our experience, the Warp-III procedure requires much less computational time than the MCMC sampling from the posterior. One of the crucial determinants of the computational time of Warp-III is how long it takes to evaluate the un-normalized posterior density. To maximize speed for our applications, we implemented the un-normalized posterior density functions in C++ code called from within R via Rcpp (Eddelbuettel et al., 2011). Compared to a simpler bridge sampling version which only matches the first two moments of the proposal and the posterior (e.g., Overstall & Forster, 2010), Warp-III is expected to take about twice as long for a fixed number of samples due to the mixture representation of the warping procedure which requires evaluating the un-normalized posterior twice as often as for the simpler bridge sampling version. However, Warp-III is also expected to be more accurate in case the posterior is skewed which means there might be a speed-accuracy trade-off.

Despite its computational simplicity, Warp-III should not be applied blindly. Specifically, as we demonstrated for our empirical examples, it is important to assess the variability of the resulting model comparison measure – such as posterior model probabilities or Bayes factors – by repeating the Warp-III procedure multiple times. When the measure of interest clearly favors a given model, as in our second example, some fluctuation is not necessarily concerning. However, in situations where the fluctuation influences which model is favored, researchers should either increase the number of posterior and proposal samples to decrease the variability of the estimate, or, if this solution is practically infeasible, they should acknowledge that the estimate does not support firm conclusions about the relative predictive adequacy of the models.

The accuracy of the estimate is governed not only by the number of samples but also by the overlap between the proposal and the posterior distribution. Warp-III attempts to maximize this overlap by matching the mean vector, covariance matrix, and the skew of the two distributions. However, in case the posterior

distribution exhibits multiple modes, the overlap may not be sufficiently close. Researchers should carefully check whether multi-modalities occur in their application. If this is the case, repeated runs of the Warp-III procedure could be used to obtain an impression of the stability of the estimate. Nevertheless, it should be kept in mind that Warp-III is not designed for multi-modal posterior distributions and results should be interpreted with caution. The development of bridge sampling procedures for multi-modal posterior distributions is currently ongoing (e.g., Frühwirth-Schnatter, 2004; L. Wang & Meng, 2016). Note, however, that this is not a very severe limitation of the Warp-III method, since posterior distributions are unimodal in many models used in psychology – they even converge to normal distributions under specific conditions (Dawid, 1970).

Relatedly, note that we use the unscaled effects  $\omega_i$  and the scaling parameters  $\xi$  directly in the bridge sampling procedure – but technically, these are only identified jointly. Therefore, MCMC chains for these parameters may look irregular and exhibit, for instance, multiple modes, decreasing the efficiency of the Warp-III procedure as mentioned above. Although this was not the case for our applications, we advise researchers to carefully monitor the MCMC chains of the unidentified unscaled effects and scaling parameters.

On a more theoretical note, as Equation 3.3 illustrates, Bayesian model comparison is sensitive to the choice of the prior distribution. We relied on relatively standard priors for the group-level parameters, but also established the robustness of our conclusions with a series of sensitivity analyses (see also Supplemental Materials). Nevertheless, we do not suggest that our prior choices should be considered as the gold-standard for model comparison in hierarchical MPTs. Several approaches are available for specifying theoretically justified prior distributions for cognitive models (Lee & Vanpaemel, 2018; see also Heck & Wagenmakers, 2016, for specifying order constraints in MPTs). We believe that the increasing popularity of hierarchical MPTs will enable researchers to specify informative paradigm-specific and model-specific prior distributions based on experience with the models (e.g., typical parameter ranges and effect sizes). The dependency on the prior is sometimes considered as a weakness of Bayes factor model comparisons (e.g., Aitkin, 2001). Some researchers and statisticians even conclude that due to this reason, the use of Bayes factors is not recommended (e.g., Gelman, Carlin, et al., 2014, chapter 7.4).<sup>20</sup> In contrast, we believe that the ability to incorporate prior knowledge is an advantage of Bayesian inference; we consider the prior as integral part of the model which should be chosen just as carefully as the likelihood (e.g., Vanpaemel, 2010). Ideally, researchers should pre-register their priors before data collection (Chambers, 2013, 2015) to ensure that these are used to express genuine prior knowledge and not to increase researchers’ degrees of freedom in obtaining the desired results. Note that we are not the first to advocate a Bayesian approach to hierarchical MPTs. However, to the best of our knowledge, we are the first who advocate Bayesian model comparison using posterior model/inclusion probabilities and Bayes factors and provide the tools

<sup>20</sup>Another objection is that Bayes factors are often used to compare nested models where certain values of continuous parameters are treated as “special” (since the parameters are fixed to these values). These researchers often favor continuous model expansion instead (e.g., Gelman, Carlin, et al., 2014, chapter 7.4; Gelman & Rubin, 1995).

to compute these quantities for hierarchical MPTs. Equipped with a feasible approach for computing the relevant quantities for Bayesian model comparison, one could, in principle, specify an informed prior for the models themselves in addition to the specification of the parameter prior. This way one could incorporate prior knowledge about how likely each model is or one could, if desired, incorporate a penalty for multiple comparisons as described in Scott and Berger (2010).

Although we focused exclusively on latent-trait MPTs, Warp-III is not limited to the latent-trait approach or other hierarchical MPTs, such as the beta-MPT (J. B. Smith & Batchelder, 2010) or the crossed-random effects approach (Matzke et al., 2015). Warp-III may be used to compute the marginal likelihood for a large variety of cognitive models. For instance, the simple multivariate normal bridge sampling approach has been recently applied to hierarchical reinforcement learning models (Gronau, Sarafoglou, et al., 2017). We believe that Warp-III may be especially useful for so-called sloppy models with highly correlated parameters (K. S. Brown & Sethna, 2003), including but not limited to race models of response times, which often yield skewed posterior distributions (e.g., S. D. Brown & Heathcote, 2008; Matzke, Love, & Heathcote, 2017). The Warp-III methodology also lends itself to model comparison in extensions of hierarchical cognitive models that impose on the model parameters a statistical structure such as a linear regression, factor analysis, or analysis of variance (e.g., Boehm, Steingroever, & Wagenmakers, 2018; Heck, Arnold, & Arnold, 2018; Turner, Wang, & Merkle, 2017; Vandekerckhove, 2014). The application of Warp-III to complex experimental designs is ongoing work in our lab.

Although Warp-III is a general procedure for computing the marginal likelihood, depending on the situation, other approaches may be better suited for the model comparison problem at hand. If researchers focus on non-hierarchical implementations of cognitive models, importance sampling may be an easier solution, particularly in the context of MPTs (Vandekerckhove et al., 2015). If the focus is on nested models, the Savage-Dickey density ratio is an easier and faster alternative. Lastly, if the number of models under consideration is very large, Reversible Jump MCMC (Green, 1995) might be the appropriate choice. Nevertheless, we believe that in most applications of hierarchical cognitive models, the research question concerns the comparison of a limited set of possibly non-nested models. In these situations, Warp-III provides a straightforward and accurate method for computing the marginal likelihood for a wide range of complex models.

The Supplemental Materials can be found at: <https://osf.io/rycg6/>.

---

# Computing Bayes Factors for Evidence-Accumulation Models Using Warp-III Bridge Sampling

---

## Abstract

Over the last decade, the Bayesian estimation of evidence-accumulation models has gained popularity, largely due to the advantages afforded by the Bayesian hierarchical framework. Despite recent advances in the Bayesian estimation of evidence-accumulation models, model comparison continues to rely on suboptimal procedures, such as posterior parameter inference and model selection criteria known to favor overly complex models. In this chapter we advocate model comparison for evidence-accumulation models based on the Bayes factor obtained via Warp-III bridge sampling. We demonstrate, using the Linear Ballistic Accumulator (LBA), that Warp-III sampling provides a powerful and flexible approach that can be applied to both nested and non-nested model comparisons, even in complex and high-dimensional hierarchical instantiations of the LBA. We provide an easy-to-use software implementation of the Warp-III sampler and outline a series of recommendations aimed at facilitating the use of Warp-III sampling in practical applications.

## 4.1 Introduction

Cognitive models of response times and accuracy canonically assume an accumulation process, where evidence favoring different options is summed over time until a

---

This chapter is published as Gronau, Q. F., Heathcote, A., & Matzke, D. (2020). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods*, 52, 918–937. doi: <https://doi.org/10.3758/s13428-019-01290-6>. Also available as *PsyArXiv preprint*: <https://psyarxiv.com/9g4et>

threshold is reached that triggers an associated response. The two most prominent types of evidence-accumulation models, the Diffusion Decision Model (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008) and the Linear Ballistic Accumulator (LBA; S. D. Brown & Heathcote, 2008) have been widely applied across animal and human research in biology, psychology, economics, and the neurosciences to topics including vision, attention, language, memory, cognition, emotion, development, aging, and clinical disorders (for reviews, see Donkin & Brown, 2018; M. J. Mulder, Van Maanen, & Forstmann, 2014; Ratcliff, Smith, Brown, & McKoon, 2016). Evidence-accumulation models are popular because they provide a comprehensive account of the probability of choices and the associated distribution of times to make them, and because they provide parameter estimates that directly quantify important psychological quantities, such as the quality of the evidence provided by a choice stimulus and the amount of evidence required to trigger the response.

Parameter estimation and statistical inference in the context of evidence-accumulation models can be challenging because they belong to the class of “sloppy” models with highly correlated parameters (Apgar, Witmer, White, & Tidor, 2010; Gutenkunst et al., 2007), examples of which occur widely in biology and psychology (Apgar et al., 2010; Gutenkunst et al., 2007; Heathcote et al., 2018). However, with appropriate experimental designs – critically including sufficiently high error rates and experimental trials per participant (Ratcliff & Childers, 2015) – the model parameters can be estimated reliably using error minimization and Bayesian methods.

Recently, the Bayesian estimation of evidence-accumulation models has gained popularity, largely due to the advantages afforded by the Bayesian hierarchical framework (e.g., Heathcote et al., 2018; Vandekerckhove, Tuerlinckx, & Lee, 2011; Wiecki, Sofer, & Frank, 2013). In fact, our recent literature review indicated that 19% and 21% of the 262 and 53 papers that used the DDM and the LBA, respectively, relied on Bayesian methods to estimate the model parameters.<sup>1</sup> Bayesian hierarchical methods simultaneously estimate model parameters for a group of participants assuming that the participant-level parameters are drawn from a common group-level distribution. From a statistical point of view, the group-level distribution acts as a prior that pulls (“shrinks”) the participant-level parameters to the group mean, which can result in less variable and, on average, more accurate estimates than non-hierarchical methods (Farrell & Ludwig, 2008; Gelman & Hill, 2007; Lee & Wagenmakers, 2013; Shiffrin et al., 2008). From a psychological point of view, the group-level distribution provides a model of individual differences. From this perspective, it is apparent that introducing a group-level distribution improves the model theoretically only if the group-level distribution provides a good model for the individual variation (Farrell & Lewandowsky, 2018, section 9.5).

As a result of the strong parameter correlations in evidence-accumulation models, standard Markov chain Monte Carlo samplers (MCMC; e.g., Gilks, Richardson, & Spiegelhalter, 1996) typically used for Bayesian parameter estimation can

---

<sup>1</sup>The numbers are based on a systematic literature review of published articles that fit the DDM and LBA to empirical data (Tran, 2018). A summary of the results is available at <https://osf.io/ynwpa/>.

be inefficient. Rather, samplers designed to handle high posterior correlations must be used, such as differential evolution MCMC (DE-MCMC; Turner, Sederberg, Brown, & Steyvers, 2013). This approach to Bayesian estimation is now readily available for the DDM, LBA, and other evidence-accumulation models in the “Dynamic Models of Choice” software (DMC; Heathcote et al., 2018) along with extensive tutorials and supporting functions that facilitate model diagnostics and the analysis of results.<sup>2</sup> In this chapter, we focus on the Bayesian approach because of the advantages it offers, such as a coherent inferential framework, the use of prior information, the possibility of straightforward hierarchical extensions, and the natural quantification of uncertainty in both parameter estimates and model predictions.

In typical applications of evidence-accumulation models, researchers are not only interested in parameter estimation, but often wish to assess the effects of experimental manipulations on the model parameters. For example, Strickland, Loft, Remington, and Heathcote (2018) compared non-nested LBA models that either allowed the effect of maintaining a prospective memory load (i.e., in the context of a routine ongoing task, the intent to make an alternative response to a rarely occurring stimulus) to influence only the rate of evidence accumulation or only the threshold amount of evidence required to make a response. The former model corresponds to competition for limited information-processing capacity, whereas the latter model corresponds to strategic slowing in order to avoid the ongoing task response pre-empting the prospective memory response (Heathcote, Loft, & Remington, 2015). Nested comparisons are also common in the context of evidence-accumulation models to determine which of a set of candidate experimental manipulations had an effect on a particular parameter. For example, Rae, Heathcote, Donkin, Averell, and Brown (2014) examined whether or not an emphasis on the speed vs. accuracy of responding influences evidence accumulation rates.

Despite recent advances in the Bayesian estimation of evidence-accumulation models, model comparison continues to rely on suboptimal procedures, such as posterior parameter inference based on complex models where separate model parameters are estimated for each experimental condition. In this approach, differences between parameters are often evaluated using posterior  $p$ -values (e.g., Klauer, 2010; Matzke, Boehm, & Vandekerckhove, 2018; Matzke et al., 2015; Matzke, Hughes, Badcock, Michie, & Heathcote, 2017; Osth, Jansson, Dennis, & Heathcote, 2018; J. B. Smith & Batchelder, 2010; Strickland et al., 2018; Tilman, Osth, van Ravenzwaaij, & Heathcote, 2017; Tilman, Strayer, Eidels, & Heathcote, 2017). Posterior parameter inference has at least three limitations. First, it can only be used for nested model comparison. Second, it cannot provide evidence for the absence of an effect (i.e., it cannot “prove the null”), similar to classical  $p$ -values (e.g., Wagenmakers, 2007). Third, it can result in fitting an overly complex model, which is particularly problematic in the presence of strong parameter correlations, because a real effect in one parameter can spread to create a spurious effect on other parameters (Heathcote et al., 2015).

<sup>2</sup>A file that describes the content of the DMC tutorials and the different DMC functions is available from <https://osf.io/kygr3/>.

These shortcomings can be addressed using formal model selection. This approach critically depends on the availability of a model selection criterion that properly penalizes the greater flexibility of more complex models. The Deviance Information Criterion (DIC) is one of the most commonly used model selection measures, and has the advantage that it can be easily computed from the posterior samples obtained during parameter estimation. However, the DIC is known to prefer overly complex models (Spiegelhalter et al., 2002). The more recent Widely Applicable Information Criterion (WAIC; Vehtari et al., 2017), which is also based on posterior samples, is an approximation to (leave-one-out) cross-validation and suffers from the same shortcoming (Browne, 2000). It should be noted that even as the number of observations goes to infinity, methods that approximate (leave-one-out) cross-validation will not choose the data-generating model with certainty (Shao, 1993).

Here we advocate model selection for evidence-accumulation models based on the *Bayes factor* (e.g., Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Ly et al., 2016a). The Bayes factor is the principled method of performing model selection from a Bayesian perspective and follows immediately from applying Bayes’ rule to models instead of parameters (e.g., Kass & Raftery, 1995). In contrast to model selection methods that approximate (leave-one-out) cross-validation, in general, the Bayes factor will choose the data-generating model with certainty when the number of observations goes to infinity (Bayarri et al., 2012). Although the desirability of Bayes factors has long been recognized (e.g., Jeffreys, 1939), their use has only become increasingly widespread with general linear models (e.g., ANOVA and regression; see Rouder, Morey, Speckman, & Province, 2012 and Rouder & Morey, 2012) due the availability of efficient and user-friendly software implementations in packages such as **BayesFactor** (Morey & Rouder, 2015) in R (R Core Team, 2019) and the GUI-based **JASP** (JASP Team, 2020). With this chapter, we aim to bring these advantages to the domain of evidence-accumulation models by providing an easy-to-use software implementation that uses a state-of-the-art method for computing Bayes factors.

The Bayes factor is the predictive updating factor that changes prior model odds for two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  into posterior model odds based on observed data  $\mathbf{y}$ :

$$\underbrace{\frac{p(\mathcal{M}_1 | \mathbf{y})}{p(\mathcal{M}_2 | \mathbf{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathbf{y} | \mathcal{M}_1)}{p(\mathbf{y} | \mathcal{M}_2)}}_{\text{Bayes factor BF}_{12}} \times \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{prior odds}}. \quad (4.1)$$

Continuing the example from Strickland et al. (2018), suppose that  $\mathcal{M}_1$  refers to the model in which only rates are affected by prospective-memory load and  $\mathcal{M}_2$  refers to the model in which only thresholds are affected. Different researchers may start with different prior beliefs about the relative plausibility of the two competing psychological explanations of the prospective-memory load effect. However, the change in beliefs brought about by the data (i.e., the change from prior to posterior odds which is the Bayes factor) is the same, regardless of the prior beliefs. Therefore, reporting the Bayes factor enables researchers to update their personal prior odds to posterior odds. Commonly, only the Bayes factor is reported and interpreted, since strength of evidence for the two competing models is naturally

expressed as the degree to which one should update prior beliefs about the models based on observed data. A Bayes factor of, say,  $\text{BF}_{12} = 10$  would indicate that the data are 10 times more likely under  $\mathcal{M}_1$  than  $\mathcal{M}_2$ , whereas a Bayes factor of  $\text{BF}_{12} = 0.1$  would indicate that the data are 10 times more likely under  $\mathcal{M}_2$  than  $\mathcal{M}_1$ .

As shown in Equation 4.1, the Bayes factor is the ratio of the *marginal likelihoods* of the models. The marginal likelihood is the probability of the data given a model and is obtained by integrating out the model parameters with respect to the parameters' prior distribution:

$$p(\mathbf{y} \mid \mathcal{M}) = \int_{\Theta} p(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M}) d\boldsymbol{\theta}, \quad (4.2)$$

where  $\boldsymbol{\theta}$  denotes the parameter vector for model  $\mathcal{M}$ . The marginal likelihood quantifies average predictive adequacy as follows: The likelihood  $p(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{M})$  corresponds to the predictive adequacy of a particular parameter setting  $\boldsymbol{\theta}$  under model  $\mathcal{M}$ . The average predictive adequacy (i.e., the marginal likelihood) is obtained as the weighted average of the predictive adequacies across the entire parameter space, where the weights are given by the parameters' prior probabilities. Complex models may have certain parameter settings that yield high likelihood values, however, the large parameter space may also contain many parameter settings which result in small likelihood values, lowering the weighted average. Consequently, the marginal likelihood – and the Bayes factor, which contrasts the average predictive adequacy of two models – incorporates a natural penalty for undue complexity. Interpreting the marginal likelihood as a weighted average highlights the crucial importance of the prior distribution for Bayesian model comparison.

For evidence-accumulation models, the integral in Equation 4.2 – and hence the Bayes factor – cannot be computed analytically. In these cases, four major approaches are available for computing Bayes factors: (1) approximate methods such as the Laplace approximation (e.g., Kass & Vaidyanathan, 1992); (2) the Savage-Dickey density ratio approximation of the Bayes factor (Dickey & Lientz, 1970; Wagenmakers et al., 2010); (3) transdimensional methods such as reversible jump MCMC (Green, 1995); and (4) simulation-based methods that estimate the integrals involved in the computation of the Bayes factor directly (e.g., Evans & Annis, 2019; Evans & Brown, 2018; Meng & Schilling, 2002; Meng & Wong, 1996). Approximate methods have the disadvantage that it is typically difficult to assess the approximation error, which could be particularly substantial for hierarchical evidence-accumulation models. The Savage-Dickey density ratio can only be applied to nested model comparisons. Transdimensional methods are challenging to implement, especially in hierarchical settings and for non-nested model comparisons, as explained in more detail later.

Therefore, here we advocate *Warp-III bridge sampling* (Meng & Schilling, 2002) for obtaining the Bayes factor for evidence-accumulation models. Warp-III bridge sampling is a simulation-based method that can be applied to both nested and non-nested comparisons and – once posterior samples from the competing models have been obtained – it is straightforward to implement even in hierarchical settings. As non-nested hierarchical comparisons are integral to many

applications of cognitive models, we believe that Warp-III bridge sampling provides an excellent computational tool that will greatly facilitate the use of Bayesian model comparison for evidence-accumulation models.

The chapter is organized as follows. First, we review simple Monte Carlo sampling, another simulation-based method that has been proposed for computing the Bayes factor for evidence-accumulation models. We then outline the details of Warp-III bridge sampling and illustrate its use for the single-participant as well as the hierarchical case. We focus on the LBA, but elaborate on the applicability of our approach to other evidence-accumulation models, for instance the DDM, in the Discussion. The Discussion also provides recommendations aimed at facilitating the use of Warp-III bridge sampling in practical applications. The implementation of the Warp-III bridge sampler is available at <https://osf.io/ynwpa/> and has also been incorporated into the latest DMC release.<sup>3</sup>

## 4.2 Simple Monte Carlo Sampling

A simple Monte Carlo estimator of the marginal likelihood is obtained by interpreting the integral in Equation 4.2 as an expected value with respect to the parameters' prior distribution:

$$\begin{aligned} p(\mathbf{y} \mid \mathcal{M}) &= \mathbb{E}_{p(\boldsymbol{\theta} \mid \mathcal{M})} [p(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{M})] \\ &\approx \frac{1}{N} \sum_{i=1}^N p(\mathbf{y} \mid \tilde{\boldsymbol{\theta}}_i, \mathcal{M}), \quad \text{where } \tilde{\boldsymbol{\theta}}_i \sim p(\boldsymbol{\theta} \mid \mathcal{M}). \end{aligned} \quad (4.3)$$

Thus, an estimate of the marginal likelihood can be obtained by sampling from the prior distribution and averaging the likelihood values based on the samples.

Recently, Evans and Brown (2018) proposed the use of simple Monte Carlo sampling for the computation of the Bayes factor for the LBA. This simple approach can work well if the posterior distribution is similar to the prior distribution; however, when the posterior is substantially different from the prior – as is often the case – simple Monte Carlo sampling becomes very inefficient. The reason is that only a few prior samples (i.e., those in the region where most posterior mass is located) result in substantial likelihood values so that the average in Equation 4.3 will be dominated by a small number of samples. The result is an unstable estimator, even in non-hierarchical applications. Naturally, the problem becomes more severe in hierarchical settings where the parameter space is substantially larger. Although increasing the number of prior samples may remedy the problem to a certain extent, reliable estimation of the marginal likelihood of hierarchical evidence-accumulation models using simple Monte Carlo sampling remains challenging, even with Evans and Brown's powerful GPU implementation. Given the many advantages of the Bayesian hierarchical framework for cognitive modeling (e.g., Heathcote et al., 2018; Lee, 2011; Lee & Wagenmakers, 2013; Matzke et al., 2015; Matzke, Dolan, Logan, Brown, & Wagenmakers, 2013; Shiffrin et al., 2008;

---

<sup>3</sup>This release is available at <https://osf.io/5yeh4/>. It also contains a new tutorial that explicitly explains how to use the bridge sampling functionality in DMC (i.e., `dmc_5-7_BayesFactors.R`).

Vandekerckhove et al., 2011; Wiecki et al., 2013), we believe that an alternative approach is needed.

### 4.3 Warp-III Bridge Sampling

We propose the use of Warp-III bridge sampling (Meng & Schilling, 2002, henceforth referred to as *Warp-III*) for estimating the marginal likelihood for evidence-accumulation models. Warp-III is an advanced version of bridge sampling (Gronau, Sarafoglou, et al., 2017; Meng & Wong, 1996), which is based on the following identity:

$$p(\mathbf{y} \mid \mathcal{M}) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})} [h(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M})]}{\mathbb{E}_{p(\boldsymbol{\theta} \mid \mathbf{y}, \mathcal{M})} [h(\boldsymbol{\theta}) g(\boldsymbol{\theta})]}, \quad (4.4)$$

where  $g$  is a proposal distribution and  $h$  a bridge function.

The efficiency of the bridge sampling estimator is governed by the overlap between the proposal and the posterior distribution. A simple approach for obtaining the bridge sampling estimator relies on a multivariate normal proposal distribution that matches the first two moments, the mean vector and covariance matrix, of the posterior distribution (e.g., Gronau, Sarafoglou, et al., 2017; Overstall & Forster, 2010). However, this method becomes inefficient when the posterior distribution is skewed. To remedy this problem, Warp-III aims to maximize the overlap by fixing the proposal distribution to a standard multivariate normal distribution<sup>4</sup> and then “warping” (i.e., manipulating) the posterior so that it matches not only the first two, but also the third moment of the proposal distribution (for details, see Meng & Schilling, 2002, and Gronau, Wagenmakers, Heck, & Matzke, 2019).

Figure 4.1 illustrates the warping procedure for the univariate case using hypothetical posterior samples. The solid black line in the top-left panel displays the standard normal proposal distribution and the skewed histogram displays samples from the posterior distribution. Since none of the moments of the two distributions match, applying bridge sampling to these distributions can be called Warp-0 (i.e., the number indicates how many moments have been matched). The histogram in the top-right panel displays the same posterior samples after subtracting their mean from each sample. This manipulation matches the first moment of the two distributions; the posterior samples are now zero-centered, just like the proposal distribution. This is called Warp-I. In the bottom-right panel, the posterior samples are additionally divided by their standard deviation. This manipulation matches the first two moments of the distributions; the posterior samples are now zero-centered with variance 1, just like the proposal distribution. This is called Warp-II. Finally, the bottom-left panel displays the posterior samples after assigning a minus sign with probability 0.5 to each sample. This manipulation achieves symmetry and matches the first three moments of the distributions; the posterior samples are now symmetric and zero-centered with variance 1, just like the proposal distribution. This is called Warp-III. Note how successively matching the moments of the two distributions has increased the overlap between the posterior

<sup>4</sup>Other proposal distributions, such as a multivariate  $t$ -distribution, are also conceivable.

#### 4. COMPUTING BAYES FACTORS FOR EVIDENCE-ACCUMULATION MODELS USING WARP-III BRIDGE SAMPLING

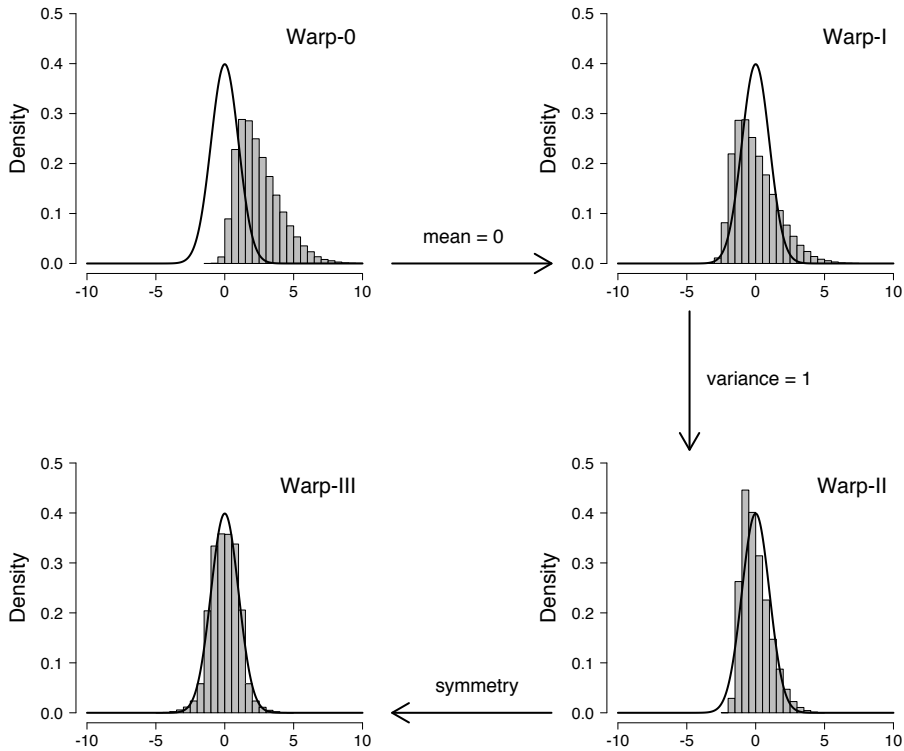


Figure 4.1: Illustration of the warping procedure. The black solid line shows the standard normal proposal distribution and the gray histogram shows the posterior samples. Available at <https://tinyurl.com/y7owvsz3> under CC license <https://creativecommons.org/licenses/by/2.0/>.

and the proposal distribution.<sup>5</sup> We have found that the improvement afforded by Warp-III can be crucial for efficient application of bridge sampling to evidence-accumulation models, particularly in situations where the posteriors are skewed, as is often the case with only a small number of observations per participant.

The bridge function  $h$  is chosen such that it minimizes the relative mean-square error of the resulting estimator (Meng & Wong, 1996). Using this “optimal” bridge function,<sup>6</sup> the estimator of the marginal likelihood is obtained by updating

<sup>5</sup>The warping procedure assumes that all parameters are allowed to range across the entire real line; if this is not the case, appropriate transformations can be applied to fulfill this requirement. Note that the resulting expressions need to be adjusted by the relevant Jacobian term.

<sup>6</sup>Note that this choice is only optimal if the samples from the posterior distribution are independent which is not the case when using MCMC methods. To account for this fact, we replace  $N_1$  when computing  $s_1$  and  $s_2$  by an effective sample size – the median effective sample

an initial guess of the marginal likelihood until convergence. The estimate at iteration  $t + 1$  is given by:<sup>7</sup>

$$\hat{p}(\mathbf{y} \mid \mathcal{M})^{(t+1)} = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}(\mathbf{y} \mid \mathcal{M})^{(t)}}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}(\mathbf{y} \mid \mathcal{M})^{(t)}}}, \quad (4.5)$$

where  $s_k = \frac{N_k}{N_1 + N_2}$  for  $k \in \{1, 2\}$ ,

$$l_{1,j} = \frac{\frac{|\mathbf{R}|}{2} [q(2\boldsymbol{\mu} - \boldsymbol{\theta}_j^*) + q(\boldsymbol{\theta}_j^*)]}{g(\mathbf{R}^{-1}(\boldsymbol{\theta}_j^* - \boldsymbol{\mu}))}, \quad (4.6)$$

and

$$l_{2,i} = \frac{\frac{|\mathbf{R}|}{2} [q(\boldsymbol{\mu} - \mathbf{R}\tilde{\boldsymbol{\theta}}_i) + q(\boldsymbol{\mu} + \mathbf{R}\tilde{\boldsymbol{\theta}}_i)]}{g(\tilde{\boldsymbol{\theta}}_i)}. \quad (4.7)$$

$\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \dots, \boldsymbol{\theta}_{N_1}^*\}$  are  $N_1$  draws from the posterior distribution,  $\{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_{N_2}\}$  are  $N_2$  draws from the standard normal proposal distribution, and  $q(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M})$  denotes the un-normalized posterior density function. Furthermore,  $\boldsymbol{\mu}$  corresponds to the posterior mean vector and  $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{R}^\top$  corresponds to the posterior covariance matrix ( $\mathbf{R}$  is obtained via a Cholesky decomposition of the posterior covariance matrix). The posterior mean vector and covariance matrix can be estimated using the posterior samples. In practice, we split the posterior samples in two halves; the first half is used to estimate  $\boldsymbol{\mu}$  and  $\mathbf{R}$  and the second half is used in the iterative scheme in Equation 4.5.

Computing  $l_{1,j}$  and  $l_{2,i}$  is the computationally most expensive part of the method; fortunately, these quantities can be computed completely in parallel. Note also that  $l_{1,j}$  and  $l_{2,i}$  only need to be computed once before the updating scheme is started. Hence, with these quantities in hand, running the updating scheme is quick and typically converges in fewer than 20 or 30 iterations. Although our implementation relies on a fixed starting value, it is also possible to start the updating scheme from an informed guess of the marginal likelihood, for instance, based on a normal approximation to the posterior distribution. We have found that the value of the initial guess usually does not influence the resulting estimator substantially, but a good starting value may reduce the number of iterations needed to reach convergence. Moreover, as we show later, an appropriately chosen starting value is crucial in rare cases when the iterative scheme seemingly does not converge.<sup>8</sup>

It can be shown that the simple Monte Carlo estimator described in the previous section is a special case of Equation 4.4 obtained by using a bridge function other than the optimal one (e.g., Gronau, Sarafoglou, et al., 2017, Appendix A).

size across all posterior components – obtained using the `coda` R package (Plummer et al., 2006).

<sup>7</sup>Note that in practice, we always run the iterative scheme in a more numerically stable way with respect to  $\hat{r}^{(t)} = \text{const} \times \hat{p}(\mathbf{y} \mid \mathcal{M})^{(t)}$  (for details, see Gronau, Sarafoglou, et al., 2017, Appendix B).

<sup>8</sup>In principle, convergence is guaranteed (Meng & Wong, 1996), however, convergence may be so slow that it is infeasible to wait in practice.

Therefore, Warp-III that relies on the optimal bridge function must perform better in terms of the relative mean-square error of the estimator than the simple Monte Carlo approach. This will be illustrated in the next section, where we apply Warp-III sampling to a nested model comparison problem and compare its performance to three alternative methods, including simple Monte Carlo sampling.

## 4.4 Simulation Study I: Nested Model Comparison for the Single-Participant Case

As a first example, we computed the Bayes factor for a nested model comparison problem in the LBA by approximating the marginal likelihood of the two models using Warp-III sampling. To verify the correctness of our Warp-III implementation, we also computed the Bayes factor using three alternative methods: (1) simple Monte Carlo sampling; (2) the Savage-Dickey density ratio; and (3) a simple version of reversible jump MCMC (RJMCMC; Green, 1995) as described in Barker and Link (2013). We included the latter two approaches because they provide conceptually different methods for Bayes factor computations than the simulation-based Warp-III and simple Monte Carlo. The details of the Savage-Dickey and the RJMCMC methods are provided in the Appendix.

### 4.4.1 Models and Data

We considered a data set generated from the LBA for a single participant performing a simple choice task with two stimuli and two corresponding responses. As shown in Figure 4.2, the LBA assumes a race among a set of deterministic evidence-accumulation processes, with one runner per response option. The choice is determined by the winner of the race.

On each trial, accumulation begins at a starting point drawn – independently for each accumulator – from a uniform distribution with width  $A$ .  $A$  may vary between accumulators, but here we assume it is the same. The evidence total increases linearly at rate  $v$  that is drawn independently for each accumulator from a normal distribution, which we assume here is truncated below at zero (Heathcote & Love, 2012). The accumulator that matches the stimulus has mean rate  $v_{\text{true}}$  and standard deviation  $s_{\text{true}}$ , and the mismatching accumulator  $v_{\text{false}}$  and  $s_{\text{false}}$ . In principle, there could be different  $v_{\text{true}}$  and  $v_{\text{false}}$  values for each stimulus, but here we assume they are the same. The first accumulator to reach its threshold ( $b$ ) – again potentially differing between accumulators but assumed to be the same here – triggers the corresponding response. We estimate threshold in terms of a positive quantity,  $B$ , which quantifies the gap between the threshold and the upper bound of the start-point noise (i.e.,  $B = b - A$ ). Response time (RT) is equal to the time taken to reach threshold plus non-decision time,  $t_0$ , which is the sum of the time to initially encode the stimulus and the time to produce a motor response.

We estimated the Bayes factor to compare two nested LBA models. The first, which we refer to as the *full* model, featured a starting point parameter  $A$ , a threshold parameter  $B$ , mean drift rate parameters for the matching and

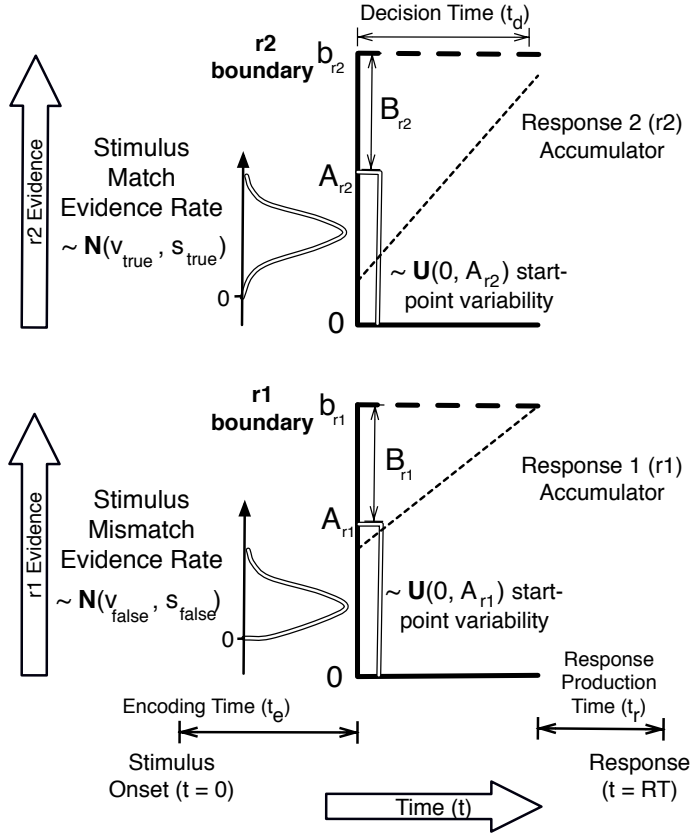


Figure 4.2: *Graphical representation of the Linear Ballistic Accumulator for two possible responses ( $r1$  and  $r2$ ) corresponding to two stimuli ( $s1$  and  $s2$ ).* The figure illustrates a case where  $s2$  is presented and the sampled rate for the  $r2$  accumulator is greater than the sampled rate for the  $r1$  accumulator, i.e., the accumulation path (dashed line) is steeper for  $r2$  than for  $r1$ . However, as the sampled starting point for  $r1$  is higher than for  $r2$ , the  $r1$  accumulator has a sufficient head start to get to its threshold first after time  $t_d$ . The resulting response is an error, with  $RT = t_0 + t_d$ . Available at <https://tinyurl.com/yc4n8lpm> under CC license <https://creativecommons.org/licenses/by/2.0/>.

mismatching accumulators,  $v_{\text{true}}$  and  $v_{\text{false}}$ , and a non-decision time parameter  $t_0$ . In order to identify the model, one accumulator parameter must be fixed (Donkin, Brown, & Heathcote, 2009); here we assumed that the standard deviations of the drift rate distributions were fixed to 1. In later simulations, we make only the minimum required assumption of fixing one parameter, in particular assuming  $s_{\text{true}} = 1$ . We generated a data set with 250 trials per stimulus (i.e., a total of 500 trials) from the full model using the following parameter values:  $A = 0.5$ ,  $B = 1$ ,

$v_{\text{true}} = 4$ ,  $v_{\text{false}} = 3$ , and  $t_0 = 0.2$ .

The full model was compared to a restricted model in which  $v_{\text{true}}$  was fixed to 3.55. The value 3.55 yields a Bayes factor close to one (equivalently, log Bayes factor of zero) and was chosen for two reasons. First, this value facilitates the implementation of the Savage-Dickey density ratio. The Savage-Dickey method relies on estimating the posterior density at the test value, which can be unreliable when the test value falls in the tail of the posterior distribution. We circumvented this problem by using a test value in the restricted model ( $v_{\text{true}} = 3.55$ ) relatively close to the generating parameter in the full model ( $v_{\text{true}} = 4$ ).

Second, this value makes discriminating between the models difficult, and allows us to point out the difference between inference and model inversion (Lee, 2018). Although the data have been generated from the full model, a Bayes factor close to one indicates that the data are just as likely under the restricted model as under the full model. This may at first appear as an undesirable property of the Bayes factor. This reasoning, however, confuses inference and model inversion. Model inversion means that if the data are generated from model  $\mathcal{M}_1$  and one fits the data-generating model  $\mathcal{M}_1$  and an alternative model  $\mathcal{M}_2$ , one is able to identify the data-generating model  $\mathcal{M}_1$  based on a model selection measure of interest. Consider, however, the following example. Suppose we are interested in comparing a null model which assumes that there is no difference in non-decision time  $t_0$  between two groups to an alternative model which allows the effect size to be different from zero. Suppose further that the alternative model is the data-generating model and we simulate data for a small number of synthetic participants assuming a small non-zero effect size, resulting in an observed effect size that, for this sample of participants, happens to be approximately zero. As a result, the simpler null model can account for the observed data almost equally well as the more complex data-generating model and may be favored on the ground of parsimony. As more observations are generated from the alternative model, however, it will become clear that the effect size is non-zero, and the support for the simpler null model will decrease – equivalently, the support for the more complex alternative model will increase. Hence, with a large enough number of observations, model inversion may be fulfilled.

This discussion highlights why the Bayes factor for the simulated LBA data set is indifferent: the number of trials is relatively small and the misspecified simpler model fixes  $v_{\text{true}}$  to 3.55 which is close to the data-generating value of 4. Therefore, the slight misspecification of the simpler restricted model is almost perfectly balanced out by its parsimony advantage compared to the more complex full model. The example is meant as a reminder that Bayesian inference conditions on the data at hand and that it may be reasonable to obtain evidence in favor of a different model than the data-generating one for certain data sets. Therefore, although one can assess the predictive adequacy of two competing models for the observed data using the Bayes factor (Wagenmakers, Marsman, et al., 2018), the Bayes factor should not be expected to necessarily recover a data-generating model in a simulation study. Nevertheless, as the number of observations grows large, the Bayes factor should select the correct model, a property known as model selection consistency (Bayarri et al., 2012).

#### 4.4.2 Prior Distributions

We used the following prior distributions for the different parameter types:

$$\begin{aligned}
 A &\sim \mathcal{N}_+(1, 1) \\
 B &\sim \mathcal{N}_+(1, 1) \\
 v_{\text{true}} &\sim \mathcal{N}(2, 3^2) \\
 v_{\text{false}} &\sim \mathcal{N}(1, 3^2) \\
 t_0 &\sim \mathcal{N}_{(0.1, \infty)}(0.3, 0.25^2),
 \end{aligned} \tag{4.8}$$

where  $\mathcal{N}(\mu, \sigma^2)$  denotes a normal distribution with mean  $\mu$  and variance  $\sigma^2$ ,  $\mathcal{N}_+(\mu, \sigma^2)$  denotes a normal distribution truncated to allow only positive values, and  $\mathcal{N}_{(x, y)}(\mu, \sigma^2)$  denotes a normal distribution with lower truncation  $x$  and upper truncation  $y$ . In the full model, we specified a prior distribution for all parameters, including  $v_{\text{true}}$ . In the restricted model, we specified a prior distribution for all parameters except  $v_{\text{true}}$ , as  $v_{\text{true}}$  was fixed to 3.55.

The priors in Equation 4.8 were taken from Heathcote et al. (2018). Although we believe that these priors provide a reasonable set up based on our experience with the LBA parameter ranges, they may be replaced by empirically informed priors in future applications. We also acknowledge that our prior choices are for many parameters wider than the ones used by Evans and Brown (2018); this may make the simple Monte Carlo method less efficient than when used in combination with the Evans-Brown priors.

#### 4.4.3 Parameter Estimation and Model Comparison

We used the DE-MCMC algorithm, as implemented in the DMC software (<https://osf.io/pbwx8/>) to estimate the model parameters. We set the number of MCMC chains to three times the number of model parameters; for the full model we ran 15 and for the restricted model we ran 12 chains with over-dispersed start values. In order to reduce auto-correlation, we thinned each MCMC chain to retain only every 10<sup>th</sup> posterior sample. During the burn-in period, the probability of a migration step was set to 5%; after burn-in, migration was turned off and only crossover steps were performed. Convergence of the MCMC chains was assessed by visual inspection and the  $\hat{R}$  statistic (Brooks & Gelman, 1998), which was below 1.05 for all parameters.<sup>9</sup> We obtained 10 independent sets of posterior samples for both the full and the restricted model, which were used to assess the uncertainty of the Bayes factor estimates.

Once the posterior samples were obtained, we computed the Bayes factor in favor of the full model using the Warp-III, the simple Monte Carlo, the Savage-Dickey, and the RJMCMC methods. The implementations of the four approaches are available at <https://osf.io/ynwpa/>. To assess the uncertainty of the Bayes factor estimates, we repeated each procedure 10 times for each model. For the

---

<sup>9</sup>It has been pointed out that  $\hat{R}$  is not a perfect indicator of convergence in certain scenarios (e.g., Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2019). For a recent proposal of an improved  $\hat{R}$ , see Vehtari, Gelman, et al. (2019).

#### 4. COMPUTING BAYES FACTORS FOR EVIDENCE-ACCUMULATION MODELS USING WARP-III BRIDGE SAMPLING

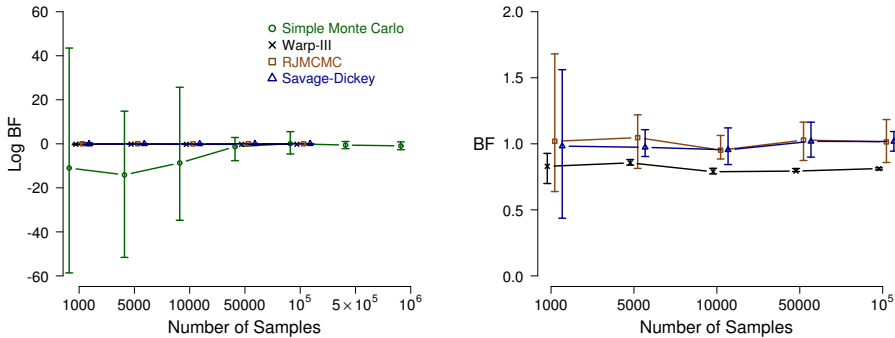


Figure 4.3: *Bayes factor estimates for the single-participant case as a function of the number of samples.* The left panel displays the *log* Bayes factor estimates computed using the Warp-III (black crosses), simple Monte Carlo (green circles), Savage-Dickey (blue triangles), and RJMCMC (brown squares) methods. The right panel displays the Bayes factors estimates computed using the Warp-III (black crosses), Savage-Dickey (blue triangles), and RJMCMC (brown squares) methods (i.e., omitting the simple Monte Carlo estimates and displaying the results on the Bayes factor and not *log* Bayes factor scale). For Warp-III, the *x*-axis corresponds to the number of posterior samples (collapsed across all chains) used for computing the marginal likelihood for each model. For simple Monte Carlo, it corresponds to the number of prior samples used for computing the marginal likelihoods. For Savage-Dickey, it corresponds to the number of posterior samples used to estimate the density of the posterior distribution at the test value (i.e., 3.55). For RJMCMC, it corresponds to the number of posterior samples used from each model (for details, see the Appendix). The symbols (i.e., crosses, circles, triangles, squares) indicate the median (log) Bayes factor estimates and bars indicate the range of the estimates across the 10 repetitions. Available at <https://tinyurl.com/y5brs44a> under CC license <https://creativecommons.org/licenses/by/2.0/>.

Warp-III, Savage-Dickey, and RJMCMC methods, we used a fresh set of posterior samples for each repetition.

#### 4.4.4 Results

The left panel of Figure 4.3 displays estimates of the *log* Bayes factor as a function of the number of samples. Note that we included an order of magnitude more samples for the simple Monte Carlo method in order to produce results that are comparable to estimates provided by the other methods. The right panel of Figure 4.3 zooms in on the results obtained with the Warp-III, Savage-Dickey, and RJMCMC methods and omits the simple Monte Carlo estimates; this panel shows the Bayes factor and *not* the *log* Bayes factor to facilitate interpretation.

All four methods eventually converged to a log Bayes factor estimate close to zero (equivalently, a Bayes factor estimate close to one). As the number of samples increased, the uncertainty of the estimates decreased. For this example, Warp-III resulted in the smallest uncertainty intervals. The Warp-III, Savage-Dickey, and RJMCMC methods resulted in stable Bayes factor estimates already with 1,000 samples. Although the three methods numerically did not yield the exact same Bayes factors, they all produced estimates close to one with relatively small uncertainty. The simple Monte Carlo method was clearly the least efficient; it produced wide uncertainty intervals and took approximately 50,000-100,000 samples to converge to the estimates from the other methods. Note that the number of samples required by the different methods for the stable and reliable estimation of the Bayes factor may vary depending on the characteristics of the specific example and should not be interpreted as a guideline.

Although in this particular example we were able to obtain stable and accurate Bayes factor estimates with all four methods, this is not necessarily the case for more complicated – non-nested and hierarchical – model selection problems. The Savage-Dickey method cannot be used for non-nested model comparison. Moreover, the Savage-Dickey estimate of the Bayes factor becomes very unstable if the test value falls in the tail of the posterior distribution because density estimates in the tails of the posterior are highly variable. Similarly, the RJMCMC approach cannot be easily generalized to situations involving non-nested comparisons. RJMCMC exploits the relations between the parameters of the models; however, if the models are non-nested it might be impossible to relate the two sets of parameters. Even generalizing RJMCMC to nested hierarchical comparisons is challenging because it involves linking a large number of parameters, especially if the vector of participant-level parameters differs between the two models for each participant. Furthermore, as a result of the strong parameter correlations in evidence-accumulation models, fixing one parameter in nested model comparisons can lead to substantial changes in the other parameters, making it even more difficult to efficiently link the competing models. Because of these challenges associated with non-nested and hierarchical model comparisons, we believe that the Savage-Dickey density ratio and RJMCMC methods are not suited as general model selection tools for evidence-accumulation models and will not be considered further.

The simple Monte Carlo and the Warp-III method can be used for both nested and non-nested model comparisons because they consider one model at a time.<sup>10</sup> In Warp-III, this also allows us to use a convenient proposal distribution chosen to maximize the overlap between the proposal and the posterior, which leads to a substantial gain in efficiency relative to simple Monte Carlo sampling. The inefficiency of simple Monte Carlo in our straightforward single-participant example suggests that this method is infeasible in many practical applications of hierarchical evidence-accumulation models. First, as also acknowledged by Evans and Brown (2018), simple Monte Carlo can result in highly variable Bayes factor es-

<sup>10</sup>In its original form, bridge sampling has been proposed to estimate the Bayes factor directly. In line with, for instance, Overstall and Forster (2010) here we advocate a version that estimates one marginal likelihood at a time (see also, Meng & Schilling, 2002, section 1.3).

timates in hierarchical settings. Second, the number of samples needed to obtain stable estimates with simple Monte Carlo sampling can quickly become unmanageable. This was indeed the case when we tried to apply it to the hierarchical model comparison problems outlined in the next section.<sup>11</sup>

## 4.5 Simulation Study II: Nested and Non-nested Model Comparison for the Hierarchical Case

As a second example, we considered eight LBA data sets that featured observations from multiple participants generated and fit using the hierarchical approach. We investigated the performance of Warp-III for two nested and two non-nested model comparison problems.

### 4.5.1 Models and Data

We simulated a design with four cells, two conditions that differed in a particular parameter crossed with two stimuli, and two possible responses. In the nested case, we compared a model that allowed only mean drift rate  $v_{\text{true}}$  to be different across conditions (i.e.,  $V$ -model) to a null model that featured one common  $v_{\text{true}}$  parameter for both conditions (i.e.,  $0$ -model). In the non-nested case, we compared the  $V$ -model to a model that allowed only threshold  $B$  to be different across conditions (i.e.,  $B$ -model). Note that we made these comparisons in both directions, for example, we computed the Bayes factor for the  $V$ -model vs.  $B$ -model comparison when the  $V$ -model generated the data, and computed the Bayes factor for the  $B$ -model vs.  $V$ -model comparison when the  $B$ -model generated the data.

We generated new data sets from both models in each comparison. We used two different combinations of the number of participants ( $n$ ) and the number of trials per cell ( $k$ ), both with 4,000 data points in total. Thus, overall there were eight different data sets: one for each of the four comparisons at each group size. In the first combination, we simulated data using  $n = 20$  with  $k = 200$ , corresponding to a smaller group of participants each measured fairly well. In the second combination, we simulated data using  $n = 80$  with  $k = 50$ , corresponding to a larger group of participants each measured at or below the lower bound of  $k$  required for acceptable individual estimation. These two cases exemplified either an emphasis on individual or group estimation. In the former case, the number of participants was at the lower bound of  $n$  required for acceptable estimation of the group-level parameters. In the latter case, estimation of the participant-specific parameters relied heavily on the additional constraint provided by the hierarchical structure.

To generate the data sets, we used normal group-level distributions for each parameter (truncated below to allow only positive values), specified the location ( $\mu$ ) and scale ( $\sigma$ ) of the group-level distributions, and then simulated participant-specific parameters from these normal distributions. Subsequently, the participant-specific parameters were used to generate trials for each participant.

---

<sup>11</sup>We thank Nathan Evans for attempting to apply simple Monte Carlo sampling to one of our hierarchical model comparison examples.

To ensure identifiability, the standard deviation of the drift rate corresponding to the accumulator for the correct response,  $s_{\text{true}}$ , was fixed to one for every participant.

To generate data from the  $V$ -model, we used the following  $\mu$  parameters (where bracketed superscripts indicate experimental condition):  $\mu_A = 1$ ,  $\mu_B = 0.4$ ,  $\mu_{v_{\text{true}}^{(1)}} = 4$ ,  $\mu_{v_{\text{true}}^{(2)}} = 3$ ,  $\mu_{v_{\text{false}}} = 1$ ,  $\mu_{s_{\text{false}}} = 1$ , and  $\mu_{t_0} = 0.3$ . For the 0-model, we used  $\mu_A = 1$ ,  $\mu_B = 0.4$ ,  $\mu_{v_{\text{true}}} = 3$ ,  $\mu_{v_{\text{false}}} = 1$ ,  $\mu_{s_{\text{false}}} = 1$ , and  $\mu_{t_0} = 0.3$ . For the  $B$ -model, we used  $\mu_A = 1$ ,  $\mu_{B^{(1)}} = 0.3$ ,  $\mu_{B^{(2)}} = 0.7$ ,  $\mu_{v_{\text{true}}} = 3.5$ ,  $\mu_{v_{\text{false}}} = 1$ ,  $\mu_{s_{\text{false}}} = 1$ , and  $\mu_{t_0} = 0.3$ . The data-generating  $\sigma$  parameters were obtained by dividing the  $\mu$  parameters by 10, resulting in appreciable but not excessive individual differences in the participant-specific parameters.

## 4.5.2 Prior Distributions

We used zero-bounded truncated normal group-level distributions to model individual differences in the parameters. We used the following prior distributions for the group-level parameters:

$$\begin{aligned}
 \mu_A, \sigma_A &\sim \mathcal{N}_+(1, 1) \\
 \mu_B, \sigma_B &\sim \mathcal{N}_+(0.4, 0.4^2) \\
 \mu_{v_{\text{true}}}, \sigma_{v_{\text{true}}} &\sim \mathcal{N}_+(3, 3^2) \\
 \mu_{v_{\text{false}}}, \sigma_{v_{\text{false}}} &\sim \mathcal{N}_+(1, 1) \\
 \mu_{s_{\text{false}}}, \sigma_{s_{\text{false}}} &\sim \mathcal{N}_+(1, 1) \\
 \mu_{t_0}, \sigma_{t_0} &\sim \mathcal{N}_+(0.3, 0.3^2).
 \end{aligned} \tag{4.9}$$

As for the single-participant case, we believe that the priors provide a reasonable set up but they may be replaced by empirically informed priors in future applications.

## 4.5.3 Parameter Estimation and Model Comparison

We used the DE-MCMC algorithm, as implemented in the DMC software to estimate the model parameters. We first estimated parameters separately for each synthetic participant, similar to our previous single-participant example. The result of this phase provided the starting values for the hierarchical analysis. For each model, we set the number of MCMC chains to three times the number of participant-specific parameters. We thinned each MCMC chain to retain only every  $10^{\text{th}}$  posterior sample. Burn-in was accomplished by DMC's `h.run.unstuck.dmc` function with a 5% migration probability. We then used the `h.run.converge.dmc` function with no migration until 250 iterations were obtained that appeared to be converged to the stationary distribution ( $\hat{R} < 1.1$ ). Further iterations were then added using the `h.run.dmc` function until we obtained approximately 100,000 posterior samples per parameter (the exact number of samples varied because the number of MCMC chains varied among the different models). With this very large number of samples,  $\hat{R}$  was very close to 1 for all parameters at both the group and participant levels. We obtained 10 independent

#### 4. COMPUTING BAYES FACTORS FOR EVIDENCE-ACCUMULATION MODELS USING WARP-III BRIDGE SAMPLING

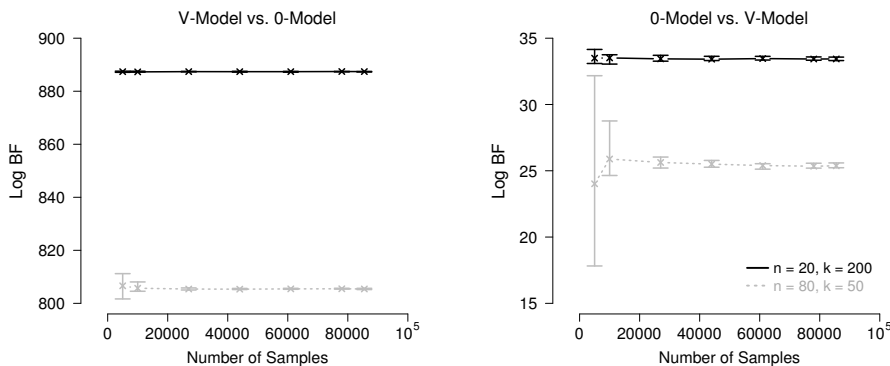


Figure 4.4: *Log Bayes factor estimates obtained with Warp-III sampling for the nested hierarchical model comparisons as a function of the number of posterior samples (collapsed across all chains) used for computing the marginal likelihood for each model.* Crosses indicate the median log Bayes factor estimates and bars indicate the range of the estimates across the 10 repetitions. The left panel shows results for the data sets generated from the V-model; the right panel shows results for the data sets generated from the 0-model. Results for  $n = 20$  with  $k = 200$  are displayed in black; results for  $n = 80$  with  $k = 50$  are displayed in gray with dotted lines. The log Bayes factor is expressed in favor of the data-generating model. Available at <https://tinyurl.com/yxgsgjaj> under CC license <https://creativecommons.org/licenses/by/2.0/>.

sets of posterior samples for each model, which were used to assess the uncertainty of the Bayes factor estimates.

Once the posterior samples were obtained, we computed the Bayes factor in favor of the data-generating models using Warp-III.<sup>12</sup> For each model, we assessed the uncertainty of the estimates by running the Warp-III sampler 10 times using a fresh set of posterior samples for each repetition.

#### 4.5.4 Results

Figure 4.4 shows the log Bayes factor estimates obtained with Warp-III sampling as a function of the number of samples for the nested comparisons and Figure 4.5 shows the results for the non-nested comparisons.<sup>13</sup> The log Bayes factors are expressed in favor of the data-generating models.

<sup>12</sup>We provide R code for an exemplary hierarchical model (i.e., code for the  $B$ -model with data generated from the  $B$ -model using  $n = 20$ ,  $k = 200$ ) at <https://osf.io/ynwpa/>. The reason why we only provide code for one of the hierarchical examples is that (1) the data sets are simulated and one example is sufficient to show how to apply the method (the other examples are obtained via trivial changes to the code), (2) the corresponding files are *very* large. Files for the other examples are available upon request.

<sup>13</sup>More fine-grained versions of Figure 4.4 and Figure 4.5 are available at <https://osf.io/ynwpa/>.

#### 4.5. Simulation Study II: Nested and Non-nested Model Comparison for the Hierarchical Case

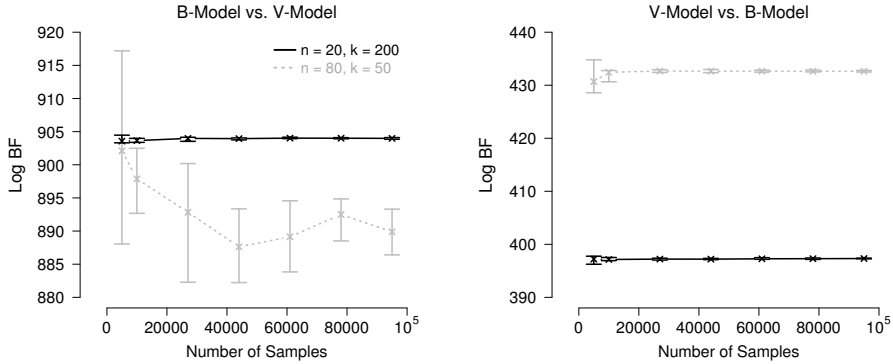


Figure 4.5: *Log Bayes factor estimates obtained with Warp-III sampling for the non-nested hierarchical model comparisons as a function of the number of posterior samples (collapsed across all chains) used for computing the marginal likelihood for each model.* Crosses indicate the median log Bayes factor estimates and bars indicate the range of the estimates across the 10 repetitions. The left panel shows results for the data sets generated from the *B*-model; the right panel shows results for the data sets generated from the *V*-model. Results for  $n = 20$  with  $k = 200$  are displayed in black; results for  $n = 80$  with  $k = 50$  are displayed in gray with dotted lines. The log Bayes factor is expressed in favor of the data-generating model. Available at <https://tinyurl.com/y3f71263> under CC license <https://creativecommons.org/licenses/by/2.0/>.

The figures illustrate that Warp-III resulted in stable Bayes factor estimates in favor of the data-generating model with narrow uncertainty intervals in all but one case, the non-nested *B*-model vs. *V*-model comparison for the  $n = 80$  with  $k = 50$  data set. For this data set, the iterative scheme from Equation 4.5 initially did not seem to converge, but instead oscillated between two different values, say  $x_1$  and  $x_2$ . We were able to achieve convergence by stopping the iterative scheme and re-starting it with the initial guess of the marginal likelihood set to the geometric mean of the two values between which the estimate initially oscillated (i.e., the square root of the product of  $x_1$  and  $x_2$ ). Although this approach enabled us to obtain an estimate of the marginal likelihood, the uncertainty of this estimate was noticeably larger than for the other cases. Nevertheless, this estimate was sufficiently certain to conclude that the Bayes factor clearly favored the *B*-model.<sup>14</sup>

The results show that the hierarchical model comparisons required substantially more samples than the single-participant case. Note also that more samples

<sup>14</sup>Note that in practice, very large log Bayes factor estimates as in this case (e.g., 880 – 920) yield the same conclusion independent of the exact number: overwhelming evidence for the favored model. However, when the estimated Bayes factor is closer to 1 (equivalently, log Bayes factor closer to 0), it is more important that the Bayes factor is estimated precisely as this may influence which model is favored (see, e.g., the single-participant example and the following example).

were needed for the  $n = 80$  with  $k = 50$  data sets than for the  $n = 20$  with  $k = 200$  data sets to obtain comparable uncertainty intervals. The reason is that the number of participants,  $n$ , determines how many participant-specific parameters need to be integrated out, whereas the number of trials per cell,  $k$ , does not affect the number of model parameters. Therefore, increasing the number of participants increases the dimensionality of the integral in Equation 4.2 that is estimated via Warp-III. It is likely that the greater difficulty in obtaining well-behaved participant-specific parameter estimates with  $k = 50$  has also contributed to the larger uncertainty intervals.

All Bayes factors yielded overwhelming evidence for the data-generating model, including the ones computed for the data sets generated from the nested 0-model (i.e., right panel of Figure 4.4). Note, however, that the magnitude of the Bayes factors for these nested examples is smaller than for the other examples. This result is not unexpected: the  $V$ -model can account for all data sets that the 0-model can account for and, additionally, also for data sets that show a difference in  $v_{\text{true}}$  between conditions. Therefore, the Bayes factor can only favor the 0-model due to parsimony and not because it describes the data better than the  $V$ -model. Note also that although the Bayes factors clearly favored the data-generating models, this may not necessarily be the case in other examples. As outlined in our earlier discussion of model inversion, Bayesian inference conditions on the data at hand and it may be reasonable to obtain evidence in favor of a different model than the data-generating one for certain data sets.

## 4.6 Simulation Study III: Estimating Equivocal Bayes Factors for the Hierarchical Case

In the previous section, it was demonstrated that Warp-III yields stable and precise Bayes factor estimates for different hierarchical examples. Many of these Bayes factor estimates were very large and it could be argued that for large Bayes factors, obtaining very precise estimates is not crucial since the qualitative conclusion (“overwhelming evidence”) will not change unless the estimation uncertainty is extremely large. In this section, we demonstrate that Warp-III is also able to provide precise estimates of a Bayes factor close to 1 for the hierarchical case. Estimating Bayes factors in this range precisely is important since a large estimation uncertainty would make it difficult to judge which model is favored.

### 4.6.1 Models and Data

For this example, we reused the data set generated from the  $B$ -model with  $n = 20$  and  $k = 200$  described in the previous section. We compared the data-generating  $B$ -model to a restricted  $B_{\text{res}}$ -model. The  $B_{\text{res}}$ -model was identical to the  $B$ -model except that the group-level parameter  $\mu_{v_{\text{false}}}$  was fixed to 1.24. This value was chosen to yield a Bayes factor close to 1.<sup>15</sup>

---

<sup>15</sup>This model comparison may be regarded as artificial, however, the main goal of the example is to demonstrate that, even in the hierarchical setting, a Bayes factor of about 1 can be estimated precisely using Warp-III.

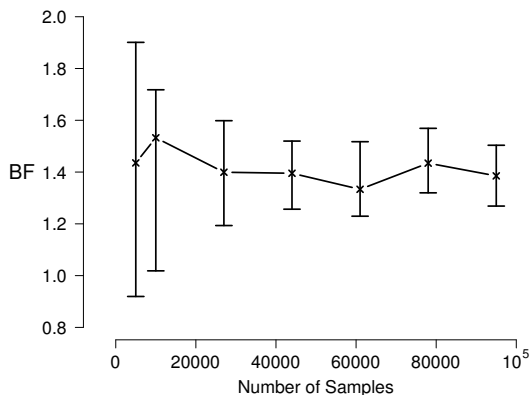


Figure 4.6: *Bayes factor estimates obtained with Warp-III sampling for the  $B$ -model vs.  $B_{\text{res}}$ -model example as a function of the number of posterior samples (collapsed across all chains) used for computing the marginal likelihood for each model.* Crosses indicate the median Bayes factor estimates and bars indicate the range of the estimates across the 10 repetitions. The data set was generated from the  $B$ -model with  $n = 20$  and  $k = 200$  and is identical to the one used in the left-panel of Figure 4.5. The Bayes factor is expressed in favor of the data-generating model. Available at <https://tinyurl.com/y599st45> under CC license <https://creativecommons.org/licenses/by/2.0/>.

### 4.6.2 Prior Distributions

The prior distributions were identical to the ones used in the previous hierarchical example. Note that for the  $B_{\text{res}}$ -model, the group-level parameter  $\mu_{v_{\text{false}}}$  was fixed to 1.24 and was not assigned a prior distribution.

### 4.6.3 Parameter Estimation and Model Comparison

Parameter estimation and model comparison was conducted in an analogous manner to the previous hierarchical example. Note that we reused the log marginal likelihood estimates for the  $B$ -model from the previous example which was based on the exact same data set.

### 4.6.4 Results

Figure 4.6 shows the Bayes factor (*not* log Bayes factor) estimates obtained with Warp-III sampling as a function of the number of samples. The Bayes factor is expressed in favor of the data-generating  $B$ -model. The figure illustrates that Warp-III resulted in stable Bayes factor estimates with narrow uncertainty intervals. The estimated Bayes factor is slightly larger than 1 indicating that the

data-generating  $B$ -model is slightly favored. Nevertheless, a Bayes factor close to 1 indicates that none of the models is favored in a compelling fashion by the data at hand; the evidence is ambiguous.

### 4.7 Discussion

Over the last decade, the Bayesian estimation of evidence-accumulation models has gained momentum (e.g., Heathcote et al., 2018; Vandekerckhove et al., 2011; Wiecki et al., 2013). This increase in popularity is largely attributable to the advantages afforded by the Bayesian hierarchical framework that allows researchers to obtain well-constrained parameter estimates even in situations with relatively few observations per participant. Despite recent advances in the Bayesian estimation of evidence-accumulation models, model comparison continues to rely on suboptimal procedures, such as posterior parameter inference and model selection criteria known to favor overly complex models.

In this chapter, therefore, we advocated model selection for evidence-accumulation models based on the Bayes factor (e.g., Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Ly et al., 2016a). The Bayes factor is given by the ratio of the marginal likelihoods of the competing models and thus enables the quantification of relative evidence on a continuous scale (e.g., Wagenmakers, Marsman, et al., 2018). The Bayes factor implements a trade-off between parsimony and goodness-of-fit (Jefferys & Berger, 1992; Myung & Pitt, 1997) and is considered as “the standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis & Raftery, 1997, p. 648). Bayes factors enable the computation of posterior model probabilities, which provide an intuitive metric for comparison among models. Bayes factors also enable Bayesian model averaging, which avoids the need to make categorical decisions between models and which produces better calibrated predictions (e.g., Hoeting et al., 1999). Bayes factors are well suited for the type of model comparison problems that are faced by cognitive modelers because they do not favor overly complex models, and so guard against the proliferation of “crud factors” that plague psychology (Meehl, 1990).

Despite the advantages afforded by the Bayesian framework, Bayes factors are rarely, if ever, used for evidence-accumulation models, largely because of the computational challenges involved in the evaluation of the marginal likelihood. Here we advocated Warp-III bridge sampling (Meng & Schilling, 2002) for computing the marginal likelihood – and hence the Bayes factor – for evidence-accumulation models. We believe that Warp-III is well suited for cognitive models in general and evidence-accumulation models in particular because, as we have shown, it can be straightforwardly applied to hierarchical models and non-nested comparisons, unlike the simple Monte Carlo and the Savage-Dickey approaches. Moreover, Warp-III is relatively easy to implement, and requires only the posterior samples routinely collected during parameter estimation. In contrast to transdimensional MCMC methods, such as RJMCMC, it does not require changing the sampling algorithm or linking the competing models, which can be problematic for hierarchical and non-nested models. We have shown that Warp-III bridge sampling is

practically feasible even in complex and high-dimensional hierarchical instantiations of the Linear Ballistic Accumulator (LBA; S. D. Brown & Heathcote, 2008). Although we encountered a challenging case with scarce participant-level data (left panel of Figure 4.5), even in that case we were able to detect and ameliorate the convergence problem.

Once the posterior samples are obtained, computing the marginal likelihood for the single-participant case using Warp-III is relatively fast. For each repetition, it took approximately 13 minutes to run the Warp-III sampler with 100,000 posterior samples, using four CPU cores on our servers. As these servers are old and the individual cores relatively slow given they are embedded in 16-core chips, more modern quad-core laptops will achieve the task in a much shorter time. Naturally, in the hierarchical setting, the computational burden is higher and strongly depends on the number of participants. For instance, for the  $V$ -model vs.  $B$ -model comparison (right panel in Figure 4.5) in combination with  $n = 20$  and  $k = 200$ , running the Warp-III sampler with 95,000 posterior samples took approximately 7 hours, using four CPU cores on our servers. In contrast, for the  $n = 80$  and  $k = 50$  case, the computational time was approximately 25 hours. However, it is important to note that it was not necessary to collect such a high number of posterior samples. For the individual case, the Bayes factor estimate was precise and stable after only 1,000 samples. For most hierarchical comparisons, we obtained well-behaved Bayes factor estimates with approximately 20,000-30,000 samples. Note also that the computational time strongly depends on the specific programming language used for evaluating the likelihood and the prior. Our implementation relies on R (R Core Team, 2019), but integrating the Warp-III sampler with Lin and Heathcote’s (2017) C++ implementation of the LBA and the DDM is expected to speed up sampling by an order of magnitude. In summary, although Warp-III is computationally more intensive than using model selection criteria such as the DIC (Spiegelhalter et al., 2002), in standard applications of evidence-accumulation models, the computational costs are manageable, even using personal computers. We believe that the computational costs of Warp-III are a small price to pay for the advantages afforded by the use of principled Bayesian model selection techniques. Where practical issues are faced due to the need to select among a large number of models, researchers may consider an initial triage using easy-to-compute alternatives, such as DIC, in order to obtain a candidate set for model selection based on Bayes factors (for related approaches, see Madigan & Raftery, 1994, and Overstall & Forster, 2010).

As many evidence-accumulation models have analytic likelihoods, and so are amenable to MCMC methods for obtaining posterior distributions, Warp-III sampling is not limited to the LBA, but may be readily applied to other models, such as the Diffusion Decision Model (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008). Heathcote et al.’s (2018) DMC software enables the hierarchical MCMC-based estimation of not only the LBA and the DDM, but also a variety of other models including single-boundary and racing diffusion models (Leite & Ratcliff, 2010; Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014; Tilman, Strayer, et al., 2017), lognormal race models (Heathcote & Love, 2012; Rouder, Province, Morey, Gómez, & Heathcote, 2015), as well as race models of the stop-signal paradigm (Matzke et al., 2013; Matzke, Love, & Heathcote, 2017). Our easy-to-use R-implementation

of the Warp-III sampler enables the computation of the marginal likelihood of any model implemented in the DMC software. When analytic likelihoods are not available, approximate Bayesian computation may be used to enable MCMC sampling, opening up the possibility to explore more complex and realistic cognitive process models (Holmes, Trueblood, & Heathcote, 2016; Turner & Sederberg, 2014), although this approach remains challenging (e.g., Lin & Heathcote, 2019). Future research should investigate the performance of simulation-based methods, such as Warp-III, in the context of models without analytic likelihood.

As illustrated in our single-participant example, the Bayes factor will not necessarily select a data-generating model. In contrast, as explained in detail before, it might be the case that the Bayes factor favors a model different than the data-generating one for certain data sets. However, in the single-participant example and in the final hierarchical example, the Bayes factor did not clearly favor a model different than the data-generating one but was approximately one, meaning that both models were about equally likely. Thus, another advantage of Bayes factors is that they allow one to disentangle evidence of absence (i.e., the Bayes factor favors the simpler model) and absence of evidence (i.e., the Bayes factor is approximately one).

It is crucial to acknowledge that the Bayes factor critically depends on the prior distribution of the model parameters. We emphasize that the priors we used in the present chapter are not the gold standard for the LBA. We are presently developing empirically informed prior distributions for the LBA and the DDM based on archival data sets. In the meantime, we recommend that researchers develop their own empirically based priors (perhaps through pilot work or analysis of related archival data sets) in LBA applications. For the DDM, the distributions of parameter values in Matzke and Wagenmakers (2009) already provide reasonable priors. We see the development of theoretically and empirically informed prior distributions as necessary part of the maturation of any well-specified quantitative model, consistent with the position of Lee and Vanpaemel (2018).

### 4.7.1 Practical Recommendations

In this final section, we provide recommendations about the use of Warp-III sampling in practical applications. Our recommendations should not be interpreted as strict guidelines, but rather as suggestions based on our experience of using Warp-III in the context of cognitive models in general and evidence-accumulation models in particular.

#### 4.7.1.1 How to Assess the Uncertainty and Stability of the Estimate

Once the data have been observed and the model (i.e., the likelihood and the prior) have been specified, there is a single *true* marginal likelihood corresponding to a particular data-model combination. However, for (hierarchical) evidence-accumulation models, the true marginal likelihood cannot be computed analytically and must be estimated. As with all estimates, the marginal likelihood provided by Warp-III is uncertain and may vary even for the same data-model

combination. Consequently, it is crucial to assess and report the uncertainty of the estimate and investigate the degree to which uncertainty affects conclusions.

Our recommendation is to assess the uncertainty directly for the quantity of interest. For instance, when conclusions are based on the Bayes factor, researchers should assess the uncertainty of the Bayes factor; when conclusions are based on posterior model probabilities, researchers should assess the uncertainty of the posterior model probabilities. To do so, we recommend researchers to compute the quantity of interest repeatedly based on independent runs of Warp-III. For example, when one is interested in estimating the Bayes factor, one should repeatedly (1) draw fresh posterior samples from the competing models; (2) use Warp-III to estimate the marginal likelihood of the models; and (3) compute the resulting Bayes factor. The uncertainty of the estimate can then be assessed by considering the empirical variability of the Bayes factor estimates across the repetitions. The empirical assessment of uncertainty is generally considered as the gold standard, even when approximate errors are available such as for the simple multivariate normal bridge sampling estimator (e.g., Frühwirth-Schnatter, 2004).<sup>16</sup>

We find it useful to not only assess the uncertainty, but also to investigate whether the estimate of the quantity of interest (e.g., Bayes factor) has stabilized. As our simulations demonstrated, when successively increasing the number of samples, the estimate becomes more precise and – after some initial fluctuation – tends to stabilize. One way to assess stability is to compute the quantity of interest using batches of the available posterior samples, as we have done in our simulations. However, we acknowledge that this process can be time consuming. A crude alternative is to compute the estimate with the corresponding uncertainty based on (at least) three different samples sizes, for instance, (a)  $\frac{1}{3}$ , (b)  $\frac{2}{3}$ , and (c) all of the posterior samples. Considering the sequence of these three estimates allows one to get an idea about whether the estimate has stabilized.

#### 4.7.1.2 How Many Samples Are Required for Precise and Stable Estimates

Assessing the uncertainty and stability of the estimate is a natural and – in our opinion – the best approach to determine the number of samples required for reliable conclusions. Note that the required level of precision and stability depends on the particular application. For instance, for one of our non-nested hierarchical examples (left panel in Figure 4.5), the Bayes factor estimates were relatively uncertain and fluctuated quite substantially even in the high-sample region. However, given that all of the estimates provided overwhelming evidence for the *B*-model, the achieved accuracy and stability were sufficiently high to conclude that the *B*-model was clearly favored over the *V*-model. In contrast, in situations when the Bayes factor estimates do not provide compelling evidence for either model (for instance, when the Bayes factor estimates are varying around 1), it is crucial to obtain more precise and stable estimates to ensure that fluctuations do not influence which of the two models is favored or whether it is concluded that the evidence is

<sup>16</sup> Another complication with approximate errors for separate marginal likelihood estimates is that it is not completely straightforward to derive an approximate error for the resulting Bayes factor estimate.

equivocal. The single-participant and the final hierarchical example indicate that it is possible to obtain precise and stable Warp-III Bayes factor estimates also for this Bayes factor range.

Given these considerations, combined with the fact that the quality of the estimate depends on factors such as the number of participants and the complexity of the models, we are unable to provide general recommendations about the number of samples necessary for the reliable application of Warp-III sampling. Warp-III requires more posterior samples than one would typically collect for the purpose of parameter estimation. In our experience, a minimum of 1,000-2,000 posterior samples (collapsed across chains) typically provides a reasonable starting point in single-participant applications. In hierarchical applications, we recommend at least 10,000-20,000 samples. Nevertheless, as with all simulation-based methods, the more samples, the better. Note that our recommendations assume that the posterior samples are not highly auto-correlated; the degree of thinning in our simulations resulted in posterior samples that were virtually uncorrelated. Although autocorrelation is not itself necessarily a problem for parameter estimation, it does reduce the effective number of samples, and when large numbers of samples are required it is practically efficient to thin the samples, at least to the degree that there is little loss of effective sample size. Warp-III also benefits from having posterior samples with low autocorrelation. One reason is that the “optimal” bridge function is only optimal in case the posterior samples are independent and identically distributed which is not the case when using MCMC methods. However, some autocorrelation may not be too worrisome since, in our implementation, we use an effective sample size in this bridge function.

### 4.7.1.3 When to Use Simple Bridge Sampling and When to Use Warp-III Sampling

The Warp-III estimator is an advanced version of the “simple” multivariate normal bridge sampling estimator (e.g., Overstall & Forster, 2010). Warp-III matches the first three moments of the posterior and the proposal distribution; the multivariate normal approach – which is equivalent to Warp-II – matches only the first two moments of the distributions. As the precision of the estimate of the marginal likelihood is governed by the overlap between the posterior and the proposal distribution, the Warp-III estimate is at least as precise as the estimate computed using simple bridge sampling.<sup>17</sup> With symmetric posterior distributions, the advantage of Warp-III diminishes, but nothing is lost in terms of precision relative to simple bridge sampling. In contrast, with skewed posterior distributions, Warp-III results in more precise estimates because it is able to match the posterior and the proposal more closely. Note that both Warp-III and simple bridge sampling assume that the posterior samples are allowed to range across the entire real line. Hence, the skew of the posterior distributions must be assessed after the appropriate transformations. This does not mean that sampling from the posterior distributions must occur with all parameters transformed to the real line.

---

<sup>17</sup>For multi-modal posterior distributions, both simple bridge sampling and Warp-III sampling may result in insufficient overlap between the posterior and proposal distribution, and should be used with caution.

Table 4.1: Overview of the transformations used in the Warp-III implementation.  $\theta_i$  denotes a parameter and  $\omega_i$  denotes the corresponding new parameter that is obtained after having transformed  $\theta_i$  to the real line.  $l$  denotes a parameter lower bound and  $u$  denotes an upper bound.  $\Phi(\cdot)$  denotes the cumulative distribution function and  $\phi(\cdot)$  the probability density function of the normal distribution. The table displays the parameter type, the corresponding transformation, inverse-transformation, and the relevant Jacobian contribution.

Type	Transformation	Inv.-Transformation	Jacobian Contribution
unbounded	$\omega_i = \theta_i$	$\theta_i = \omega_i$	$\left  \frac{\partial \theta_i}{\partial \omega_i} \right  = 1$
lower-bounded	$\omega_i = \log(\theta_i - l)$	$\theta_i = \exp(\omega_i) + l$	$\left  \frac{\partial \theta_i}{\partial \omega_i} \right  = \exp(\omega_i)$
upper-bounded	$\omega_i = \log(u - \theta_i)$	$\theta_i = u - \exp(\omega_i)$	$\left  \frac{\partial \theta_i}{\partial \omega_i} \right  = \exp(\omega_i)$
double-bounded	$\omega_i = \Phi^{-1}\left(\frac{\theta_i - l}{u - l}\right)$	$\theta_i = (u - l)\Phi(\omega_i) + l$	$\left  \frac{\partial \theta_i}{\partial \omega_i} \right  = (u - l)\phi(\omega_i)$

In fact, in our simulations, only the  $v$  parameters were sampled on the real line; all other parameters were transformed to the real line after the posterior samples have been obtained. Our R-implementation of the Warp-III sampler automatically applies the appropriate transformations to the posterior samples obtained with the DMC software. Specifically, the implementation assumes that each posterior component can be transformed separately<sup>18</sup> and distinguishes between four different parameter types: (1) unbounded parameters, (2) lower-bounded parameters, (3) upper-bounded parameters, and (4) double-bounded parameters (i.e., parameters that have a lower and an upper bound). Table 4.1 displays the transformations that are used for the different parameter types. After having detected the parameter type, an appropriate transformation is applied and the expressions are adjusted by the relevant Jacobian contribution (see Table 4.1).

In general, Warp-III is a more powerful tool than simple bridge sampling for estimating the marginal likelihood, but the gain in precision depends on the particular application. A potential advantage of simple bridge sampling is its relative speed. Warp-III results in a mixture representation which requires one to evaluate the un-normalized posterior twice as often as in simple bridge sampling (e.g., Gronau, Wagenmakers, et al., 2019; Overstall, 2010). This implies a speed-accuracy trade-off: simple bridge sampling may be less precise but faster; Warp-III may be more precise but slower. Of course, one may increase the precision of the simple bridge sampling estimate by increasing the number of posterior samples. However, this approach neglects the fact that – in evidence-accumulator models in particular – obtaining the posterior samples typically takes substantially longer than computing the marginal likelihood using Warp-III. Therefore, although simple bridge sampling is faster for a given (initial) set of posterior samples, it is not necessarily true that it is more efficient to run the simpler version based on

<sup>18</sup>Consequently, the code would need to be adjusted to allow for covariance matrix parameters or probability vector parameters where constraints apply jointly to several components.

additional posterior samples than to run Warp-III on the initial set of samples to obtain comparable precision. Furthermore, we expect that the problem of seemingly non-converging estimates may be more frequent when using simple bridge sampling. Although this can be addressed by restarting the iterative scheme from an appropriately chosen start value, as shown in the left panel of Figure 4.5, this solution substantially increases the uncertainty of the estimate.

In situations where the joint posterior is exactly multivariate normal,<sup>19</sup> simple bridge sampling is clearly more efficient than Warp-III. However, it is challenging to assess multivariate normality in the high-dimensional spaces regularly encountered in hierarchical evidence-accumulation models. Although evaluating the marginal posterior distributions is feasible in most standard applications, normality of the marginals – which is often not the case for evidence-accumulation models applied to scarce data – does not necessarily imply that the joint posterior is multivariate normal. In sum, if one expects multivariate normal posterior distributions, simple bridge sampling is more efficient and should be preferred. Whenever this is not the case, we recommend Warp-III sampling.

## 4.8 Conclusion

In this chapter we advocated Warp-III bridge sampling as a general method for estimating the marginal likelihood – and hence the Bayes factor – for evidence-accumulation models. We demonstrated that Warp-III sampling provides a powerful and flexible approach that can be applied to both nested and non-nested model comparisons and – once posterior samples from the competing models have been obtained – it is straightforward to implement even in hierarchical settings. We believe that our easy-to-use and freely available implementation of Warp-III sampling will greatly facilitate the use of principled Bayesian model selection in practical applications of evidence-accumulation models.

R scripts for reproducing the results presented in this chapter are available at <https://osf.io/ynwpa/>.

---

<sup>19</sup>As before, multivariate normality should hold for the appropriately transformed posterior distribution.

## 4.A Savage-Dickey Density Ratio

Suppose that the parameter vector  $\theta$  can be partitioned into a set of nuisance parameters  $\zeta$  and test-relevant parameters  $\eta$  so that  $\theta = (\zeta, \eta)$ . The Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010) can then be used to compute the Bayes factor for testing whether  $\eta$  is equal to a constant  $\eta_0$  in the presence of nuisance parameters  $\zeta$ . Concretely, the Bayes factor compares model  $\mathcal{M}_0$  which assigns  $\zeta$  the prior density  $p_0(\zeta)$  and fixes  $\eta$  to the constant  $\eta_0$  to model  $\mathcal{M}_1$  which assigns  $\zeta$  and  $\eta$  the joint prior density  $p_1(\zeta, \eta)$ . The Savage-Dickey density ratio representation of the Bayes factor is then given by

$$\text{BF}_{01} = \frac{p_1(\eta_0 \mid \mathbf{y})}{p_1(\eta_0)}, \quad (4.10)$$

where  $p_1(\eta_0 \mid \mathbf{y})$  denotes the marginal posterior density of  $\eta$  under  $\mathcal{M}_1$  evaluated at  $\eta_0$  and  $p_1(\eta_0)$  denotes the marginal prior density of  $\eta$  under  $\mathcal{M}_1$  evaluated at  $\eta_0$ . Note that this representation is only valid in case  $p_1(\zeta \mid \eta_0) = p_0(\zeta)$ . Hence, conditional on  $\eta = \eta_0$ , the prior density for  $\zeta$  under  $\mathcal{M}_1$  must be identical to the prior density of  $\zeta$  under  $\mathcal{M}_0$ .<sup>20</sup> In our single-participant example, this assumption holds since the prior under  $\mathcal{M}_1$  is given by  $p_1(\zeta, \eta) = p_0(\zeta)p_1(\eta)$ . We used a logspline density estimator (Koopberg, 2016) to estimate the marginal posterior density at the point of interest.

## 4.B Reversible Jump Markov Chain Monte Carlo

Reversible jump Markov chain Monte Carlo (RJMCMC; Green, 1995) refers to an MCMC sampler on an enlarged state space which incorporates a model indicator  $M$  as an additional unknown. The posterior of the model indicator  $M$  can be used to estimate posterior model probabilities and posterior model odds. An estimate of the Bayes factor can be obtained by dividing the estimated posterior model odds by the known prior model odds. Barker and Link (2013) described a version of RJMCMC that represents the process intuitively as a Gibbs sampler where updates of the model indicator  $M$  are alternated with updates of a “palette” parameter vector  $\psi$ . The palette vector  $\psi$  has dimension  $d = \max\{\dim(\theta_k)\}$  where  $\theta_k$  denotes the parameter vector for model  $M_k$ ,  $k = 1, 2, \dots, K$  and  $K$  denotes the number of models under consideration.<sup>21</sup> Each model’s parameter vector  $\theta_k$  can be obtained from the palette vector  $\psi$  by a known invertible mapping  $g_k(\psi) = \xi_k = (\theta_k, \mathbf{u}_k)$ , where  $\mathbf{u}_k$  denotes a vector of auxiliary variables which is redundant to model  $M_k$  but ensures that the dimensionality of  $\psi$  and  $\xi_k$  matches.

The full-conditional distributions for the Gibbs sampler are determined by the joint model  $p(\mathbf{y}, \psi, M) = p(\mathbf{y} \mid \psi, M)p(\psi \mid M)p(M)$ . The model prior  $p(M)$  is set by the researcher and evaluating the likelihood  $p(\mathbf{y} \mid \psi, M)$  for a specific model  $M_k$  is straightforward since the model-specific parameter vector  $\theta_k$  can

<sup>20</sup>Verdinelli and Wasserman (1995) proposed a generalization of the Savage-Dickey density ratio that relaxes this assumption.

<sup>21</sup>Technically,  $d \geq \max\{\dim(\theta_k)\}$ , that is, the dimensionality of  $\psi$  could be larger than the maximum dimensionality of the model parameter vectors, however, this is uncommon in practice.

be obtained from  $\boldsymbol{\psi}$  using the function  $g_k$ . The prior  $p(\boldsymbol{\psi} \mid M)$  is obtained by applying the change of variables theorem. Recall that  $\boldsymbol{\psi} = g_k^{-1}(\boldsymbol{\xi}_k)$  and  $\boldsymbol{\xi}_k = (\boldsymbol{\theta}_k, \mathbf{u}_k)$ . Furthermore, note that the prior  $p(\boldsymbol{\xi}_k \mid M_k) = p(\boldsymbol{\theta}_k, \mathbf{u}_k \mid M_k)$  factorizes as  $p(\boldsymbol{\xi}_k \mid M_k) = p(\boldsymbol{\theta}_k \mid M_k) p(\mathbf{u}_k \mid \boldsymbol{\theta}_k, M_k)$ .<sup>22</sup> For clarity of what follows, let  $f_k(\boldsymbol{\xi}_k) = p(\boldsymbol{\xi}_k \mid M_k)$ . The implied prior on  $\boldsymbol{\psi}$  under model  $M_k$  is then given by

$$p(\boldsymbol{\psi} \mid M_k) = f_k(g_k(\boldsymbol{\psi})) \left| \frac{\partial g_k(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \right|, \quad (4.11)$$

where  $\left| \frac{\partial g_k(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \right|$  denotes the Jacobian determinant of the transformation. The Gibbs sampler can then be implemented by alternating between 1) drawing  $\boldsymbol{\psi}$  from the full-conditional distribution  $p(\boldsymbol{\psi} \mid M, \mathbf{y})$  and 2) drawing  $M$  from the full-conditional distribution  $p(M \mid \boldsymbol{\psi}, \mathbf{y})$ . Drawing  $\boldsymbol{\psi}$  from  $p(\boldsymbol{\psi} \mid M, \mathbf{y})$  is accomplished as follows: one first draws  $\boldsymbol{\theta}_k$  from the model-specific posterior  $p(\boldsymbol{\theta}_k \mid M_k, \mathbf{y})$ , then samples  $\mathbf{u}_k$  from  $p(\mathbf{u}_k \mid \boldsymbol{\theta}_k, M_k)$ , sets  $\boldsymbol{\xi}_k = (\boldsymbol{\theta}_k, \mathbf{u}_k)$ , and then computes  $\boldsymbol{\psi} = g_k^{-1}(\boldsymbol{\xi}_k)$ . This means that one can conveniently post-process previously obtained model-specific posterior samples since a sample from  $p(\boldsymbol{\theta}_k \mid M_k, \mathbf{y})$  can be obtained by selecting randomly a draw from stored model-specific MCMC output. The full-conditional distribution for the model indicator  $M$  is a categorical distribution, where  $M_k$  is sampled with probability

$$p(M_k \mid \boldsymbol{\psi}, \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\psi}, M_k) p(\boldsymbol{\psi} \mid M_k) p(M_k)}{\sum_{j=1}^K p(\mathbf{y} \mid \boldsymbol{\psi}, M_j) p(\boldsymbol{\psi} \mid M_j) p(M_j)}. \quad (4.12)$$

We used the marginalized version of the Gibbs sampler described in section 2.3 of Barker and Link (2013). This marginalized version estimates the transition matrix  $\boldsymbol{\Phi} = (\{\phi_{ij}\})$ , where  $\phi_{ij} = p(M^{(b+1)} = M_j \mid M^{(b)} = M_i)$  and  $M^{(b)}$  denotes the sampled value for  $M$  at iteration  $b$  of the Gibbs sampler. The marginalized version does not require one to draw  $M$ ; instead, one estimates  $\boldsymbol{\Phi}$  directly, one row at a time. The  $i$ th row of  $\boldsymbol{\Phi}$  is estimated by repeatedly 1) drawing  $\boldsymbol{\psi}$  given model  $M_i$  from  $p(\boldsymbol{\psi} \mid M_i, \mathbf{y})$  and 2) using the drawn  $\boldsymbol{\psi}$  to compute  $p(M_j \mid \boldsymbol{\psi}, \mathbf{y})$ ,  $j = 1, 2, \dots, K$ . A Rao-Blackwellized estimate of the  $i$ th row of  $\boldsymbol{\Phi}$  is then given by the average of the vector  $(p(M_1 \mid \boldsymbol{\psi}, \mathbf{y}), p(M_2 \mid \boldsymbol{\psi}, \mathbf{y}), \dots, p(M_K \mid \boldsymbol{\psi}, \mathbf{y}))$  across draws from  $p(\boldsymbol{\psi} \mid M_i, \mathbf{y})$ . This process is repeated for all models  $M_i$ ,  $i = 1, 2, \dots, K$  to obtain an estimate of all rows of the transition matrix  $\boldsymbol{\Phi}$ . An estimate of the posterior model probabilities is then obtained by normalizing the left eigenvector of the estimated transition matrix corresponding to the eigenvalue 1. An advantage of this marginalized version is that instead of sampling models according to their posterior model probabilities, one can fix the number of samples for each model.

We applied this marginalized Gibbs sampler RJMCMC version to our single-participant example. The dimensionality of  $\boldsymbol{\psi}$  was equal to the number of parameters of the full model. Under the full model, we simply set  $\boldsymbol{\psi} = \boldsymbol{\theta}_{\text{full}}$ . Under the null model, there was one parameter less since  $v_{\text{true}}$  was fixed. Hence, the dimensionality of the auxiliary variable vector  $\mathbf{u}_k = \mathbf{u}$  was one for the null model

---

<sup>22</sup>Typically, the distribution of the auxiliary variable vector  $\mathbf{u}_k$  is assumed to be conditionally independent of  $\boldsymbol{\theta}_k$  so that  $p(\mathbf{u}_k \mid \boldsymbol{\theta}_k, M_k) = p(\mathbf{u}_k \mid M_k)$ .

and we set  $\boldsymbol{\psi} = (\boldsymbol{\theta}_{\text{null}}, u)$ . The auxiliary variable  $u$  was proposed from a distribution constructed based on a logspline fit (Kooperberg, 2016) to the posterior samples for  $v_{\text{true}}$  under the full model. Therefore, to relate the palette vector  $\boldsymbol{\psi}$  to the model parameters (and the auxiliary variable for the null model), we used the identity mapping for both models (i.e.,  $g_k$  was the identity function for both models); consequently, the Jacobian determinants of the transformations were equal to one.



# Bayesian Inference for Multidimensional Scaling Representations with Psychologically-Interpretable Metrics

---

## Abstract

Multidimensional scaling (MDS) models represent stimuli as points in a space consisting of a number of psychological dimensions, such that the distance between pairs of points corresponds to the dissimilarity between the stimuli. Two fundamental challenges in inferring MDS representations from data involve inferring the appropriate number of dimensions, and the metric structure of the space used to measure distance. We approach both challenges as Bayesian model-selection problems. Treating MDS as a generative model, we define priors needed for model identifiability under metrics corresponding to psychologically separable and psychologically integral stimulus domains. We then apply a differential evolution Markov-chain Monte Carlo (DE-MCMC) method for parameter inference, and a Warp-III method for model selection. We apply these methods to five previous data sets, which collectively test the ability of the methods to infer an appropriate dimensionality and to infer whether stimuli are psychologically separable or integral. We demonstrate that our methods produce sensible results, but note a number of remaining technical challenges that need to be solved before the method can easily and generally be applied. We also note the theoretical promise of the generative modeling perspective, discussing new and extended models of MDS representation that could be developed.

---

This chapter is published as Gronau, Q. F., & Lee, M. D. (2020). Bayesian inference for multidimensional scaling representations with psychologically interpretable metrics. *Computational Brain & Behavior*, 3, 322–340. doi: <https://doi.org/10.1007/s42113-020-00082-y>. Also available as *PsyArXiv preprint*: <https://psyarxiv.com/5zmep/>

## 5.1 Introduction

Multidimensional scaling (MDS) was developed in the 1950s in cognitive psychology as a statistical method for making inferences about human mental representations (Kruskal, 1964; Shepard, 1957, 1962). MDS models the similarities or psychological proximities between pairs of stimuli, representing each stimulus as a point in a multidimensional space, such that more similar stimuli are nearer each other. The core psychological motivation is that the similarities reflect the basic cognitive process of generalization. Generalization can be thought of as the ability to treat two stimuli as being the same, and has been argued to serve as a basis for the mental organization of knowledge, and the capability of the mind to make adaptive predictions about properties and consequences (Shepard, 1987). For these reasons, mental representations found via MDS methods have been and remain widely used in cognitive process models of identification, categorization, and decision making (e.g., Nosofsky, 1992).

Soon after its development in cognitive psychology, however, MDS algorithms found application as a statistical method that produces a low-dimensional representation of a set of objects, based on a measure of the similarities between them. As a data reduction or visualization method, MDS has been applied in the natural, biological, and human sciences, with application areas as diverse as representing the similarities of skulls in archaeology, the tastes of colas in marketing, and the voting patterns of senators in politics (e.g., Borg & Groenen, 1997; Cox & Cox, 1994; Schiffman, Reynolds, & Young, 1981).

Whether viewed as a model of psychological representation or a data-reduction method, a foundational challenge in MDS modeling is determining the dimensionality  $M$  of the representational space. In his 1974 Presidential Address to the Psychometric Society, Roger Shepard identified six basic challenges for MDS, the third of which was: “The problem of determining the proper number of dimensions for the coordinate embedding space” (Shepard, 1974, p. 377). A number of methods for solving the problem of MDS dimensionality have been developed in both statistics and psychology. The most common approach is a scree test that aims to identify an “elbow” in the goodness-of-fit as dimensionality increases (Cox & Cox, 1994; Kruskal, 1964; Schiffman et al., 1981). Steyvers (2006) suggests the use of cross-validation methods, although this approach does not seem to be widely used.

Since choosing the correct dimensionality of an MDS is naturally regarded as a model-selection problem – that is, choosing between a one-dimensional versus two-dimensional versus three-dimensional representation, and so on – the statistically principled approach offered by Bayes factors should provide a solution (Kass & Raftery, 1995). Along these lines, Lee (2001) implements an approach based on the Bayesian Information Criterion (BIC). The difference between BIC values for representations with different dimensionality provides a crude approximation to the Bayes factor. Oh and Raftery (2001) provide a different approach to approximation by computing the marginal likelihoods of different representations using plug-in point estimates for the stimulus locations. This is an approximation because the exact Bayes factor requires an integration across the stimulus location parameters. Oh (2012) develops a method based on spike-and-slab priors, in which

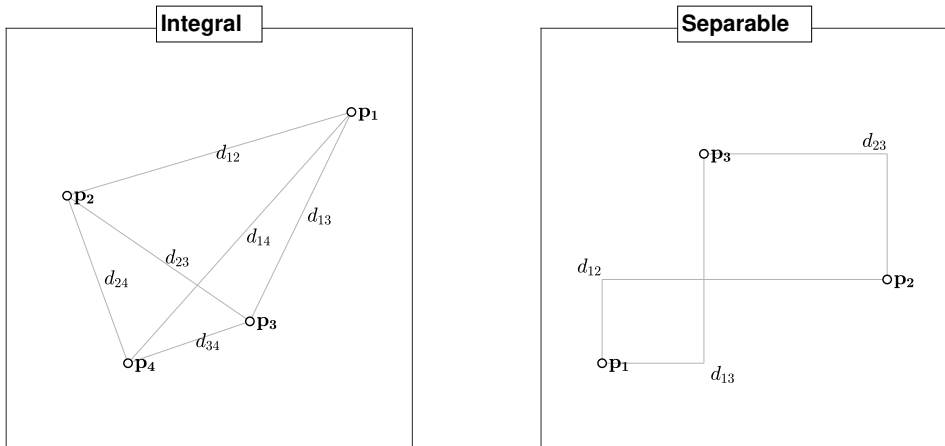


Figure 5.1: MDS representations with integral and separable metric structures.

the dimensionality is determined by the marginal posterior probabilities for each dimension that the coordinate locations are not zero for all stimuli.

From the perspective of MDS as psychological models however, none of these approaches qualify as being principled and complete. The key issue is that the theory of mental representation developed by Shepard (1957, 1987, 1991) emphasizes the role that the metric structure of the space plays in capturing key psychological properties of the stimuli. In particular, the idea is that different metrics capture the theoretical and empirical distinction between separable and integral stimuli (Attneave, 1950; Garner, 1974). Separable stimuli are those for which the component dimensions can be attended to separately. An example is different shapes of different sizes, since it is possible for people to attend selectively to either shape or the size. Integral stimuli, by contrast, are those for which the component dimensions cannot be attended to independently. The standard example is color, since it is typically not possible for people to attend selectively to the underlying hue, saturation, and brightness components.

Figure 5.1 shows how different metric structures are used to represent integral and separable stimuli. In the left panel, there are four stimuli, represented by the points  $\mathbf{p}_1, \dots, \mathbf{p}_4$ . The pairwise distances between these points, such as  $d_{12}$  between the first point and the second point, are modeled using the Euclidean metric, and so correspond to standard straight lines. In the right panel, there are three stimuli, and the pairwise distances between them are modeled according to the city-block metric. Intuitively, this corresponds to comparing the stimuli on each underlying dimension independently, then adding those dissimilarities to get an overall measure of dissimilarity.

Admittedly, this account of integrality and separability is a theoretical and empirical caricature, and much more nuanced and detailed accounts are possible (Shepard, 1991; Tversky & Gati, 1982). The point is that psychological representations based on MDS need to make assumptions about the metric structure of

the space, and use metrics other than the Euclidean metric. As Jäkel, Schölkopf, and Wichmann (2008, p. 2) point out, from the origins of MDS as a psychological model “There was no a priori reason to believe that mental representations should be Euclidean.” Previous methods for determining the dimensionality of MDS representations using Bayesian model selection, however, have either been insensitive to the metric structure of the representation (Lee, 2001), or have focused on the Euclidean metric (Oh, 2012; Oh & Raftery, 2001).

The use of non-Euclidean metrics raises another challenge, related to inferring MDS representations themselves. There is evidence that it can be computationally difficult to find multidimensional city-block MDS representations (Groenen, Heiser, & Meulman, 1998; Hubert, Arabie, & Hesson-McInnis, 1992), as well as finding unidimensional MDS representations (Mair & Leeuw, 2014). Given that these difficulties stem from basic geometric properties of the MDS representations, it seems likely they will continue to present an issue for Bayesian methods of inference.

Finally, there is the challenge of inferring the appropriate metric structure for an MDS representation. Shepard (1991) reviews the original statistical approach to this problem, which involved applying non-metric MDS algorithms for a large number of different metrics, and choosing the one with the best goodness-of-fit. As Lee (2008) pointed out, this approach neglects to account for the component of model complexity that arises from the functional form of parameter interaction (Pitt, Kim, Navarro, & Myung, 2006), which is often the only difference between MDS models using different psychologically-interpretable metrics. Lee (2008) developed a Bayesian approach in which the possible metrics correspond to a parameter that is inferred jointly with the coordinate location parameters that represent the stimuli. Okada and Shigemasu (2010) developed and tested this approach further, and showed it is capable of recovering the correct metric in simulation studies. Both the Lee (2008) and Okada and Shigemasu (2010) methods, however, failed to resolve basic challenges in model identifiability that arise from treating the choice of metric structure as a parameter inference problem. It is possible these identifiability issues could be addressed by considering the choice as a model-selection problem, and restricting the set of possibilities to a few interpretable metrics.

Accordingly, the goals of this chapter are to examine the implementation of MDS models that use psychologically-interpretable metrics, including both the Euclidean and a non-Euclidean metric, and explore the possibility of inferring the appropriate dimensionality and metric structure of these representations using Bayesian model-selection methods. The structure of the remainder of the chapter is as follows. In the next section, we define MDS models, and address the issue of model identifiability under different metrics. Consistent with previous literature, we argue that the city-block metric presents fundamental problems in making MDS representations identifiable. This leads to the development of joint prior distributions on the stimulus location parameters for the Euclidean metric, and non-Euclidean metrics other than the city-block metric. With these priors established, we apply an approach to Bayesian inference using differential evolution Markov-chain Monte Carlo (DE-MCMC) computational sampling methods. The DE-MCMC method helps address the difficulties inherent in inferring MDS repre-

sentations, which are especially evident in non-Euclidean cases. We then use the Warp-III bridge sampling method to approximate the marginal densities needed to determine Bayes factors. We apply the method to five previously studied data sets, differing in the type of stimuli and expected dimensionality of their MDS representation. For all five applications, the method makes sensible inferences about dimensionality, and produces interpretable stimulus representations. We conclude with a discussion of remaining statistical and computational challenges, and potential directions for refining and extending the approach.

## 5.2 MDS Model Identifiability

### 5.2.1 The Identifiability Problem

Formally, suppose there are  $N$  stimuli to be represented, based on observed proximity data from  $P$  participants, with  $d_{ijk}$  measuring the proximity between the  $i$ th and  $j$ th stimulus provided by the  $k$ th participant. We assume these observed proximities are normalized to lie between 0 and 1. The point representing the  $i$ th stimulus in a  $M$ -dimensional space is  $\mathbf{p}_i = (p_{i1}, \dots, p_{iM})$  and the distance between points  $\mathbf{p}_i$  and  $\mathbf{p}_j$  is measured by the Minkowski metric with metric parameter  $r$ , so that

$$\hat{d}_{ijk} = \left( \sum_{m=1}^M |p_{im} - p_{jm}|^r \right)^{1/r}. \quad (5.1)$$

The Minkowski metric has special cases of the city-block metric when  $r = 1$  and the Euclidean metric when  $r = 2$ . Values of  $r$  between 1 and 2 can potentially be interpreted as intermediate assumptions about the independence of stimulus dimensions between the end-point of complete separability and complete integrality.

The goal of MDS is for the modeled distances  $\hat{d}_{ijk}$  to correspond to the observed proximities  $d_{ijk}$ . We use the probabilistic model

$$d_{ijk} \sim \text{Gaussian} \left( \hat{d}_{ijk}, \frac{1}{\sigma^2} \right), \quad (5.2)$$

where  $\sigma$  is the standard deviation with which the observed proximities are measured.<sup>1</sup> It is assumed to be the same for all of the proximities, and is given a prior

$$\sigma \sim \text{TruncatedGaussian} \left( 0.15, \frac{1}{0.2^2} \right) T(0, \infty), \quad (5.3)$$

where the  $T(0, \infty)$  indicates the sampled value is truncated to be a positive real number. This is an informative prior (Lee & Vanpaemel, 2018), consistent with previous data and modeling. Intuitively,  $\sigma$  corresponds to the average standard deviation of different individual ratings of the same pair of stimuli. Empirical estimates of this standard deviation in previous data tend to range from about 0.1

---

<sup>1</sup>We parameterize the Gaussian distribution in terms of mean and precision parameters, for consistency with the JAGS graphical modeling language.

to about 0.2 (Lee, 2001; Lee & Pope, 2003).<sup>2</sup> Accordingly, the prior is centered on 0.15, but allows a wide range of possibilities.

We note that this MDS model does not incorporate individual differences. It is assumed that the same point  $\mathbf{p}_i$  represents the  $i$ th stimulus for all participants. We also emphasize, however, that individual-level proximity data  $d_{ijk}$  are modeled, rather than averaged or aggregated data across participants. The problems inherent in averaging data have long been understood (Estes, 1956), and have been studied in the specific cognitive modeling context provided by MDS representations (Lee & Pope, 2003). Our approach is to require the same underlying MDS representation to provide an account of each individual proximity matrix.

To complete the generative model, a straightforward approach would be to give all of the coordinate locations for the representational points uniform priors  $p_{im} \sim \text{Uniform}(-1, 1)$ . These priors, however, made the model non-identifiable, because the distances between points are invariant under transformations (Borg & Groenen, 1997, Ch. 2). The distances between points are preserved under translation, reflection, axes permutation (for non-Euclidean metrics), and rotation (for the Euclidean metric). A principled Bayesian approach for controlling these invariances to ensure model identifiability constrains the coordinate location parameters through a joint prior distribution that depends on the assumed metric.

### 5.2.2 Previous Approaches

Existing MDS modeling methods that use Bayesian inference almost always rely on post-processing to address the issue of identifiability. The method developed by Lee (2008) post-processes posterior samples of the coordinate location parameters to control for translation, reflection, and permutation. For example, to control for translation, the method zero centers every posterior sample of the sets of coordinate location. The Lee (2008) method does not control for rotation, which is problematic, because the method also attempts to infer the  $r$  metric parameter, and so the inferred representational space can have a Euclidean metric, which requires rotational invariance.

Most other methods, in contrast, assume the MDS space is Euclidean. The post-processing of the coordinate location parameters used by both Oh and Raftery (2001) and Oh (2012) assumes a Euclidean space and controls for translation, reflection, and rotation. Okada and Mayekawa (2018) extend the approach developed by Okada (2012), which relies on Procrustes analysis. Post-processing uses a loss function to align posterior samples of the coordinate location, but again assumes a Euclidean space.

Besides the lack of flexibility in the nature of the distance metric, post-processing methods have the effect of implementing modeling assumptions without explicitly specifying those assumptions as part of the model. While this is often practical, it is theoretically inelegant, and contrary to the goals of generative modeling. Ideally, the constraints required for model identifiability should be part of the model itself. In the case of MDS models, these constraints are naturally imposed through the specification of a joint prior over the coordinate location

---

<sup>2</sup>See also the data repository at <https://osf.io/ey9vp/>

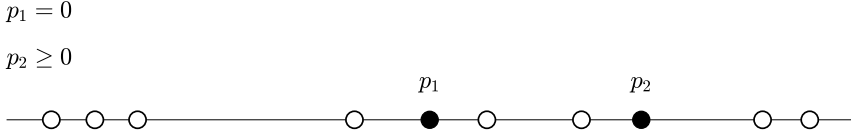


Figure 5.2: Identification constraints for a one-dimensional representation.

parameters that addresses the transformational invariances, removes the need for post-processing, and makes bridge sampling feasible.

This generative approach is used by the “parameter fixing” method considered by Okada and Mayekawa (2018), who evaluate it as a contrast with the Procrustes methods that are their focus. Parameter fixing corresponds to setting a structured joint prior over the coordinate location parameters. Okada and Mayekawa (2018) define the appropriate prior for a Euclidean space using results provided by Bakker and Poole (2013), which were derived using an analytic method based on matrix properties.

Our goal is to extend this approach to include non-Euclidean representations. We start by considering one-dimensional MDS representations, before considering multidimensional representations in both Euclidean and non-Euclidean metric spaces. We take a geometric approach to identifying the required joint priors for invariance constraints, complementing the non-geometric approach of Bakker and Poole (2013) for the Euclidean metric.

### 5.2.3 One-dimensional Representation

For a one-dimensional representation, all of the psychologically-interpretable metrics we consider give the same distances. The required constraints on the points are shown in Figure 5.2, with one point fixed at the origin to control translation, and second point restricted to be positive to control reflection.

These constraints can be formalized by a joint prior with

$$\begin{aligned} p_1 &= 0 \\ p_2 &\sim \text{Uniform}(0, 1) \\ p_3, \dots, p_N &\sim \text{Uniform}(-1, 1). \end{aligned} \tag{5.4}$$

### 5.2.4 Euclidean Multidimensional Representations

Figure 5.3 shows the constraints needed to identify Euclidean MDS representations in two and three dimensions. In the two-dimensional case, the first point  $\mathbf{p}_1$  is fixed at the origin, to control translation, the second point  $\mathbf{p}_2$  is constrained to the positive  $x$ -axis, to control reflection in the  $y$ -axis and rotation, and the third point  $\mathbf{p}_3$  is constrained to have a positive  $y$ -value to control reflection in the  $x$ -axis. The same logic is applied in the three-dimensional case, with  $\mathbf{p}_1$  controlling translation,  $\mathbf{p}_2$  and  $\mathbf{p}_3$  controlling reflection and rotation in successive axes, and  $\mathbf{p}_4$  controlling the final reflection.

## 5. BAYESIAN INFERENCE FOR MULTIDIMENSIONAL SCALING REPRESENTATIONS WITH PSYCHOLOGICALLY-INTERPRETABLE METRICS

---

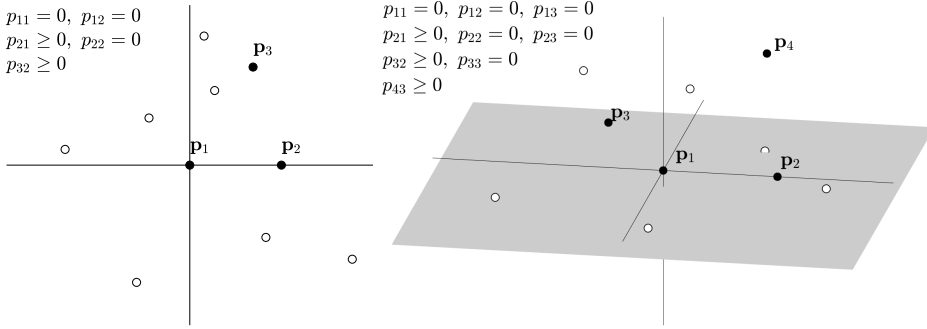


Figure 5.3: Identification constraints for Euclidean representations in two dimensions (left) and three dimensions (right).

These are the first two cases of a general pattern, clear by induction, that applies to a  $M$ -dimensional representation, and corresponds to the matrix result provided by Bakker and Poole (2013). An intuitive presentation of the inductive pattern is shown below, where “0” denotes fixing a coordinate location to zero, “+” denotes constraining it to be positive, and “ $\mathcal{R}$ ” denotes imposing no constraint.

	1	2	3	4		$D$
	Dim	Dim	Dim	Dim		Dim
$p_1$	0	0	0	0	...	0
$p_2$	+	0	0	0	...	0
$p_3$	$\mathcal{R}$	+	0	0	...	0
$p_4$	$\mathcal{R}$	$\mathcal{R}$	+	0	...	0
$p_5$	$\mathcal{R}$	$\mathcal{R}$	$\mathcal{R}$	+	...	0

Formally, these constraints in  $D$  dimensions correspond to the joint prior

$$\begin{aligned}
 p_{11}, \dots, p_{1D} &= 0 \\
 p_{21} &\sim \text{Uniform}(0, 1) \\
 p_{22}, \dots, p_{2D} &= 0 \\
 p_{31} &\sim \text{Uniform}(-1, 1) \\
 p_{32} &\sim \text{Uniform}(0, 1) \\
 p_{33}, \dots, p_{3D} &= 0 \\
 p_{41}, p_{42} &\sim \text{Uniform}(-1, 1) \\
 p_{43} &\sim \text{Uniform}(0, 1) \\
 p_{44}, \dots, p_{4D} &= 0 \\
 &\dots
 \end{aligned} \tag{5.5}$$

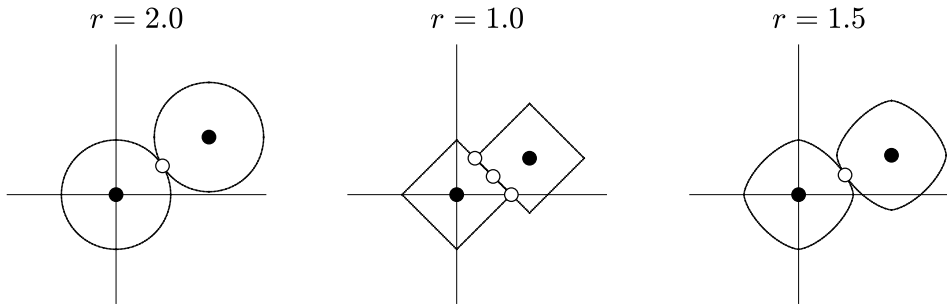


Figure 5.4: The nature of iso-distance curves and the identifiability of mid-points for the three Minkowski metrics corresponding to  $r = 2$  (Euclidean),  $r = 1$  (city-block), and  $r = 1.5$ .

### 5.2.5 Non-Euclidean Multidimensional Representations

Finding constraints for invariance in non-Euclidean metrics is more complicated, and is especially difficult for the city-block metric. The basic geometric problem was noted as early as Arnold (1971), and discussed in Shepard's (1974, Figure 10) presidential address. A simple demonstration of the fundamental problem is provided by Figure 5.4. The three panels correspond to Euclidean ( $r = 2$ ), city-block ( $r = 1$ ), and a general non-Euclidean ( $r = 1.5$ ) metric, and show unit iso-distance contours around the same two points in each metric, shown as black dots. These iso-distance contours are the “unit circles” of each metric, showing all the points in the space that are the same distance from the two points. For the Euclidean metric, these contours are familiar circles, and coincide at only one point, shown by the white dot. This means that there is a unique point in the space that is equally distant from the two points shown by black dots. In the context of an MDS representation, a stimulus that is equally different to both of the points can be uniquely identified.

For the city-block case, however, the iso-distance contours are diamonds, and there are infinitely many points that are equally different. Three specific possibilities are shown by white dots, but clearly any point along the line where the iso-distance contours coincide is possible. In the context of an MDS representation, this means that there is a fundamental difficulty in identifying a stimulus that is equally different to both of the points. This basic problem is not, in general, solved by the introduction of additional stimuli that provide additional constraints. Indeed, the problem compounds for potential city-block representations with many stimuli. Bortz (1974, see, especially, Figures 2 and 3) provides compelling examples, and the same point is emphasized in the seminal text by Borg and Groenen (1997, pp. 369–372).

Figure 5.5 provides a concrete example, based on the more general configuration examined by Borg and Groenen (1997, Figure 17.6). Each panel shows a representation of six fictitious people in terms of two underlying dimensions. The city-block distance between each pair of people is identical in both configu-

## 5. BAYESIAN INFERENCE FOR MULTIDIMENSIONAL SCALING REPRESENTATIONS WITH PSYCHOLOGICALLY-INTERPRETABLE METRICS

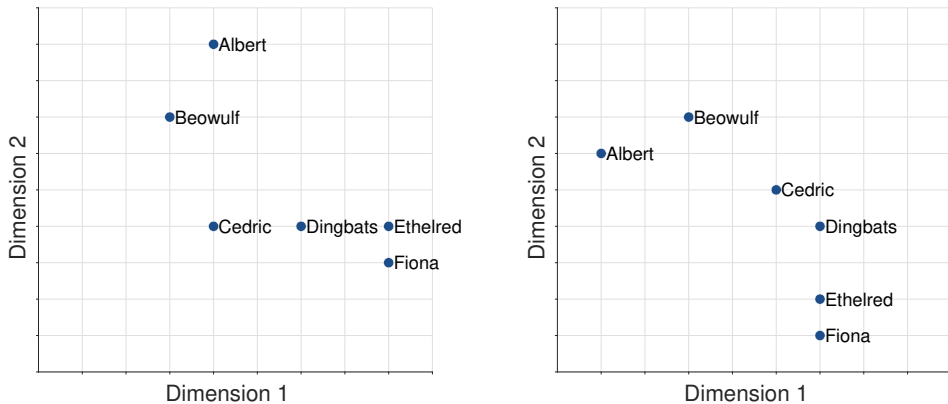


Figure 5.5: Two city-block representations of six fictitious people in terms of two dimensions. Both representations have identical proximity matrices.

rations. This means, of course, that this proximity matrix is equally consistent with both representations, and either could be inferred from the data. But, the two representations are substantively different, in non-trivial ways. The representations do not differ simply by changing the axes, and have basic structural differences. For example: Cedric, Dingbats, and Ethelred are co-linear in the first representation, but not in the second, where Dingbats, Ethelred and Fiona become co-linear; the ordering of Albert and Beowulf changes on both dimensions between the configurations; and so on. In fact, once the lack of invariance revealed by the Borg and Groenen (1997, Figure 17.6) analysis is understood, it is clear that many additional representations for the proximity between the six people could be constructed, supporting a wide range of different meaningful interpretations.

A practical approach for identifying city-block representations, used by Nosofsky (1985), relies on determining the values of some stimuli on some dimensions, by means external to the MDS modeling. Ultimately, this strategy can solve the problem, if it is possible to find the values of every stimulus on every dimension. But, Figure 5.5 suggests the strategy may not be effective in situations where the identification of just a few stimuli is possible. In both representations, Dingbats is at the same location, consistent with values on dimensions having been externally determined, yet the locations of the remaining stimuli are under-determined. In addition, if, for example, Albert was additionally identified as being located in the position shown in the first representation, that would constrain the inference about Beowulf and Cedric, but would not constrain Ethelred and Fiona, who could still be inferred to be at either of the possibilities shown in the two representations. Thus, while the addition of stimuli, or the identification of dimension values for some stimuli, may work in some specific circumstances, we do not believe either represents a general approach to making city-block MDS representations identifiable.

We do not know how to solve the problem of MDS model invariance for the

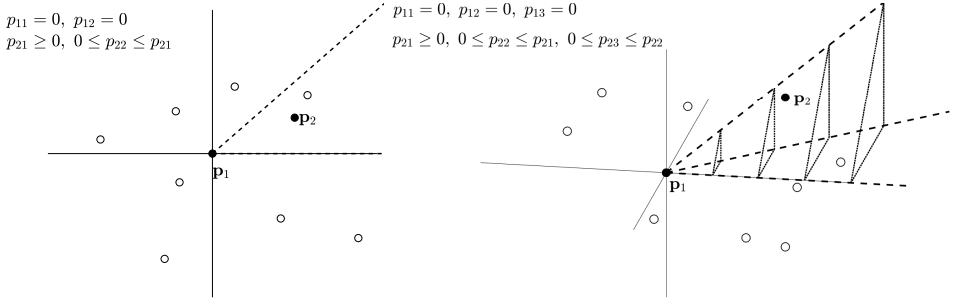


Figure 5.6: Identification constraints for non-Euclidean representations in two dimensions (left) and three dimensions (right).

city-block metric. As the right-most panel of Figure 5.4 makes clear, however, the problem does not occur for Minkowski-metric parameters  $r > 1$ . For the  $r = 1.5$  metric, the iso-distance contours again coincide at only one point. The asymmetry of these contours makes clear they do not have the rotational invariance of the Euclidean  $r = 2$  metric. In this way, general non-Euclidean metrics, such as  $r = 1.5$ , capture the psychological idea that the dimensions in an MDS representation have meaning and allow selective attention, while avoiding the degenerate lack of identifiability inherent in the city-block metric.

Figure 5.6 shows the constraints needed to identify these sort of non-Euclidean MDS representations in two and three dimensions. In the two-dimensional case, the first point  $\mathbf{p}_1$  is once again fixed at the origin, to control translation, the second point  $\mathbf{p}_2$  is constrained to the positive quadrant to control reflection. In addition, the constraint that  $p_{22} \leq p_{21}$  is imposed, requiring the value of the second stimulus on the  $y$ -axis not to be larger than its value on the  $x$ -axis. This constraint controls for axis permutation, preventing the two dimensions from being swapped, and so allocates a specific underlying stimulus dimension to each axis. The three-dimensional case extends this logic by requiring that the  $z$ -axis value of the second point be positive, to prevent reflection, and be less than the value of the second point on the  $y$ -axis, to prevent permutation.

These first two cases once again make clear a general pattern, in which the coordinate values of the second point are positive and order constrained.<sup>3</sup> Formally, the constraints for non-city-block but non-Euclidean  $D$  dimensions are

$$\begin{aligned}
 p_{11}, \dots, p_{1D} &= 0 \\
 p_{21}, \dots, p_{2D} &\sim \text{Uniform}(0, 1) : \quad p_{21} \geq \dots \geq p_{2D} \\
 p_{31}, \dots, p_{3D} &\sim \text{Uniform}(-1, 1) \\
 &\dots
 \end{aligned} \tag{5.6}$$

<sup>3</sup>These order constraints can be imposed either in decreasing manner, as shown in Figure 5.6 for easier visualization, or in an increasing manner, as they are in our code.

### 5.3 Bayesian MDS Inference via DE-MCMC

When posterior samples for MDS models are obtained using conventional Markov-chain Monte Carlo algorithms (MCMC; e.g., Gamerman & Lopes, 2006) it can occur that chains get stuck in local maxima. In our experience, the reason is typically that the stimuli that are constrained are similar to each other. To prevent local maxima, we implemented a heuristic to order the stimuli in a way that those defining the constraints are dissimilar. We motivate and describe this heuristic in detail in Appendix A. In addition, to improve sampling, we used the differential evolution Markov-chain Monte Carlo algorithm (DE-MCMC; e.g., Heathcote et al., 2018; Turner et al., 2013) that helps to guide the chains to regions of high posterior density.

DE-MCMC is a population-based MCMC algorithm that generates efficient proposals via a population of interacting chains (Turner et al., 2013). One strength of the algorithm is that it works well for highly correlated target distributions. However, we used DE-MCMC primarily for the reason that the interacting chains can guide each other to regions of high posterior density which helps to avoid the issue of chains getting stuck in local maxima. Specifically, during burn-in, we used a *migration step* that remedies the problem of outlier chains in an effective manner (for details, see Turner et al., 2013, Appendix B). We found that the combination of the ordering heuristic and DE-MCMC provides effective sampling consistently for the Euclidean metric, and is partially effective for non-Euclidean metrics.

### 5.4 Bayesian Model Comparison via Bridge Sampling

#### 5.4.1 Marginal Likelihood

Comparing MDS models with different dimensions and metrics via Bayes factors and posterior model probabilities requires the computation of the marginal likelihood for all of the models,  $\mathcal{M}_{m,r}$ , being considered where  $m$  denotes the dimensionality and  $r$  the metric. Let  $\mathbf{D}$  denote the observed data (i.e., the pair-wise dissimilarity ratings  $d_{ijk}$ ) and  $\mathbf{P}$  denote the  $N \times m$  matrix with the latent stimulus coordinates for each stimulus. The marginal likelihood for model  $\mathcal{M}_{m,r}$  corresponds to the normalizing constant of the joint posterior distribution for  $\theta = (\mathbf{P}, \sigma)$ :

$$\begin{aligned}
 p(\mathbf{D} \mid \mathcal{M}_{m,r}) &= \int q(\theta \mid \mathbf{D}, \mathcal{M}_{m,r}) d\theta \\
 &= \int \int \underbrace{p(\mathbf{D} \mid \mathbf{P}, \sigma, \mathcal{M}_{m,r})}_{\text{Likelihood}} \underbrace{p(\mathbf{P} \mid \mathcal{M}_{m,r})}_{\text{Joint Prior on Stimulus Locations}} \underbrace{p(\sigma \mid \mathcal{M}_{m,r})}_{\text{Prior on Imprecision}} d\mathbf{P} d\sigma,
 \end{aligned} \tag{5.7}$$

where  $q(\theta \mid \mathbf{D}, \mathcal{M}_{m,r})$  denotes the unnormalized joint posterior density.

### 5.4.2 Bridge Sampling

Since the marginal likelihood in Equation 5.7 is not available analytically, we use Warp-III bridge sampling (Meng & Schilling, 2002) to estimate this potentially high-dimensional integral. Bridge sampling (Meng & Wong, 1996; for a recent tutorial see Gronau, Sarafoglou, et al., 2017) is based on the following identity:

$$p(\mathbf{D} \mid \mathcal{M}_{m,r}) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})} [h(\boldsymbol{\theta}) q(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r})]}{\mathbb{E}_{p(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r})} [h(\boldsymbol{\theta}) g(\boldsymbol{\theta})]}, \quad (5.8)$$

where the numerator is an expected value with respect to a proposal distribution  $g(\boldsymbol{\theta})$ , the denominator is an expected value with respect to the parameter posterior distribution  $p(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r})$ , and  $h(\boldsymbol{\theta})$  is a function such that  $0 < \left| \int h(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \right| < \infty$ . The bridge sampling estimate is obtained by sampling from the proposal distribution  $g(\boldsymbol{\theta})$  and the posterior distribution  $p(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r})$  to approximate the two expected values. Meng and Wong (1996) showed that the optimal choice for  $h(\boldsymbol{\theta})$  is given by

$$h_o(\boldsymbol{\theta}) \propto [s_1 q(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r}) + s_2 p(\mathbf{D} \mid \mathcal{M}_{m,r}) g(\boldsymbol{\theta})]^{-1}, \quad (5.9)$$

where  $s_i = n_i / (n_1 + n_2)$ ,  $i \in \{1, 2\}$ ,  $n_1$  denotes the number of samples from the posterior  $p(\boldsymbol{\theta} \mid \mathbf{D}, \mathcal{M}_{m,r})$ , and  $n_2$  denotes the number of samples from the proposal  $g(\boldsymbol{\theta})$ . The optimal choice for  $h(\boldsymbol{\theta})$  depends on the marginal likelihood of interest. Therefore, in practice, the bridge sampling estimate is obtained via an iterative scheme, presented below, that updates an initial guess of the marginal likelihood until convergence.

The variability of the bridge sampling estimate is governed not only by the number of samples, but also, crucially, by the overlap between the proposal and the posterior distribution. To obtain estimates with low variability, it is therefore prudent to maximize the overlap between these two distributions. The Warp-III approach attempts to create a large overlap by fixing the proposal to a standard multivariate Gaussian distribution and then manipulating (i.e., “warping”) the posterior in a way that matches the first three moments of the two distributions.<sup>4</sup> Crucially, the warping procedure retains the normalizing constant of the posterior (i.e., the marginal likelihood of interest).

A prerequisite for the warping procedure is that all elements of the parameter vector are allowed to range across the entire real line. This can be achieved via a change-of-variables of the form  $\boldsymbol{\zeta} = f(\boldsymbol{\theta})$ , where  $f$  is a suitable<sup>5</sup> vector-valued function that transforms the constrained elements of  $\boldsymbol{\theta}$  so that all elements of  $\boldsymbol{\zeta}$  are unconstrained.<sup>6</sup> The Warp-III procedure is based on the following stochastic transformation of the unconstrained parameter vector  $\boldsymbol{\zeta}$ :

$$\boldsymbol{\eta} = b \mathbf{C}^{-1} (\boldsymbol{\zeta} - \boldsymbol{\mu}), \quad (5.10)$$

<sup>4</sup>Note that other proposal distributions are conceivable. The only constraints are that the proposal has a zero mean vector, an identity covariance matrix, and exhibits no skewness.

<sup>5</sup>The function  $f$  needs to be one-to-one and its inverse  $f^{-1}$  needs to have a well-defined Jacobian.

<sup>6</sup>We use a function  $f$  that applies a log transformation to  $\sigma$  and (scaled) probit transformations to the non-zero elements of  $\mathbf{P}$ . The transformation for the ordered coordinates of the second stimulus for the non-Euclidean case is described in Appendix B. Note that it is irrelevant whether the coordinates are ordered as decreasing, as shown in Figure 5.6 for easier visualization,

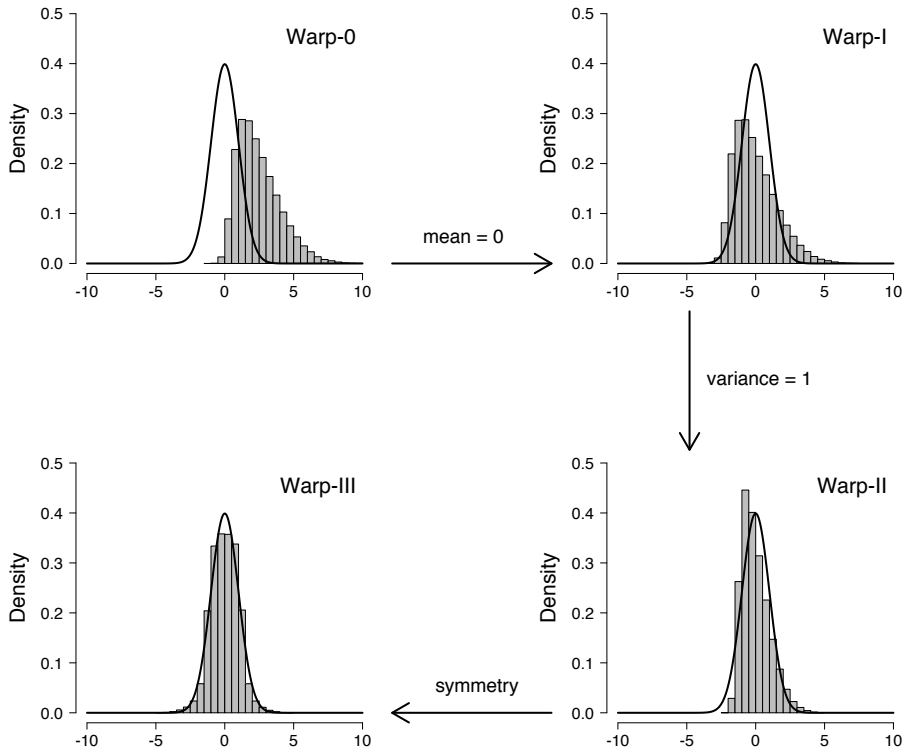


Figure 5.7: illustration of the Warp-III procedure. The black solid line shows the standard Gaussian proposal distribution and the gray histogram shows synthetic posterior samples. Available at <https://tinyurl.com/y7owvsz3> under CC license <https://creativecommons.org/licenses/by/2.0/>.

where  $b \sim \text{Bernoulli}(0.5)$  on  $\{-1, 1\}$ ,  $\boldsymbol{\mu}$  denotes the expected value vector of the posterior samples, and  $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^\top$  denotes the posterior covariance matrix (i.e.,  $\boldsymbol{C}$  is obtained via a Cholesky decomposition).

Figure 5.7 illustrates the warping approach for the univariate case. In the upper-left panel, the solid line corresponds to the standard Gaussian proposal distribution and the gray histogram depicts synthetic posterior samples. Subtracting the posterior mean from all samples matches the first moment of the proposal and the posterior distribution, as shown in the upper-right panel. Dividing all samples by the posterior standard deviation matches the second moment of the two distributions, as shown in the lower-right panel. Finally, attaching a minus sign

---

or increasing, as implemented in our code. The transformation described in the appendix assumes the latter. These transformations can be applied after having obtained posterior samples for  $\boldsymbol{\theta}$ . Furthermore, where necessary, the expressions are adjusted by the relevant Jacobian term  $|\det \mathcal{J}_{f^{-1}}(\boldsymbol{\zeta})|$ .

with probability 0.5 to the posterior samples achieves symmetry and thus matches the third moment of the proposal and the posterior distribution, as shown in the lower-left panel.

The Warp-III bridge sampling estimate based on  $h_o(\boldsymbol{\theta})$  is computed via an iterative scheme where the value of the estimate at iteration  $t$  is given by (for more details see Gronau, Wagenmakers, et al., 2019):

$$\hat{p}(\mathbf{D} \mid \mathcal{M}_{m,r})^{(t+1)} = \frac{\frac{1}{n_2} \sum_{i=1}^{n_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}(\mathbf{D} \mid \mathcal{M}_{m,r})^{(t)}}}{\frac{1}{n_1} \sum_{j=1}^{n_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}(\mathbf{D} \mid \mathcal{M}_{m,r})^{(t)}}}, \quad (5.11)$$

with

$$l_{1,j} = \frac{\frac{|\hat{\mathbf{C}}|}{2} [q(2\hat{\boldsymbol{\mu}} - \boldsymbol{\zeta}_j^* \mid \mathbf{D}, \mathcal{M}_{m,r}) + q(\boldsymbol{\zeta}_j^* \mid \mathbf{D}, \mathcal{M}_{m,r})]}{g(\hat{\mathbf{C}}^{-1}(\boldsymbol{\zeta}_j^* - \hat{\boldsymbol{\mu}}))}, \quad (5.12)$$

and

$$l_{2,i} = \frac{\frac{|\hat{\mathbf{C}}|}{2} [q(\hat{\boldsymbol{\mu}} - \hat{\mathbf{C}}\hat{\boldsymbol{\eta}}_i \mid \mathbf{D}, \mathcal{M}_{m,r}) + q(\hat{\boldsymbol{\mu}} + \hat{\mathbf{C}}\hat{\boldsymbol{\eta}}_i \mid \mathbf{D}, \mathcal{M}_{m,r})]}{g(\hat{\boldsymbol{\eta}}_i)}. \quad (5.13)$$

In Equations 5.12–5.13,  $q(\cdot \mid \mathbf{D}, \mathcal{M}_{m,r})$  denotes the unnormalized posterior density with respect to the unconstrained parameter vector  $\boldsymbol{\zeta}$ ,  $\{\boldsymbol{\zeta}_1^*, \boldsymbol{\zeta}_2^*, \dots, \boldsymbol{\zeta}_{n_1}^*\}$  denote  $n_1$  posterior samples, and  $\{\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2, \dots, \hat{\boldsymbol{\eta}}_{n_2}\}$  denote  $n_2$  samples from the standard multivariate Gaussian proposal distribution. To compute the Warp-III estimate one obtains  $2n_1$  posterior samples: the first half of these samples is used to approximate  $\boldsymbol{\mu}$  and  $\mathbf{C}$  with their sample versions  $\hat{\boldsymbol{\mu}}$  and  $\hat{\mathbf{C}}$ , the second half of the posterior samples is used in the iterative scheme (i.e., Equation 5.11). We use the `bridgesampling` R package (Gronau, Singmann, & Wagenmakers, 2020) to compute the bridge sampling estimate in Equation 5.11.

## 5.5 Applications

In this section, we present applications of our method to five existing data sets. For each application, we describe the stimuli and the nature of the data, as well as make clear our expectations about the MDS representation that will be inferred. In particular, we state our expectations about both the dimensionality and metric structure of the representation whenever possible. The results we present are based on considering MDS models up to and beyond this expected dimensionality, so that the inference our method makes is clear. Where possible, we apply our method under the assumption that the metric space is both Euclidean ( $r = 2$ ) and non-Euclidean ( $r = 1.5$ ) so that an inference can also be made about the integrality or separability of the stimulus domain. For some applications, we were unable to generate samples with acceptable convergence for the  $r = 1.5$  metric. In those cases, we only report results assuming the  $r = 2$  metric.

### 5.5.1 Line Length

Our first application involves the similarity judgments between nine lines of equally increasing length provided by 27 participants, as reported in Cohen, Nosofsky,

## 5. BAYESIAN INFERENCE FOR MULTIDIMENSIONAL SCALING REPRESENTATIONS WITH PSYCHOLOGICALLY-INTERPRETABLE METRICS

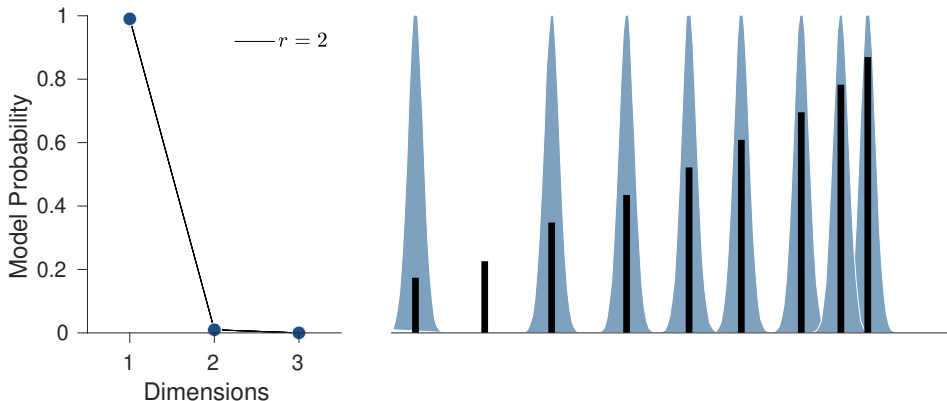


Figure 5.8: Results for line-length similarity data from Cohen et al. (2001). The left panel shows the posterior model probabilities for one- through three-dimensional MDS representations. The right panel shows the inferred one-dimensional representation with black lines showing the line stimuli at their inferred locations and blue histograms showing the marginal posterior distributions for these locations.

and Zaki (2001). We expect these stimuli to have a one-dimensional MDS representation, corresponding to line length. Because the Minkowski metrics are all equivalent in a one-dimensional space, we do not have any expectations about the metric structure. Thus, we applied our method to these data by assuming a Euclidean metric.<sup>7</sup> As for all of our applications, we used 15 chains and 500 burn-in samples. During burn-in, the probability of a migration step was set to 0.05. After burn-in, migration was switched off, and the algorithm was run for 9,000 iterations. We only retained every third sample so that we ended up with 3,000 samples per chain for further use (i.e., a total of 45,000 samples collapsed across chains).

The left panel of Figure 5.8 shows posterior model probabilities, assuming equal prior probabilities, for one-, two-, and three-dimensional MDS representations. To assess the stability of the posterior model probability estimates, we ran the Warp-III procedure five times based on new samples from the proposal distribution (we always used the same set of posterior samples). These five repetitions are drawn as separate lines but, in this case, the results are so similar that they are visually indistinguishable. Because of the assumptions of equal prior probabilities, the ratio of any pair of posterior probabilities is naturally interpreted as a Bayes factor. The key result is that the expected one-dimensional representation is inferred, with a posterior probability near one.

The right panel of Figure 5.8 shows the inferred one-dimensional MDS representation. The black lines show the stimuli in terms of their physical line lengths,

<sup>7</sup>We note, however, for completeness that we had difficulty with convergence using the  $r = 1.5$  metric for these data.

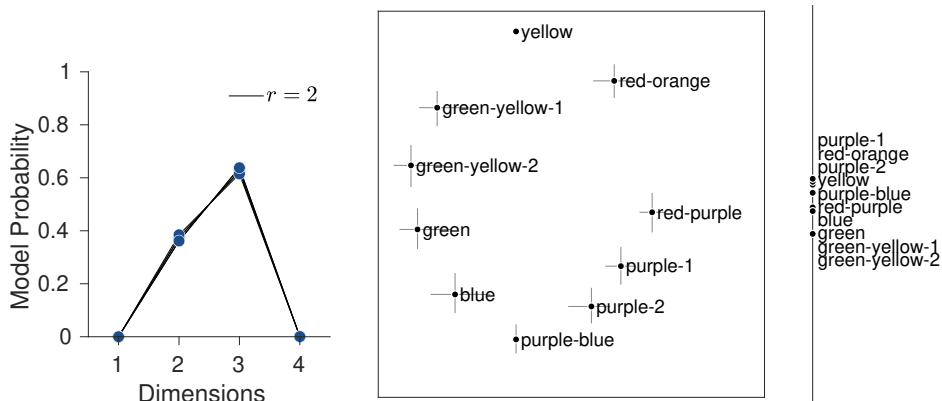


Figure 5.9: Results for color similarity data from color-normal subjects reported by Helm (1964). The left panel shows the posterior probabilities for one- through four-dimensional MDS representations. The right panel shows the inferred three-dimensional representation, with two dimensions shown as a two-dimensional plot in the center, and the third dimension shown along an axis to the right. Circular markers and labels show the inferred locations of each stimulus and error bars show 95% credible intervals for the marginal posterior distribution for each dimension.

located at the posterior mean of their location in the psychological space. The blue histograms show the marginal posterior distributions for each line stimulus. The MDS representation arranges the line stimuli in order of their length, but they are not evenly spaced, despite the lines increasing in constant physical increments. Instead, the psychological representation shows compression for the longer lines, consistent with basic psychophysics (Fechner, 1966 [1860]). This compression is large enough that the posterior distributions begin to overlap for the longest line stimuli.

### 5.5.2 Colors

Our second application considers classic data reported by Helm (1964), involving the similarities between ten colors. The experimental procedure involved trials in which participants were presented with physical tiles of three different colors, and moved one of the tiles to reflect their perceived overall similarity of the color of this tile to the colors of the other two tiles. Based on these responses, Helm (1964) calculated measures of pairwise similarities between the colors that have previously been considered in the MDS literature (e.g. Borg & Groenen, 1997; Carroll & Wish, 1974). We consider only the data from the ten participants with normal color vision.

We expect the MDS representation to use the Euclidean metric, consistent with the integral nature of the color stimulus domain. We also expect a two-dimensional representation, following the color circle found by previous MDS analyses of these

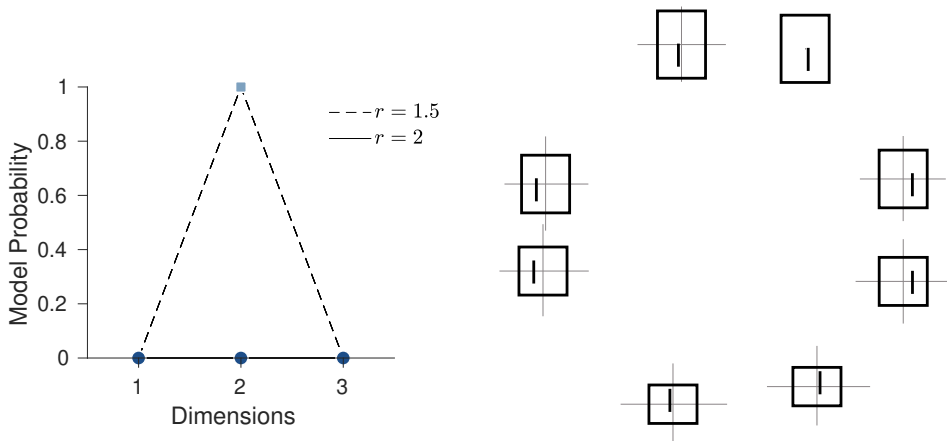


Figure 5.10: Results for rectangles with interior line segments data reported by Kruschke (1993). The left panel shows the posterior probabilities for one- through three-dimensional MDS representations, for both the Minkowski metrics with  $r = 1.5$  and  $r = 2$ . The right panel shows the inferred two-dimensional representation. The stimuli are shown at their inferred locations and error bars show 95% credible intervals for the marginal posterior distribution for each dimension.

and other color similarity data, such as the Shepard (1962) original MDS analysis of data reported by Ekman (1954).

Figure 5.9 shows the results of applying our method, assuming a Euclidean metric. This was a case in which we were unable to generate samples with acceptable convergence for the  $r = 1.5$  metric. For the Euclidean metric, there is uncertainty regarding the dimensionality, with a three-dimensional representation having probability a little over 0.6 and a two-dimensional representation having almost all of the remaining probability. The inferred three-dimensional representation is shown by pairing the first two dimensions as a two-dimensional plot in the center of Figure 5.9, and showing the remaining third dimension separately to the right along an axis. Because of our ordering heuristic, the yellow and purple-blue stimuli were fixed at the origin and on the first axis. These assignments mean that the first two dimensions effectively represent the expected color circle that “bends” the visible physical spectrum from red to purple colors into a circle that reflects the psychological similarity between the end points. The third dimension, which we did not expect, could correspond to something like luminance, since low luminance purple-like colors are generally located at one end of the dimension and high luminance yellow-like colors are generally located at the other end.

### 5.5.3 Rectangles with Line Segments

Our third application involves data reported by Kruschke (1993) involving the similarity between eight geometric stimuli. These stimuli consisted of rectangles

with interior line segments, and varied in terms of the height of the rectangle and the horizontal location of the line segment. A total of 50 participants provided similarity ratings on a nine-point scale for all 28 stimulus pairs. Based on the original (Kruschke, 1993) and subsequent (e.g., Lee, 2001, 2008) analyses of these data, we expect a two-dimensional MDS representation. We also expect the two stimulus dimensions to be psychologically separable.

Figure 5.10 shows the results of applying our method assuming both the  $r = 1.5$  and  $r = 2$  metrics. It is clear that a two-dimensional representation with the separable  $r = 1.5$  metric is inferred. It has essentially all of the posterior probability, with one- and three-dimensional  $r = 1.5$  representations, and all of the  $r = 2$  representations having essentially no posterior probability. The inferred representation closely matches the ways in which the stimuli physically vary, with each psychological axis corresponding to an interpretable stimulus dimension. The horizontal axis corresponds to the position of the line segment and the vertical axis corresponds to the height of the rectangle.

#### 5.5.4 Shepard Circles

Our fourth application involves data collected by Treat, McFall, Viken, and Kruschke (2001), involving the similarity between nine geometric stimuli known as “Shepard circles”. These stimuli consist of a closed semi-circle with an interior ray from the center to the perimeter. The nine stimuli are constructed by exhaustively varying three different radius lengths and three different angles for the internal ray. As for the rectangles with line segments, we expect a separable two-dimensional MDS representation. For these stimuli, we expect the dimensions to correspond to the radius and angle dimensions.

Figure 5.11 shows the results of applying our method assuming both the  $r = 1.5$  and  $r = 2$  metrics.<sup>8</sup> It is clear, once again, that a two-dimensional representation with the separable  $r = 1.5$  metric is inferred. The inferred representation also again closely matches the ways in which the stimuli physically vary, with the horizontal axis corresponding to the radius of the semi-circle and the vertical axis corresponding to the angle of the ray.

#### 5.5.5 Colored Shapes

Our final application considers similarity data for nine colored shape stimuli collected by Lee and Navarro (2002). The stimuli were circles, squares, and triangles that were colored red, green, and blue. The data were collected from 20 participants, each of whom rated the similarity of each pair of stimuli on a five-point scale.

Following the previous analysis in Lee and Navarro (2002), we expect a four-dimensional representation. This representation is best understood as being the product of a pair of two-dimensional representations, with one representing the similarities between the shapes, and the other representing the similarities between the colors. There are only three shapes and three colors, and neither set

<sup>8</sup>For these stimuli, we did not have access to information about the precise physical values of the radius and angles, and so the depictions in Figure 5.11 are approximate.

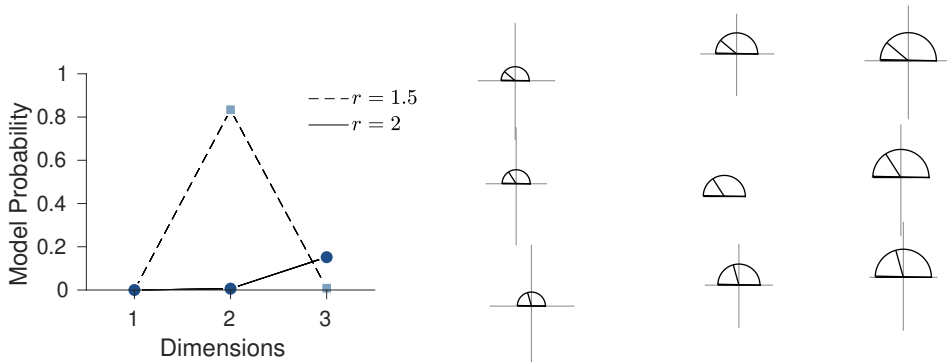


Figure 5.11: Results for the Shepard circles data collected by Treat et al. (2001). The left panel shows the posterior probabilities for one- through three-dimensional MDS representations, for both the Minkowski metrics with  $r = 1.5$  and  $r = 2$ . The right panel shows the inferred two-dimensional representation. The stimuli are shown at their inferred locations and error bars show 95% credible intervals for the marginal posterior distribution for each dimension.

of three has a natural ordering. Instead, the circle, square, and triangle are all approximately equally different from one another, and the same is true of the red, green, and blue colors. These equal similarities are naturally represented by two-dimensional approximately equilateral triangles. The four-dimensional representation we expect is simply the independent combination of these two two-dimensional sub-spaces.

Our expectations for the metric structure of the MDS representations are less straightforward. Theoretically, the interaction between the shape and color dimensions is a classic example of a separable relationship. The metric structure within the color sub-space, however, is theoretically integral, as for the previous application. Countering these theoretical expectations is the fact that there are only three values for the color and shape dimensions present in the stimulus set. The corresponding approximately equilateral triangles could be equally well accommodated by any of the Minkowski metrics we are considering. Thus, from a statistical perspective – without regard to the theory of separable and integral stimuli – we expect the simplest metric to be inferred. Since all metrics should be able to fit the data, the one with the smallest functional form complexity should be preferred.

We found that this was a third case in which we were unable to generate samples with acceptable convergence for the  $r = 1.5$  metric. Accordingly, Figure 5.12 shows the results of applying our method assuming the Euclidean metric. A four-dimensional representation is clearly favored. This representation is shown in terms of two two-dimensional subspaces, and has the expected structure. The middle panel of Figure 5.12 shows a subspace that captures the similarity relationships between the red, green, and blue colors. The right panel shows a subspace

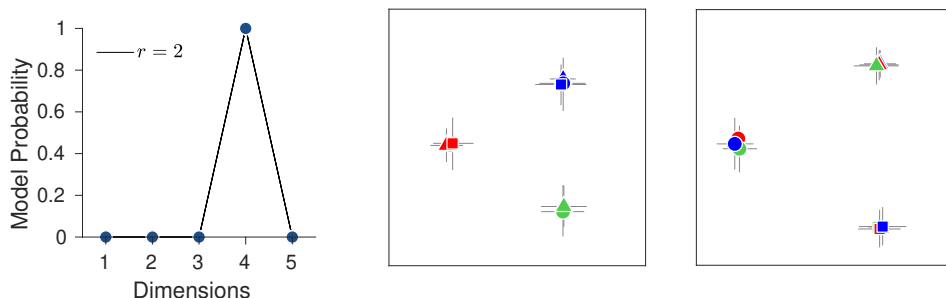


Figure 5.12: Results for colored shapes data reported by Lee and Navarro (2002). The left panel shows the posterior probabilities for one- through five-dimensional MDS representations for the Euclidean metric. The middle and right panels show the inferred four-dimensional representation, with two dimensions shown in each panel. The colored shapes show the inferred locations of each stimulus and error bars show 95% credible intervals for the marginal posterior distribution for each dimension.

that captures the similarity relationships between the circle, square, and triangle shapes. These subspaces were found using an orthogonal Procrustes method (Borg & Groenen, 1997, p. 162). In particular, we solved for the orthogonal transformation matrix that most closely mapped the inferred coordinate locations to the expected representational structure, defined as the product of two subspaces each with an equilateral triangle configuration.

## 5.6 Discussion

Collectively, the five applications demonstrate that our method is able to make reasonable inferences about MDS representations. The inferred number of dimensions, and the inferred stimulus locations, generally matched theoretical expectations, with the exception of the color application. In addition, where inferences about whether a Euclidean or non-Euclidean metric structure were made, they matched theoretical expectations. It is interesting to note that all of the applications for which non-Euclidean metrics made inference difficult involved stimulus domains for which the expectation was that the Euclidean metric was appropriate.

We also think that the five applications serve to demonstrate the usefulness of our approach to determining dimensionality and metric structure. Our approach is to treat these determinations as Bayesian model-selection problems and use Bayes factors to make inferences. Complete Bayes factors have not been used in this way previously to determine either dimensionality or metric structure, and our introduction of the Warp-III method to solve the difficult computational approximation problems involved represents progress on these long-standing challenges in MDS modeling.

Despite this progress, we think the greatest contribution of the current work is to highlight fundamental challenges in MDS models of mental representation,

and suggest new avenues for theoretical development. The challenges largely stem from our insistence on fully Bayesian inference, which has enormous advantages in terms of reaching complete, coherent, and principled conclusions, but also raises technical hurdles. The opportunities largely stem from our adoption of a generative modeling approach (Lee, 2018). In particular, we think there are many remaining possibilities relating to the use of different metrics in MDS representations, and that there is an opportunity to extend the generative approach to develop more complete cognitive process models for inferring MDS representations. We conclude by discussing some of these challenges and opportunities.

### 5.6.1 Technical Challenges

Developing a generative MDS model in a Bayesian setting required the key issue of identifiability and invariance to be solved in terms of prior information, rather than more heuristically through post-processing. We used an existing solution to this challenge for the Euclidean metric, and proposed a solution for psychologically-interpretable non-Euclidean metrics with  $1 < r < 2$ . We also highlighted, however, the fundamental intractability of MDS representations using the city-block metric. This intractability has been documented before (Bortz 1974; Frank 2006, Figure 5.4; Shepard 1974, Figure 11), but has not prevented the use of MDS representations inferred based on the city-block metric in the cognitive modeling literature (e.g., Kruschke, 1993; Lee & Wetzels, 2010).

Our current approach to determining the appropriate metric treats this inference as a model-selection problem, and only considers the possibilities  $r = 1.5$  and  $r = 2$ . Allowing for other metrics is theoretically interesting, but computationally difficult. One obvious cost is the need to generate posterior probabilities across a larger set of candidate models. But it also seems likely that some models will be difficult to make inferences about. We tried our DE-MCMC approach for  $r = 1.1$  on a number of data sets, and were not able to achieve satisfactory convergence. Furthermore, as explained above, for a few of the applications we were also not able to achieve satisfactory convergence for  $r = 1.5$ . These challenging cases involved stimulus domains for which the expectation was that the Euclidean metric was appropriate, which leads to a speculative suggestion that failure is related to model mis-specification. This is a potential example of a general aspect of Bayesian model comparison that can be computationally challenging: in order to rule out models that are likely mis-specified, one needs to be able to infer them well enough that they can be part of the model comparison. Although we believe that DE-MCMC is a powerful sampling algorithm which substantially helps alleviate the issue of non-converging chains, future research should explore different sampling algorithms that may perform better, particularly for non-Euclidean metrics.

Collectively, these technical challenges mean that our approach cannot currently be applied to large naturalistic stimulus domains. For example, Nosofsky, Sanders, Meagher, and Douglas (2018) consider MDS representations based on sparse matrices of pairwise similarity judgments for a set of 360 images of rocks, and Hebart, Zheng, Pereira, and Baker (2020) report extensive crowd-sourced triadic comparison similarity data for 1854 images of real-world objects. Being able to

determine the dimensionality, metric structure, and psychological representations of MDS representations of these domains using the Bayesian framework would potentially offer deep insight into how people represent the real-world stimuli. The successful applications we presented – in which there were clear expectations about dimensionality, metric, and representational structure – provide a basis for believing the Bayesian framework can provide this insight to situations where answers must be inferred from data, if and when the computational technical hurdles are overcome.

### 5.6.2 Other Representations

We did not consider Minkowski metrics with  $r < 1$ . This possibility has been proposed as a way of representing stimulus domains in which the component dimensions compete for attention (Shepard, 1987, 1991; Tversky & Gati, 1982). The identifiability constraints for this metric present an open research challenge, and it is not clear how well DE-MCMC sampling methods will perform in inferring representations.

There is also the possibility of moving beyond the Minkowski family of metrics. In his presidential address, Shepard (1974, Figure 11) presented a taxonomy of metric spaces, each of which makes different fundamental representational assumptions that could be appropriate for at least some stimulus domains. There has been relatively little work in exploring these possibilities. Lindman and Caelli (1978) investigated MDS representations using Riemannian spaces with constant curvature, and Cox and Cox (1991) presented compelling applications for a special case of this approach involving MDS representations on a sphere.

A new idea raised by our application to the colored shape stimuli involves the possibility of different metric structures within the same representation. These stimuli involved two sorts of stimulus dimensions: those representing color, which are usually considered to be integral, and those representing qualitatively different shapes, which seems more separable. Certainly the interaction between the color dimensions and the shape dimensions would be expected to be separable, since it seems likely people can selectively attend to either the color or the shape of a stimulus, depending upon the cognitive context. This suggests a generalization of the MDS models in which each pair of dimensions is associated with a metric.

Finally, there are alternative representational models, which do not assume stimuli are represented by values on dimensions, that can compete with or complement MDS models. These alternatives include feature-based representations (Tversky, 1977), such as those found by additive clustering and related methods (Shepard & Arabie, 1979) and special cases such as tree-based models (Corter, 1996; Shepard, 1980). One attraction of the Warp-III approach we used is that it could estimate Bayes factors between fundamentally different sorts of representations – such as comparing dimensional and featural representations – since it operates directly on posterior samples for each model applied independently to the data. Even further, Navarro and Lee (2003) proposed a hybrid model of stimulus representation that combined both dimensions and features, and it would be conceptually elegant to choose between all of the candidate models, with various combinations of dimensions and features, using our methods. Navarro and Lee

(2003) used an approximate analytic approach for this purpose, which would be significantly improved by an approach based on Bayes factors.

### 5.6.3 MDS Cognitive Process Models

Our modeling approach is generative, but is based on an extremely simple cognitive model. In essence, we assume that all participants have the same MDS representation, and produce dissimilarity judgments for pairs of stimuli that directly reflect the distances between those stimuli in the representation. It is likely that much better generative models can be developed by considering more realistic processing assumptions, and especially by including individual differences.

One example, involving the line length application, was presented in a preliminary form by Lee (2014). A simple plot of the raw behavioral data suggests that one of the 27 participants appears to have reversed the scale that was used to judge similarity. This means that their judgments contaminate the inference of the MDS representation. Lee (2014) used a simple latent-mixture model extension of the basic MDS generative model, in which either the scale was used correctly or reversed. One participant was inferred to have reversed the scale, as expected. Perhaps more importantly, however, the resulting inference about the one-dimensional MDS representation was shown to have less uncertainty than the one shown in Figure 5.8. In this way, the introduction of individual differences in the cognitive process of similarity judgment helped decontaminate the inference about the representation of stimuli.

The same basic generative approach could support much more general cognitive process modeling using MDS representations. The hierarchical, latent mixture, and common cause model structures advocated by Lee (2018) could allow for rich accounts of individual differences in judgment processes or stimulus representations, and allow for models that extend beyond the judgment of similarity to other cognitive capabilities like categorization and inference. As one example, Ennis (1992) considers extended assumptions about MDS representations that allow for the noisy representation of perceptual stimuli, which could be incorporated by adding hierarchical structure to the coordinate locations. As another example, there are extensions of the basic MDS model we considered that allow for structured individual differences, such as INDSCAL (Carroll, 1972; Carroll & Chang, 1970). These would be easy to implement within our generative modeling framework. A model like INDSCAL, which assumes individuals weight the latent stimulus dimensions differently, relies on the appropriate number of dimensions being inferred, and evidence that the stimulus domain is separable. In this way, the potential of our method to make these inferences is especially important. As a final example, the rectangle and line segment stimuli are used by Kruschke (1993) to study category learning, but the similarity data and category learning data are analyzed independently. In effect, the similarity data are used to generate the MDS representation, and that representation is then assumed to provide the fixed basis for category learning. An alternative approach would be to infer the MDS representation jointly from both the similarity judgments and the category learning choices. This sort of flexibility raises the possibility of tackling more

complicated cognitive phenomena, such as the ability to adapt representations in response to changes in the external environment, or the current context or goals.

#### **5.6.4 Conclusion**

We adopted a Bayesian model selection approach to the problem of determining the dimensionality and metric structure of MDS representations, while considering psychologically-interpretable Euclidean and non-Euclidean metrics. Our methods for inferring the representations, and choosing their dimensionality and metric structure show the promise of the approach, but computational challenges remain a barrier in terms of an easy-to-use general capability. Our methods and applications also show the promise of placing MDS representations in a generative cognitive modeling framework, offering the possibility of new models of how people represent stimuli, and how those representations help guide behavior.

All code is available at <https://osf.io/82g3r/>.

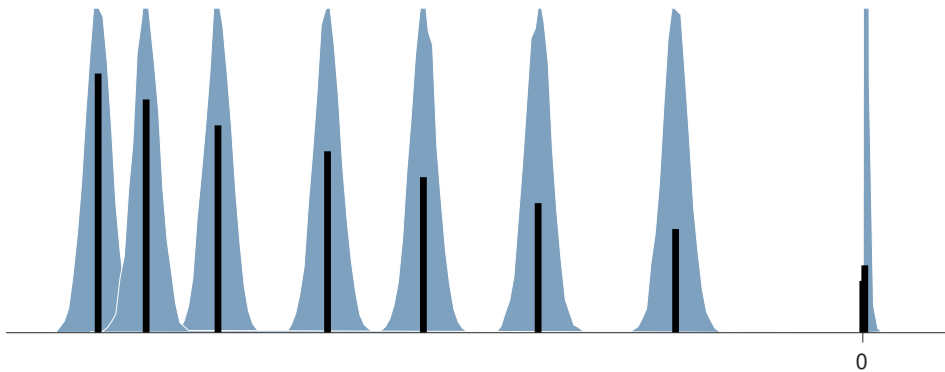


Figure 5.13: A suboptimal one-dimensional representation of the line-length similarity data from Cohen et al. (2001), motivating the need for the ordering heuristic. The black lines show the stimuli at their inferred locations in the representation, and the blue histograms show the marginal posterior distributions for these locations.

## 5.A The Ordering Heuristic

Figure 5.13 provides a concrete example to motivate the need for the ordering heuristic. It is clear this is an inferior representation to the one presented in Figure 5.8. In Figure 5.13, the first and second line stimuli, which are the two shortest, are located at almost the same point, rather than being appropriately spaced to reflect their psychological dissimilarity. Consistent with this intuition, the posterior density is worse for the representation in Figure 5.13 than the representation in Figure 5.8.

This suboptimality is caused by the naive application of the constraints identified in Figure 5.2 for a one-dimensional representation. The first stimulus is fixed at the origin, and the second stimulus is constrained to be positive. It is clear from Figure 5.13 that the second stimulus is indeed inferred to be positive, but is extremely close to zero, with the remaining longer line stimuli “flipping” to negative values in the MDS space. This configuration still satisfies the proximity data reasonably well, because the required distance between the first two stimuli is small, and the distances from the first and second stimuli to all of the others is approximately conserved. Thus, it is the choice of the two similar stimuli as those that are constrained that leads to this potential for a local maximum and suboptimal representation.

Accordingly, we developed an ordering heuristic to try and assign the constraints for the various dimensionalities and metrics to stimuli that are sufficiently dissimilar. Because higher dimensionalities place constraints on more than two stimuli, the general approach is to order all of the stimuli. Our heuristic for doing this is based on the across participants averaged pairwise dissimilarity ratings. The first two stimuli are chosen to be the ones with the largest averaged pairwise dissimilarity. The remaining stimuli are chosen, one at a time, by considering the

minimum averaged pairwise dissimilarity to the already selected stimuli. Specifically, the next stimulus is always chosen to be the one with the maximum value for the minimum averaged pairwise dissimilarity to the already selected stimuli.

We used this ordering heuristic for the colors and colored shapes applications. For the line length application, we used the heuristic as described but then, in an additional step, switched the first stimulus with the second stimulus. This switch helped prevent the posterior for the ninth stimulus, corresponding to the longest line, push against the upper bound of 1. For the rectangles with interior line segments and Shepard circles applications, we used the heuristic as a starting point, but we then reordered some of the stimuli manually since it seemed to help with convergence.

## 5.B Transformation Ordered Vector (0-1 Bounded)

The constrained vector  $\mathbf{x}$ ,  $0 \leq x_1 \leq x_2 \leq \dots \leq x_K \leq 1$ , can be transformed to an unconstrained vector  $\mathbf{y} \in \mathbb{R}^K$  as follows:

$$y_k = \begin{cases} \Phi^{-1}(x_k) & \text{if } k = 1, \\ \Phi^{-1}\left(\frac{x_k - x_{k-1}}{1 - x_{k-1}}\right) & \text{if } 1 < k \leq K, \end{cases}$$

where  $\Phi^{-1}(\cdot)$  denotes the inverse of the normal CDF. The inverse transformation is given by:

$$x_k = \begin{cases} \Phi(y_k) & \text{if } k = 1, \\ x_{k-1} + (1 - x_{k-1}) \Phi(y_k) & \text{if } 1 < k \leq K, \end{cases}$$

where  $\Phi(\cdot)$  denotes the normal CDF. Note that  $x_k$  is a function of  $y_1, y_2, \dots, y_k$  (the dependence on  $y_1, y_2, \dots, y_{k-1}$  is “hidden” in  $x_{k-1}$ ). Crucially,  $x_k$  does not depend on  $y_{k+1}, y_{k+2}, \dots, y_K$ . Consequently, the Jacobian matrix  $\mathcal{J}$  of the transformation is lower triangular so that its determinant  $|\mathcal{J}|$  is obtained by multiplying its diagonal entries. The diagonal entries are given by:

$$\mathcal{J}_{k,k} = \begin{cases} \phi(y_k) & \text{if } k = 1, \\ (1 - x_{k-1}) \phi(y_k) & \text{if } 1 < k \leq K, \end{cases}$$

where  $\phi(\cdot)$  denotes the normal PDF. Hence, the determinant of the Jacobian matrix is given by:

$$|\mathcal{J}| = \phi(y_1) \prod_{k=2}^K [(1 - x_{k-1}) \phi(y_k)].$$



---

# bridgesampling: An R Package for Estimating Normalizing Constants

---

## Abstract

Statistical procedures such as Bayes factor model selection and Bayesian model averaging require the computation of normalizing constants (e.g., marginal likelihoods). These normalizing constants are notoriously difficult to obtain, as they usually involve high-dimensional integrals that cannot be solved analytically. Here we introduce an R package that uses bridge sampling (Meng & Schilling, 2002; Meng & Wong, 1996) to estimate normalizing constants in a generic and easy-to-use fashion. For models implemented in **Stan**, the estimation procedure is automatic. We illustrate the functionality of the package with three examples.

## 6.1 Introduction

In many statistical applications, it is essential to obtain normalizing constants of the form

$$Z = \int_{\Theta} q(\boldsymbol{\theta}) \, d\boldsymbol{\theta}, \quad (6.1)$$

where  $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/Z$  denotes a probability density function (pdf) defined on the domain  $\Theta \subseteq \mathbb{R}^p$ . For instance, the estimation of normalizing constants plays a crucial role in free energy estimation in physics, missing data analyses in likelihood-based approaches, Bayes factor model comparisons, and Bayesian model averaging (e.g., Gelman & Meng, 1998). In this chapter, we focus on the role of the normalizing constant in Bayesian inference; however, the **bridgesampling** package can be used in any context where one desires to estimate a normalizing constant.

---

This chapter is published as Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92. doi: <https://doi.org/10.18637/jss.v092.i10>. Also available as *arXiv preprint*: <https://arxiv.org/abs/1710.08162>

In Bayesian inference, the normalizing constant of the joint posterior distribution is involved in (a) parameter estimation, where the normalizing constant ensures that the posterior integrates to one; (b) Bayes factor model comparison, where the ratio of normalizing constants quantifies the data-induced change in beliefs concerning the relative plausibility of two competing models (e.g., Kass & Raftery, 1995); (c) Bayesian model averaging, where the normalizing constant is required to obtain posterior model probabilities (BMA; Hoeting et al., 1999).

For Bayesian parameter estimation, the need to compute the normalizing constant can usually be circumvented by the use of sampling approaches such as Markov chain Monte Carlo (MCMC; e.g., Gamerman & Lopes, 2006). However, for Bayes factor model comparison and BMA, the normalizing constant of the joint posterior distribution – in this context usually called *marginal likelihood* – remains of essential importance. This is evident from the fact that the posterior model probability of model  $\mathcal{M}_i$ ,  $i \in \{1, 2, \dots, m\}$ , given data  $\mathbf{y}$  is obtained as

$$\underbrace{p(\mathcal{M}_i | \mathbf{y})}_{\text{posterior model probability}} = \underbrace{\frac{p(\mathbf{y} | \mathcal{M}_i)}{\sum_{j=1}^m p(\mathbf{y} | \mathcal{M}_j) p(\mathcal{M}_j)}}_{\text{updating factor}} \times \underbrace{p(\mathcal{M}_i)}_{\text{prior model probability}}, \quad (6.2)$$

where  $p(\mathbf{y} | \mathcal{M}_i)$  denotes the *marginal likelihood* of model  $\mathcal{M}_i$ .

If the model comparison involves only two models,  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , it is convenient to consider the odds of one model over the other. Bayes' rule yields:

$$\underbrace{\frac{p(\mathcal{M}_1 | \mathbf{y})}{p(\mathcal{M}_2 | \mathbf{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathbf{y} | \mathcal{M}_1)}{p(\mathbf{y} | \mathcal{M}_2)}}_{\text{Bayes factor BF}_{12}} \times \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{prior odds}}. \quad (6.3)$$

The change in odds brought about by the data is given by the ratio of the marginal likelihoods of the models and is known as the *Bayes factor* (Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995). Equation 6.2 and Equation 6.3 highlight that the normalizing constant of the joint posterior distribution, that is, the marginal likelihood, is required for computing both posterior model probabilities and Bayes factors.

The marginal likelihood is obtained by integrating out the model parameters with respect to their prior distribution:

$$p(\mathbf{y} | \mathcal{M}_i) = \int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}, \mathcal{M}_i) p(\boldsymbol{\theta} | \mathcal{M}_i) d\boldsymbol{\theta}. \quad (6.4)$$

The marginal likelihood implements the principle of parsimony also known as *Occam's razor* (e.g., Jefferys & Berger, 1992; Myung & Pitt, 1997; Vandekerckhove et al., 2015). Unfortunately, the marginal likelihood can be computed analytically for only a limited number of models. For more complicated models (e.g., hierarchical models), the marginal likelihood is a high-dimensional integral that usually cannot be solved analytically. This computational hurdle has complicated the application of Bayesian model comparisons for decades.

To overcome this hurdle, a range of different methods have been developed that vary in accuracy, speed, and complexity of implementation: naive Monte Carlo

estimation, importance sampling, the generalized harmonic mean estimator, Reversible Jump MCMC (Green, 1995), the product-space method (Carlin & Chib, 1995; Lodewyckx et al., 2011), Chib’s method (Chib, 1995), thermodynamic integration (e.g., Lartillot & Philippe, 2006), path sampling (Gelman & Meng, 1998), and others. The ideal method is fast, accurate, easy to implement, general, and unsupervised, allowing non-expert users to treat it as a “black box”.

In our experience, one of the most promising methods for estimating normalizing constants is bridge sampling (Meng & Schilling, 2002; Meng & Wong, 1996). Bridge sampling is a general procedure that performs accurately even in high-dimensional parameter spaces such as those that are regularly encountered in hierarchical models. In fact, simpler estimators such as the naive Monte Carlo estimator, the generalized harmonic mean estimator, and importance sampling are special sub-optimal cases of the bridge identity described in more detail below (e.g., Frühwirth-Schnatter, 2004; Gronau, Sarafoglou, et al., 2017).

In this chapter, we introduce **bridgesampling**, an R (R Core Team, 2019) package that enables the straightforward and user-friendly estimation of the marginal likelihood (and of normalizing constants more generally) via bridge sampling techniques. In general, the user needs to provide to the `bridge_sampler` function four quantities that are readily available:

- an object with posterior samples (argument `samples`);
- a function that computes the log of the unnormalized posterior density for a set of model parameters (argument `log_posterior`);
- a data object that contains the data and potentially other relevant quantities for evaluating `log_posterior` (argument `data`);
- lower and upper bounds for the parameters (arguments `lb` and `ub`, respectively).

Given these inputs, the **bridgesampling** package provides an estimate of the log marginal likelihood.

Figure 6.1 displays the steps that a user may take when using the **bridge-sampling** package. Starting from the top, the user provides the basic required arguments to the `bridge_sampler` function which then produces an estimate of the log marginal likelihood. With this estimate in hand – usually for at least two different models – the user can compute posterior model probabilities using the `post_prob` function, Bayes factors using the `bf` function, and approximate estimation errors using the `error_measures` function. A schematic call of the `bridge_sampler` function looks as follows (detailed examples are provided in the next sections):

```
R> bridge_sampler(samples = samples, log_posterior = log_posterior,
+               data = data, lb = lb, ub = ub)
```

The `bridge_sampler` function is an S3 generic which currently has methods for objects of class `mcmc`, `mcmc.list` (Plummer et al., 2006), `stanfit` (Stan Development Team, 2016), `matrix`, `rjags` (Plummer, 2016; Su & Yajima, 2015), `runjags`

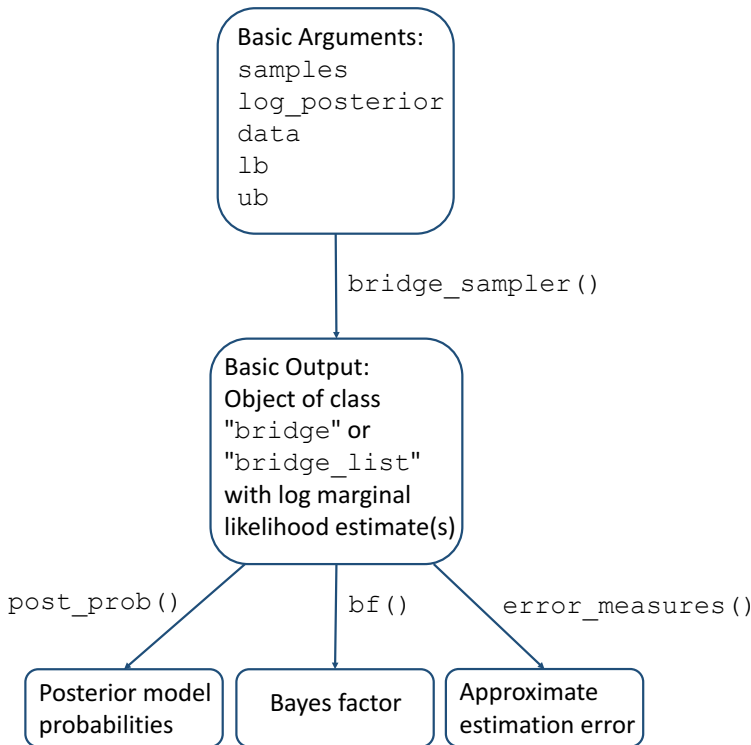


Figure 6.1: Flow chart of the steps that a user may take when using the **bridge-sampling** package. In general, the user needs to provide a posterior samples object (`samples`), a function that computes the log of the unnormalized posterior density (`log_posterior`), the data (`data`), and parameter bounds (`lb` and `ub`). The `bridge_sampler` function then produces an estimate of the log marginal likelihood. This is usually repeated for at least two different models. The user can then compute posterior model probabilities (using the `post_prob` function), Bayes factors (using the `bf` function), and approximate estimation errors (using the `error_measures` function). Note that the summary method for `bridge` objects automatically invokes the `error_measures` function. Figure available at <https://tinyurl.com/ybf4jxka> under CC license <https://creativecommons.org/licenses/by/2.0/>.

(Denwood, 2016), `stanreg` (Team, 2016), and for `MCMC_refClass` objects produced by `nimble` (de Valpine et al., 2017).<sup>1</sup> This allows the user to obtain posterior samples in a convenient and efficient way, for instance, via `JAGS` (Plummer, 2003) or

<sup>1</sup>We thank Ben Goodrich for adding the `stanreg` method to our package and Perry de Valpine for his help implementing the `nimble` support.

a highly customized sampler. Hence, bridge sampling does not require users to program their own MCMC routines to obtain posterior samples; this convenience is usually missing for methods such as Reversible Jump MCMC (but see Gelling, Schofield, & Barker, 2017).

When the model is specified in **Stan** (Carpenter et al., 2017; Stan Development Team, 2016) – in a way that retains the constants, as described below – obtaining the marginal likelihood is even simpler: the user only needs to pass the **stanfit** object to the **bridge\_sampler** function. The combination of **Stan** and the **bridge-sampling** package therefore produces an unsupervised, black box computation of the marginal likelihood.

This chapter is structured as follows: First we describe the implementation details of the algorithm from **bridgesampling**; second, we illustrate the functionality of the package using a simple Bayesian  $t$ -test example where posterior samples are obtained via **JAGS**. In this section, we also explain a heuristic to obtain the function that computes the log of the unnormalized posterior density in **JAGS**; third, we describe in more detail the interface to **Stan** which enables an even more automatized computation of the marginal likelihood. Fourth, we illustrate use of the **Stan** interface with two well-known examples from the Bayesian model selection literature.

## 6.2 Bridge Sampling: The Algorithm

Bridge sampling can be thought of as a generalization of simpler methods for estimating normalizing constants such as the naive Monte Carlo estimator, the generalized harmonic mean estimator, and importance sampling (e.g., Frühwirth-Schnatter, 2004; Gronau, Sarafoglou, et al., 2017). These simpler methods typically use samples from a single distribution, whereas bridge sampling combines samples from *two* distributions.<sup>2</sup> For instance, in its original formulation (Meng & Wong, 1996), bridge sampling was used to estimate a ratio of two normalizing constants such as the Bayes factor. In this scenario, the two distributions for the bridge sampler are the posteriors for each of the two models involved. However, the accuracy of the estimator depends crucially on the overlap between the two involved distributions; consequently, the accuracy can be increased by estimating a single normalizing constant at a time, using as a second distribution a convenient normalized proposal distribution that closely matches the distribution of interest (e.g., Gronau, Sarafoglou, et al., 2017; Overstall & Forster, 2010). The bridge sampling estimator of the marginal likelihood is then given by:<sup>3</sup>

$$p(\mathbf{y}) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})} [h(\boldsymbol{\theta}) p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{y})} [h(\boldsymbol{\theta}) g(\boldsymbol{\theta})]} \approx \frac{\frac{1}{n_2} \sum_{j=1}^{n_2} h(\tilde{\boldsymbol{\theta}}_j) p(\mathbf{y} | \tilde{\boldsymbol{\theta}}_j) p(\tilde{\boldsymbol{\theta}}_j)}{\frac{1}{n_1} \sum_{i=1}^{n_1} h(\boldsymbol{\theta}_i^*) g(\boldsymbol{\theta}_i^*)}, \quad (6.5)$$

<sup>2</sup>Note, however, that these simpler methods are special cases of bridge sampling (e.g., Gronau, Sarafoglou, et al., 2017, Appendix A). Hence, for particular choices of the bridge function and the proposal distribution, only samples from one distribution are used.

<sup>3</sup>We omit conditioning on the model for enhanced legibility. It should be kept in mind, however, that this yields the estimate of the marginal likelihood for a particular model  $\mathcal{M}_i$ , that is,  $p(\mathbf{y} | \mathcal{M}_i)$ .

where  $h(\boldsymbol{\theta})$  is called the *bridge function* and  $g(\boldsymbol{\theta})$  denotes the *proposal distribution*.  $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \dots, \boldsymbol{\theta}_{n_1}^*\}$  denote  $n_1$  samples from the posterior distribution  $p(\boldsymbol{\theta} | \mathbf{y})$  and  $\{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_{n_2}\}$  denote  $n_2$  samples from the proposal distribution  $g(\boldsymbol{\theta})$ .

To use bridge sampling in practice, one has to specify the bridge function  $h(\boldsymbol{\theta})$  and the proposal distribution  $g(\boldsymbol{\theta})$ . For the bridge function  $h(\boldsymbol{\theta})$ , the **bridgesampling** package implements the optimal choice presented in Meng and Wong (1996) which minimizes the relative mean-squared error of the estimator. Using this particular bridge function, the bridge sampling estimate of the marginal likelihood is obtained via an iterative scheme that updates an initial guess of the marginal likelihood  $\hat{p}(\mathbf{y})^{(0)}$  until convergence (for details, see Gronau, Sarafoglou, et al., 2017; Meng & Wong, 1996). The estimate at iteration  $t + 1$  is obtained as follows:

$$\hat{p}(\mathbf{y})^{(t+1)} = \frac{\frac{1}{n_2} \sum_{j=1}^{n_2} \frac{l_{2,j}}{s_1 l_{2,j} + s_2 \hat{p}(\mathbf{y})^{(t)}}}{\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{s_1 l_{1,i} + s_2 \hat{p}(\mathbf{y})^{(t)}}}, \quad (6.6)$$

where  $l_{1,i} = \frac{p(\mathbf{y} | \boldsymbol{\theta}_i^*) p(\boldsymbol{\theta}_i^*)}{g(\boldsymbol{\theta}_i^*)}$ , and  $l_{2,j} = \frac{p(\mathbf{y} | \tilde{\boldsymbol{\theta}}_j) p(\tilde{\boldsymbol{\theta}}_j)}{g(\tilde{\boldsymbol{\theta}}_j)}$ . In practice, a more numerically stable version of Equation 6.6 is implemented that uses logarithms in combination with the **Brobdignag** R package (Hankin, 2007) to avoid numerical under- and overflow (for details, see Gronau, Sarafoglou, et al., 2017, Appendix B).

The iterative scheme usually converges within a few iterations. Note that, crucially,  $l_{1,i}$  and  $l_{2,j}$  need only be computed once before the iterative updating scheme is started. In practice, evaluating  $l_{1,i}$  and  $l_{2,j}$  takes up most of the computational time. Luckily,  $l_{1,i}$  and  $l_{2,j}$  can be computed completely in parallel for each  $i \in \{1, 2, \dots, n_1\}$  and each  $j \in \{1, 2, \dots, n_2\}$ , respectively. That is, in contrast to MCMC procedures, the evaluation of, for instance,  $l_{1,i+1}$  does *not* require one to evaluate  $l_{1,i}$  first (since the posterior samples and proposal samples are already available). The **bridgesampling** package enables the user to compute  $l_{1,i}$  and  $l_{2,j}$  in parallel by setting the argument `cores` to an integer larger than one. On Unix/-macOS machines, this parallelization is implemented using the **parallel** package. On Windows machines this is achieved using the **snowfall** package (Knaus, 2015).<sup>4</sup>

After having specified the bridge function, one needs to choose the proposal distribution  $g(\boldsymbol{\theta})$ . The **bridgesampling** package implements two different choices: (a) a multivariate normal proposal distribution with mean vector and covariance matrix that match the respective posterior samples quantities and (b) a standard multivariate normal distribution combined with a *warped* posterior distribution.<sup>5</sup> Both choices increase the efficiency of the estimator by making the proposal and the posterior distribution as similar as possible. Note that under the optimal bridge function, the bridge sampling estimator is robust to the relative tail behavior of the posterior and the proposal distribution. This stands in sharp contrast to the importance and the generalized harmonic mean estimator for which unwanted

---

<sup>4</sup>Due to technical limitations specific to Windows, this parallelization is not available for the **stanfit** and **stanreg** methods.

<sup>5</sup>Note that other proposal distributions such as multivariate  $t$  distributions are conceivable but are currently not implemented in the **bridgesampling** package.

tail behavior produces estimators with very large or even infinite variances (e.g., Frühwirth–Schnatter, 2004; Gronau, Sarafoglou, et al., 2017; Owen & Zhou, 2000).

### 6.2.1 Option I: The Multivariate Normal Proposal Distribution

The first choice for the proposal distribution that is implemented in the **bridgesampling** package is a multivariate normal distribution with mean vector and covariance matrix that match the respective posterior samples quantities. This choice (henceforth “the normal method”) generalizes to high dimensions and accounts for potential correlations in the joint posterior distribution. This proposal distribution is obtained by setting the argument `method = "normal"` in the `bridge_sampler` function; this is the default setting. This choice assumes that all parameters are allowed to range across the entire real line. In practice, this assumption may not be fulfilled for all components of the parameter vector, however, it is usually possible to transform the parameters so that this requirement is met. This is achieved by transforming the original  $p$ -dimensional parameter vector  $\boldsymbol{\theta}$  (which may contain components that range only across a subset of  $\mathbb{R}$ ) to a new parameter vector  $\boldsymbol{\xi}$  (where all components are allowed to range across the entire real line) using a diffeomorphic vector-valued function  $f$  so that  $\boldsymbol{\xi} = f(\boldsymbol{\theta})$ . By the change-of-variable rule, the posterior density with respect to the new parameter vector  $\boldsymbol{\xi}$  is given by:

$$p(\boldsymbol{\xi} \mid \mathbf{y}) = p_{\boldsymbol{\theta}}(f^{-1}(\boldsymbol{\xi}) \mid \mathbf{y}) \left| \det [J_{f^{-1}}(\boldsymbol{\xi})] \right|, \quad (6.7)$$

where  $p_{\boldsymbol{\theta}}(f^{-1}(\boldsymbol{\xi}) \mid \mathbf{y})$  refers to the untransformed posterior density with respect to  $\boldsymbol{\theta}$  evaluated for  $f^{-1}(\boldsymbol{\xi}) = \boldsymbol{\theta}$ .  $J_{f^{-1}}(\boldsymbol{\xi})$  denotes the Jacobian matrix with the element in the  $i$ -th row and  $j$ -th column given by  $\frac{\partial \theta_i}{\partial \xi_j}$ . Crucially, the posterior density with respect to  $\boldsymbol{\xi}$  retains the normalizing constant of the posterior density with respect to  $\boldsymbol{\theta}$ ; hence, one can select a convenient transformation without changing the normalizing constant. Note that in order to apply a transformation no new samples are required; instead the original samples can simply be transformed using the function  $f$ .

In principle, users can select transformations themselves. Nevertheless, the **bridgesampling** package comes with a set of built-in transformations (see Table 6.1), allowing the user to work with the model in a familiar parameterization. When the user then supplies a named vector with lower and upper bounds for the parameters (arguments `lb` and `ub`, respectively), the package internally transforms the relevant parameters and adjusts the expressions by the Jacobian term. Furthermore, as will be elaborated upon below, when the model is fitted in **Stan**, the **bridgesampling** package takes advantage of the rich class of **Stan** transformations.

The transformations built into the **bridgesampling** package are useful whenever each component of the parameter vector can be transformed separately.<sup>6</sup> In this scenario, there are four possible cases per parameter: (a) the parameter is unbounded; (b) the parameter has a lower bound (e.g., variance parameters); (c) the parameter has an upper bound; and (d) the parameter has a lower and an

<sup>6</sup>Thanks to a recent pull request by Kees Mulder, the **bridgesampling** package now also supports a more complicated case in which multiple parameters are constrained jointly (i.e., simplex parameters). This pull request also added support for circular parameters.

Table 6.1: Overview of built-in transformations in the **bridgesampling** package.  $l$  denotes a parameter lower bound and  $u$  denotes an upper bound.  $\Phi(\cdot)$  denotes the cumulative distribution function (cdf) and  $\phi(\cdot)$  the probability density function (pdf) of the normal distribution.

Type	Transformation	Inv.-Transformation	Jacobian Contribution
unbounded	$\xi_i = \theta_i$	$\theta_i = \xi_i$	$\left  \frac{\partial \theta_i}{\partial \xi_i} \right  = 1$
lower-bounded	$\xi_i = \log(\theta_i - l)$	$\theta_i = \exp(\xi_i) + l$	$\left  \frac{\partial \theta_i}{\partial \xi_i} \right  = \exp(\xi_i)$
upper-bounded	$\xi_i = \log(u - \theta_i)$	$\theta_i = u - \exp(\xi_i)$	$\left  \frac{\partial \theta_i}{\partial \xi_i} \right  = \exp(\xi_i)$
double-bounded	$\xi_i = \Phi^{-1}\left(\frac{\theta_i - l}{u - l}\right)$	$\theta_i = (u - l)\Phi(\xi_i) + l$	$\left  \frac{\partial \theta_i}{\partial \xi_i} \right  = (u - l)\phi(\xi_i)$

upper bound (e.g., probability parameters). As shown in Table 6.1, in case (a) the identity (i.e., no) transformation is applied. In case (b) and (c), logarithmic transformations are applied to transform the parameter to the real line. In case (d) a probit transformation is applied. Note that internally, the posterior density is automatically adjusted by the relevant Jacobian term. Since each component is transformed separately, the resulting Jacobian matrix will be diagonal. This is convenient since it implies that the absolute value of the determinant is the product of the absolute values of the diagonal entries of the Jacobian matrix:

$$|\det [J_{f^{-1}}(\boldsymbol{\xi})]| = \prod_{i=1}^p \left| \frac{\partial \theta_i}{\partial \xi_i} \right|. \quad (6.8)$$

Once all posterior samples have been transformed to the real line, a multivariate normal distribution is fitted using method-of-moments. On a side note, bridge sampling may underestimate the marginal likelihood when the same posterior samples are used both for fitting the proposal distribution and for the iterative updating scheme (i.e., Equation 6.6). Hence, as recommended by Overstall and Forster (2010), the **bridgesampling** package divides each MCMC chain into two halves, using the first half for fitting the proposal distribution and the second half for the iterative updating scheme.

## 6.2.2 Option II: Warping the Posterior Distribution

The second choice for the proposal distribution that is implemented in the **bridgesampling** package is a standard multivariate normal distribution in combination with a *warped* posterior distribution. The goal is still to match the posterior and the proposal distribution as closely as possible. However, instead of manipulating the proposal distribution, it is fixed to a standard multivariate normal distribution, and the posterior distribution is manipulated (i.e., warped). Crucially, the warped posterior density retains the normalizing constant of the original posterior density. The general methodology is referred to as Warp bridge sampling (Meng & Schilling, 2002).

There exist several variants of Warp bridge sampling; in the **bridgesampling** package, we implemented Warp-III bridge sampling (Gronau, Wagenmakers, et al., 2019; Meng & Schilling, 2002; Overstall, 2010) which can be used by setting `method = "warp3"`. This version matches the first three moments of the posterior and the proposal distribution. That is, in contrast to the simpler normal method described above, Warp-III not only matches the mean vector and the covariance matrix of the two distributions, but also the skewness. Consequently, when the posterior distribution is skewed, Warp-III may result in an estimator that is less variable. When the posterior distribution is symmetric, both Warp-III and the normal method should yield estimators that are about equally efficient. Hence, in principle, Warp-III should always provide estimates that are at least as precise as the normal method. However, the Warp-III method also takes about twice as much time to execute as the normal method; the reason for this is that Warp-III sampling results in a mixture density (for details, see Gronau, Wagenmakers, et al., 2019; Overstall, 2010) which requires that the unnormalized posterior density is evaluated twice as often as in the normal method.

Figure 6.2 illustrates the intuition for the warping procedure in the univariate case. The gray histogram in the top-left panel depicts skewed posterior samples, the solid black line the standard normal proposal distribution. The Warp-III procedure effectively standardizes the posterior samples so that they have mean zero (top-right panel) and variance one (bottom-right panel), and then attaches a minus sign with probability 0.5 to the samples which achieves symmetry (bottom-left panel). This intuition naturally generalizes to the multivariate case. Starting with posterior samples that can range across the entire real line (i.e.,  $\xi$ ) the multivariate Warp-III procedure is based on the following stochastic transformation:

$$\eta = \underbrace{b}_{\text{symmetry}} \times \underbrace{\mathbf{R}^{-1}}_{\text{covariance } \mathbf{I}} \times \underbrace{(\xi - \mu)}_{\text{mean } \mathbf{0}}, \quad (6.9)$$

where  $b \sim \mathcal{B}(0.5)$  on  $\{-1, 1\}$  and  $\mu$  corresponds to the expected value of  $\xi$  (i.e., the mean vector).<sup>7</sup> The matrix  $\mathbf{R}$  is obtained via the Cholesky decomposition of the covariance matrix of  $\xi$ , denoted as  $\Sigma$ , hence,  $\Sigma = \mathbf{R}\mathbf{R}^\top$ . Bridge sampling is then applied using this warped posterior distribution in combination with a standard multivariate normal distribution.

### 6.2.3 Estimation Error

Once the marginal likelihood has been estimated, the user can obtain an estimate of the estimation error in a number of different ways. One method is to use the `error_measures` function which is an S3 generic. Note that the `summary` method for objects returned by `bridge_sampler` internally calls the `error_measures` function and thus provides a convenient summary of the estimated log marginal likelihood and the estimation uncertainty. For marginal likelihoods estimated with the "normal" method and `repetitions = 1`, the `error_measures` function provides an approximate relative mean-squared error of the marginal likelihood estimate,

<sup>7</sup> $\mathcal{B}(\theta)$  denotes a Bernoulli distribution with success probability  $\theta$ .

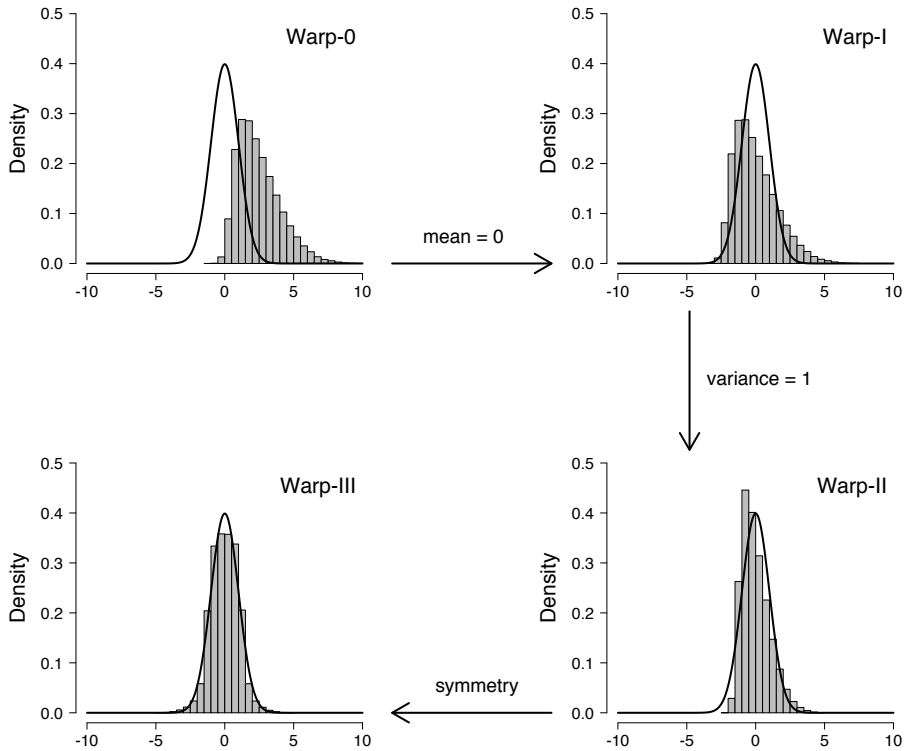


Figure 6.2: Illustration of the warping procedure. The black solid line shows the standard normal proposal distribution and the gray histogram shows the posterior samples. Available at <https://tinyurl.com/y7owvsz3> under CC license <https://creativecommons.org/licenses/by/2.0/> (see also Gronau, Heathcote, & Matzke, 2020; Gronau, Wagenmakers, et al., 2019).

an approximate coefficient of variation, and an approximate percentage error. The relative mean-squared error of the marginal likelihood estimate is given by:

$$\text{RE}^2 = \frac{\mathbb{E} \left[ (\hat{p}(\mathbf{y}) - p(\mathbf{y}))^2 \right]}{p(\mathbf{y})^2}. \quad (6.10)$$

The **bridgesampling** package computes an approximate relative mean-squared error of the marginal likelihood estimate based on the derivation by Frühwirth-Schnatter (2004) which takes into account that the samples from the proposal distribution are independent, whereas the samples from the posterior distribution may be autocorrelated (e.g., when using MCMC sampling procedures).

Under the assumption that the bridge sampling estimator  $\hat{p}(\mathbf{y})$  is unbiased, the square root of the relative mean-squared error (Equation 6.10) can be interpreted

as the coefficient of variation (i.e., the ratio of the standard deviation and the mean). To facilitate interpretation, the **bridgesampling** package also provides a percentage error which is obtained by simply converting the coefficient of variation to a percentage.

Note that the **error\_measures** function can currently not be used to obtain approximate errors for the "warp3" method with **repetitions** = 1. The reason is that, in our experience, the approximate errors appear to be unreliable in this case.

There are two further methods for assessing the uncertainty of the marginal likelihood estimate. These methods are computationally more costly than computing approximate errors, but are available for both the "normal" method and the "warp3" method. The first option is to set the **repetitions** argument of the **bridge\_sampler** function to an integer larger than one. This allows the user to obtain an empirical estimate of the variability across repeated applications of the method. Applying the **error\_measures** function to the output of the **bridge\_sampler** function that has been obtained with **repetitions** set to an integer large than one provides the user with the minimum/maximum log marginal likelihood estimate across repetitions and the interquartile range of the log marginal likelihood estimates. Note that this procedure assesses the uncertainty of the estimate conditional on the posterior samples, that is, in each repetition new samples are drawn from the proposal distribution, but the posterior samples are fixed across repetitions.

In case the user is able to easily draw new samples from the posterior distribution, the second option is to repeatedly call the **bridge\_sampler** function, each time with new posterior samples. This way, the user obtains an empirical assessment of the variability of the estimate which takes into account both uncertainty with respect to the samples from the proposal and also from the posterior distribution. If computationally feasible, we recommend this method for assessing the estimation error of the marginal likelihood.

After having outlined the underlying bridge sampling algorithm, we next demonstrate the capabilities of the **bridgesampling** package using three examples. Additional examples are available as vignettes at: <https://cran.r-project.org/package=bridgesampling>

## 6.3 Toy Example: Bayesian $T$ -test

We start with a simple statistical example: a Bayesian paired-samples  $t$ -test (Gronau, Ly, & Wagenmakers, 2020; Jeffreys, 1961; Ly et al., 2016b; Rouder, Speckman, Sun, Morey, & Iverson, 2009). We use R's **sleep** data set (Cushny & Peebles, 1905) which contains measurements for the effect of two soporific drugs on ten patients. Two different drugs were administered to the same ten patients and the dependent measure was the average number of hours of sleep gained compared to a control night in which no drug was administered. Figure 6.3 shows the increase in sleep (in hours) of the ten patients for each of the two drugs. To test whether the two drugs differ in effectiveness, we can conduct a Bayesian paired-samples  $t$ -test.

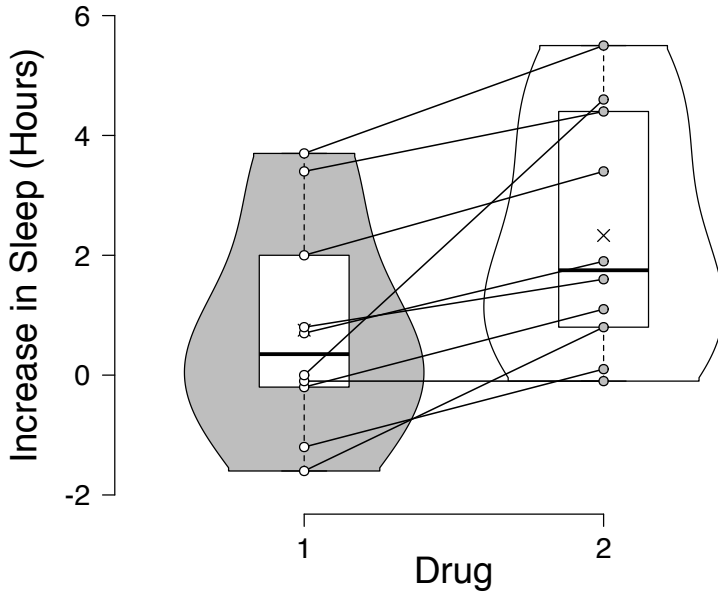


Figure 6.3: The sleep data set (Cushny & Peebles, 1905). The left violin plot displays the distribution of the increase in sleep (in hours) of the ten patients for the first drug, the right violin plot displays the distribution of the increase in sleep (in hours) of the ten patients for the second drug. Boxplots and the individual observations are superimposed. Observations for the same participant are connected by a line. Figure available at <https://tinyurl.com/yalskr23> under CC license <https://creativecommons.org/licenses/by/2.0/>.

The null hypothesis  $\mathcal{H}_0$  states that the  $n$  difference scores  $d_i$ ,  $i = 1, 2, \dots, n$ , where  $n = 10$ , follow a normal distribution with mean zero and variance  $\sigma^2$ , that is,  $d_i \sim \mathcal{N}(0, \sigma^2)$ . The alternative hypothesis  $\mathcal{H}_1$  states that the difference scores follow a normal distribution with mean  $\mu = \sigma\delta$ , where  $\delta$  denotes the standardized effect size, and variance  $\sigma^2$ , that is,  $d_i \sim \mathcal{N}(\sigma\delta, \sigma^2)$ . Jeffreys's prior is assigned to the variance  $\sigma^2$  so that  $p(\sigma^2) \propto 1/\sigma^2$  and a zero-centered Cauchy prior with scale parameter  $r = 1/\sqrt{2}$  is assigned to the standardized effect size  $\delta$  (for details, see Ly et al., 2016b; Morey & Rouder, 2015; Rouder et al., 2009).

In this example, we are interested in computing the Bayes factor  $\text{BF}_{10}$  which quantifies how much more likely the data are under  $\mathcal{H}_1$  (i.e., there is a difference between the two drugs) than under  $\mathcal{H}_0$  (i.e., there is no difference between the two drugs) by using the **bridgesampling** package. For this example, the Bayes factor can also be easily computed using the **BayesFactor** package (Morey & Rouder,

2015), allowing us to compare the results from the **bridgesampling** package to the correct answer.

The first step is to obtain posterior samples. In this example, we use **JAGS** in order to sample from the models. Here we focus on how to compute the log marginal likelihood for  $\mathcal{H}_1$ . The steps for obtaining the log marginal likelihood for  $\mathcal{H}_0$  are analogous. After having specified the model corresponding to  $\mathcal{H}_1$  as the character string `code_H1`, posterior samples can be obtained using the **R2jags** package (Su & Yajima, 2015) as follows:<sup>8</sup>

```
R> library("R2jags")
R> data("sleep")
R> y <- sleep$extra[sleep$group == 1]
R> x <- sleep$extra[sleep$group == 2]
R> d <- x - y # compute difference scores
R> n <- length(d)
R> set.seed(1)
R> jags_H1 <- jags(data = list(d = d, n = n, r = 1 / sqrt(2)),
+               parameters.to.save = c("delta", "inv_sigma2"),
+               model.file = textConnection(code_H1),
+               n.chains = 3, n.iter = 16000, n.burnin = 1000,
+               n.thin = 1)
```

Note the relatively large number of posterior samples; reliable estimates for the quantities of interest in testing usually necessitate many more posterior samples than are required for estimation. As a rule of thumb, we suggest that testing requires about an order of magnitude more posterior samples than estimation.

Next, we need to specify a function that take as input a named vector with parameter values and a data object, and returns the log of the unnormalized posterior density (i.e., the log of the integrand in Equation 6.4). This function is easily specified by inspecting the **JAGS** model. As a heuristic, one only needs to consider the model code where a “ $\sim$ ” sign appears. The log of the densities on the right-hand side of these “ $\sim$ ” symbols needs to be evaluated for the relevant quantities and then these log density values are summed.<sup>9</sup> Using this heuristic, we obtain the following unnormalized log posterior density function for  $\mathcal{H}_1$ :

```
R> log_posterior_H1 <- function(pars, data) {
+   delta <- pars["delta"] # extract parameter
+   inv_sigma2 <- pars["inv_sigma2"] # extract parameter
+   sigma <- 1 / sqrt(inv_sigma2) # convert precision to sigma
+   out <-
+   dcauchy(delta, scale = data$r, log = TRUE) + # prior
+   dgamma(inv_sigma2, 0.0001, 0.0001, log = TRUE) + # prior
+   sum(dnorm(data$d, sigma * delta,
+             sigma, log = TRUE)) # likelihood
```

<sup>8</sup>The complete code (including the **JAGS** models and the code for  $\mathcal{H}_0$ ) can be found in the supplemental material and also on the Open Science Framework: <https://osf.io/3yc8q/>.

<sup>9</sup>This heuristic assumes that the model does not include other random quantities that are generated during sampling, such as posterior predictives.

```
+   return(out)
+ }
```

The final step before we can compute the log marginal likelihoods is to specify named vectors with the parameter bounds:

```
R> lb_H1 <- rep(-Inf, 2)
R> ub_H1 <- rep(Inf, 2)
R> names(lb_H1) <- names(ub_H1) <- c("delta", "inv_sigma2")
R> lb_H1[["inv_sigma2"]] <- 0
```

The log marginal likelihood for  $\mathcal{H}_1$  can then be obtained by calling the `bridge_sampler` function as follows:

```
R> library("bridgesampling")
R> set.seed(12345)
R> bridge_H1 <- bridge_sampler(
+   samples = jags_H1,
+   log_posterior = log_posterior_H1,
+   data = list(d = d, n = n, r = 1 / sqrt(2)),
+   lb = lb_H1,
+   ub = ub_H1
+ )
```

We obtain:

```
R> print(bridge_H1)
```

```
Bridge sampling estimate of the log marginal likelihood: -27.17103
Estimate obtained in 5 iteration(s) via method "normal".
```

Note that by default, the "normal" bridge sampling method is used.

Next, we can use the `error_measures` function to obtain an approximate percentage error of the estimate:

```
R> error_measures(bridge_H1)$percentage

[1] "0.087%"
```

The small approximate percentage error indicates that the marginal likelihood has been estimated reliably. As mentioned before, we can use the `summary` method to obtain a convenient summary of the bridge sampling estimate and the estimation error. We obtain:

```
R> summary(bridge_H1)
```

```
Bridge sampling log marginal likelihood estimate
(method = "normal", repetitions = 1):
```

```
-27.17103
```

**Error Measures:**

Relative Mean-Squared Error: 7.564225e-07  
 Coefficient of Variation: 0.0008697255  
 Percentage Error: 0.087%

**Note:**

All error measures are approximate.

After having computed the log marginal likelihood estimate for  $\mathcal{H}_0$  in a similar fashion, we can compute the Bayes factor for  $\mathcal{H}_1$  over  $\mathcal{H}_0$  using the **bf** function:

```
R> bf(bridge_H1, bridge_H0)
```

Estimated Bayes factor in favor of bridge\_H1 over bridge\_H0: 17.26001

Hence, the observed data are about 17 times more likely under  $\mathcal{H}_1$  (which assigns the standardized effect size  $\delta$  a zero-centered Cauchy prior with scale  $r = 1/\sqrt{2}$ ) than under  $\mathcal{H}_0$  (which fixes  $\delta$  to zero). This is strong evidence for a difference in effectiveness between the two drugs (Jeffreys, 1939, Appendix I). The estimated Bayes factor closely matches the Bayes factor obtained with the **BayesFactor** package (i.e.,  $\text{BF}_{10} = 17.259$ ).

## 6.4 A “Black Box” Stan Interface

The previous section demonstrated how the **bridgesampling** package can be used to estimate the marginal likelihood for models coded in **JAGS**. For custom samplers, the steps needed to compute the marginal likelihood are the same. What is required is (a) an object with posterior samples; (b) a function that computes the log of the unnormalized posterior density; (c) the data; and (d) parameter bounds. A crucial step is the specification of the unnormalized log posterior density function. For applied researchers, this step may be challenging and error-prone, whereas for experienced statisticians it might be tedious and cumbersome, especially for complex models with a hierarchical structure.

In order to facilitate the computation of the marginal likelihood even further, the **bridgesampling** package contains an interface to the generic sampling software **Stan** (Carpenter et al., 2017). Assisted by the **rstan** package (Stan Development Team, 2016), this interface allows users to skip steps (b)-(d) above. Specifically, users who fit their models in **Stan** (in a way that retains the constants, as is detailed below) can obtain an estimate of the marginal likelihood by simply passing the **stanfit** object to the **bridge\_sampler** function.

The implementation of this “black box” functionality profited from the fact that, just as the **bridgesampling** package, **Stan**’s No-U-Turn sampler internally operates on unconstrained parameters (Hoffman & Gelman, 2014; Stan Development Team, 2017). The **rstan** package provides access to these unconstrained parameters and the corresponding log of the unnormalized posterior density. This

means that users can fit models with parameter types that have more complicated constraints than those currently built into **bridgesampling** (e.g., covariance/correlation matrices) without having to hand-code the appropriate transformations.

As mentioned above, in order to use the **bridgesampling** package in combination with **Stan** the models need to be implemented in a way that retains the constants. This can be achieved relatively easily: instead of writing, for instance,  $y \sim \text{normal}(\mu, \sigma)$  or  $y \sim \text{bernoulli}(\theta)$ , one needs to write

```
target += normal_lpdf(y | mu, sigma);
```

and

```
target += bernoulli_lpmf(y | theta);
```

That is, one starts with the fixed expression `target +=` which is then followed by the name of the distribution (e.g., `normal`). The name of the distribution is followed by `_lpdf` for continuous distributions and `_lpmf` for discrete distributions. Finally, in parentheses, there is the variable that was to the left of the “ $\sim$ ” sign (here, `y`), then a “`|`” sign, and finally the arguments of the distribution. This achieves that the user specifies the log target density (in this case, the log of the unnormalized posterior density) in a way that retains the constants of the involved distributions.

Note that in case the distributions are truncated, the user needs to code the correct renormalization. For instance, a normal distribution with upper truncation at `upper` is implemented as follows

```
target += normal_lpdf(y | mu, sigma) -  
          normal_lcdf(upper | mu, sigma);
```

where the function `normal_lcdf` yields the log of the cumulative distribution function (cdf) of the normal distribution. Likewise, a normal distribution with lower truncation at `lower` is obtained as

```
target += normal_lpdf(y | mu, sigma) -  
          normal_lccdf(lower | mu, sigma);
```

where `normal_lccdf` yields the log of the complementary cumulative distribution function (ccdf) of the normal distribution (i.e., the log of one minus the cumulative distribution function of the normal distribution). A normal distribution with lower truncation point `lower` and upper truncation point `upper` can be implemented as follows:

```
target += normal_lpdf(y | mu, sigma) -  
          log_diff_exp(normal_lcdf(upper | mu, sigma),  
                      normal_lcdf(lower | mu, sigma));
```

where `log_diff_exp(a, b)` is a numerically more stable version of the operation  $\log(\exp(a) - \exp(b))$ . Note that when implementing a truncated distribution, it is of course also important to give the variable of interest the correct bounds. For

instance, for the last example where  $y$  has a lower truncation at `lower` and an upper truncation at `upper` the variable  $y$  should be declared as<sup>10</sup>

```
real<lower = lower, upper = upper> y;
```

For more details about how to implement truncated distributions in **Stan** we refer the user to the **Stan** manual (Stan Development Team, 2017, section 5.3, “Truncated Distributions”).

In sum, the **bridgesampling** package enables users to obtain an estimate of the marginal likelihood for any **Stan** model (programmed to retain the constants) simply by passing the **stanfit** object to the **bridge\_sampler** function. Next we demonstrate this functionality using two prototypical examples in Bayesian model selection.

### 6.4.1 Stan Example 1: Bayesian GLMM

The first example features a generalized linear mixed model (GLMM) applied to the turtles data set (Janzen, Tucker, & Paukstis, 2000).<sup>11</sup> This data set is included in the **bridgesampling** package and contains information about 244 newborn turtles from 31 different clutches. For each turtle, the data set includes information about survival status (0 = died, 1 = survived), birth weight in grams, and clutch (family) membership (indicated by a number between one and 31). Figure 6.4 displays a scatterplot of clutch membership and birth weight. The clutches have been ordered according to mean birth weight. Dots indicate turtles who survived and red crosses indicate turtles who died. This data set has been analyzed in the context of Bayesian model selection before, allowing us to compare the results from the **bridgesampling** package to the results reported in the literature (e.g., Overstall & Forster, 2010; Sinharay & Stern, 2005).

Here we focus on the model comparison that was conducted in Sinharay and Stern (2005). The data set was analyzed using a probit regression model of the form:

$$\begin{aligned} y_i &\sim \mathcal{B}(\Phi(\alpha_0 + \alpha_1 x_i + b_{\text{clutch}_i})), & i = 1, 2, \dots, N \\ b_j &\sim \mathcal{N}(0, \sigma^2), & j = 1, 2, \dots, C, \end{aligned} \quad (6.11)$$

where  $y_i$  denotes the survival status of the  $i$ -th turtle (i.e., 0 = died, 1 = survived),  $x_i$  denotes the birth weight (in grams) of the  $i$ -th turtle,  $\text{clutch}_i \in \{1, 2, \dots, C\}$ ,  $i = 1, 2, \dots, N$ , indicates the clutch to which the  $i$ -th turtle belongs,  $C$  denotes the number of clutches, and  $b_{\text{clutch}_i}$  denotes the random effect for the clutch to which the  $i$ -th turtle belongs. Furthermore,  $\Phi(\cdot)$  denotes the cumulative distribution function (cdf) of the normal distribution. Sinharay and Stern (2005) investigated

<sup>10</sup>Note that we assumed that  $y$  is a scalar. In general,  $y$  could also be declared as a vector or an array in **Stan**. In this case, the term that is subtracted for renormalization would need to be multiplied by the number of elements of  $y$ . For example, for the case of an upper truncation and a vector  $y$  of length  $k$  the code would need to be changed to: `target += normal_lpdf(y | mu, sigma) - k * normal_lcdf(upper | mu, sigma);` For another example, see the code for “Stan Example 2”.

<sup>11</sup>Data were obtained from Overstall and Forster (2010) and made available in the **bridgesampling** package with permission from the original authors.

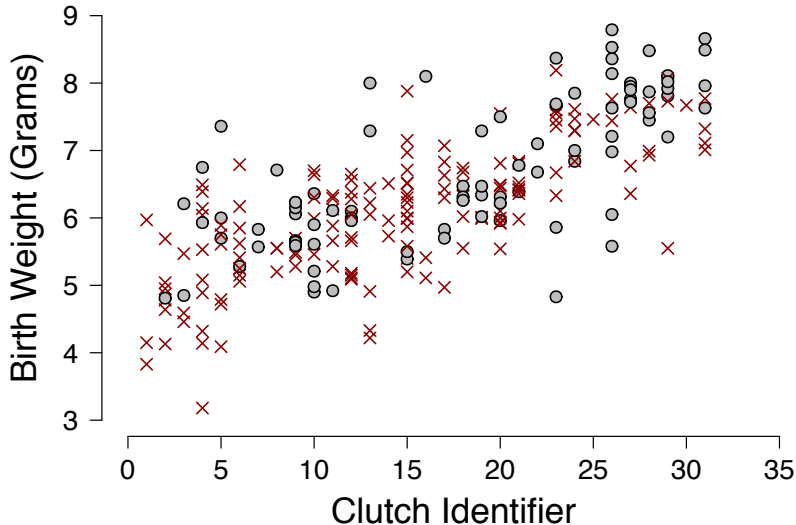


Figure 6.4: Data for 244 newborn turtles (Janzen et al., 2000). Birth weight is plotted against clutch membership. The clutches have been ordered according to their mean birth weight. Dots indicate turtles who survived and red crosses indicate turtles who died. Figure inspired by Sinharay and Stern (2005). Figure available at <https://tinyurl.com/yagfxrbw> under CC license <https://creativecommons.org/licenses/by/2.0/>.

the question whether there is an effect of clutch membership, that is, they tested the null hypothesis  $\mathcal{H}_0 : \sigma^2 = 0$ . The following priors were assigned to the model parameters:

$$\begin{aligned}\alpha_0 &\sim \mathcal{N}(0, 10), \\ \alpha_1 &\sim \mathcal{N}(0, 10), \\ p(\sigma^2) &= (1 + \sigma^2)^{-2}.\end{aligned}\tag{6.12}$$

Sinharay and Stern (2005) computed the Bayes factor in favor of the null hypothesis  $\mathcal{H}_0 : \sigma^2 = 0$  versus the alternative hypothesis  $\mathcal{H}_1 : p(\sigma^2) = (1 + \sigma^2)^{-2}$  using different methods and they reported a “true” Bayes factor of  $\text{BF}_{01} = 1.273$  (based on extensive numerical integration). Here we examine the extent to which we can reproduce the Bayes factor using the **bridgesampling** package.

After having implemented the Stan models as character strings `H0_code` and `H1_code`, the next step is to run Stan and obtain the posterior samples:<sup>12</sup>

---

<sup>12</sup>The complete code can be found in the supplemental material, on the Open Science Frame-

---

```
R> library("bridgesampling")
R> library("rstan")
R> data("turtles")
R> set.seed(1)
R> stanfit_H0 <- stan(model_code = H0_code,
+                   data = list(y = turtles$y,
+                   x = turtles$x, N = nrow(turtles)),
+                   iter = 15500, warmup = 500,
+                   chains = 4, seed = 1)
R> stanfit_H1 <- stan(model_code = H1_code,
+                   data = list(y = turtles$y,
+                   x = turtles$x, N = nrow(turtles),
+                   C = max(turtles$clutch),
+                   clutch = turtles$clutch),
+                   iter = 15500, warmup = 500,
+                   chains = 4, seed = 1)
```

With these Stan objects in hand, estimates of the log marginal likelihoods are obtained by simply passing the objects to the `bridge_sampler` function:

```
R> set.seed(1)
R> bridge_H0 <- bridge_sampler(stanfit_H0)
R> bridge_H1 <- bridge_sampler(stanfit_H1)
```

The Bayes factor in favor of  $\mathcal{H}_0$  over  $\mathcal{H}_1$  can then be obtained as follows:

```
R> bf(bridge_H0, bridge_H1)
```

Estimated Bayes factor in favor of bridge\_H0 over bridge\_H1: 1.27151

This value is close to that of 1.273 reported in Sinharay and Stern (2005). The data are only slightly more likely under  $\mathcal{H}_0$  than under  $\mathcal{H}_1$ , suggesting that the data do not warrant strong claims about whether or not clutch membership affects survival. The precision of the estimates for the marginal likelihoods can be obtained as follows:

```
R> error_measures(bridge_H0)$percentage
```

```
[1] "0.00972%"
```

```
R> error_measures(bridge_H1)$percentage
```

```
[1] "0.348%"
```

These error percentages indicate that both marginal likelihoods have been estimated accurately, but – as expected – the marginal likelihood for the more complicated model with random effects (i.e.,  $\mathcal{H}_1$ ) has the larger estimation error.

---

work (<https://osf.io/3yc8q/>), and is also available at `?turtles`. Note that the results are dependent on the compiler and the optimization settings. Thus, even with identical seeds results can differ slightly from the ones reported here.

### 6.4.2 Stan Example 2: Bayesian Factor Analysis

The second example concerns Bayesian factor analysis. In particular, we determine the number of relevant latent factors by implementing the Bayesian factor analysis model proposed by Lopes and West (2004). The model assumes that there are  $t$ ,  $t = 1, 2, \dots, T$ , observations on each of  $m$  variables. That is, each observation  $\mathbf{y}_t$  is an  $m$ -dimensional vector. The  $k$ -factor model – where  $k$  denotes the number of factors – relates each of the  $T$  observations  $\mathbf{y}_t$  to a latent  $k$ -dimensional vector  $\mathbf{f}_t$  which contains for observation  $t$  the values on the latent factors, as follows:<sup>13</sup>

$$\begin{aligned}\mathbf{y}_t \mid \mathbf{f}_t &\sim \mathcal{N}_m(\boldsymbol{\beta}\mathbf{f}_t, \boldsymbol{\Sigma}) \\ \mathbf{f}_t &\sim \mathcal{N}_k(\mathbf{0}_k, \mathbf{I}_k),\end{aligned}\tag{6.13}$$

where  $\boldsymbol{\beta}$  denotes the  $m \times k$  factor loadings matrix<sup>14</sup>,  $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2)$  denotes the  $m \times m$  diagonal matrix with residual variances,  $\mathbf{0}_k$  denotes a  $k$ -dimensional vector with zeros, and  $\mathbf{I}_k$  denotes the  $k \times k$  identity matrix. Hence, conditional on the latent factors, the observations on the  $m$  variables are assumed to be uncorrelated with each other. Marginally, however, the observations are usually not uncorrelated and they are distributed as

$$\mathbf{y}_t \sim \mathcal{N}_m(\mathbf{0}_m, \boldsymbol{\Omega}),\tag{6.14}$$

where  $\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}^\top + \boldsymbol{\Sigma}$ .

Here we reanalyze a data set that contains the changes in monthly international exchange rates for pounds sterling from January 1975 to December 1986 (West & Harrison, 1997, pp. 612–615). Currencies tracked are US Dollar (US), Canadian Dollar (CAN), Japanese Yen (JAP), French Franc (FRA), Italian Lira (ITA), and the (West) German Mark (GER). Figure 6.5 displays the data.<sup>15</sup> Using different computational methods, including bridge sampling, Lopes and West (2004) estimated the marginal likelihoods and posterior model probabilities for a factor model with one, two, and three factors. As before, this allows us to compare the results from the **bridgesampling** package to the results reported in the literature. To identify the model, the factor loading matrix  $\boldsymbol{\beta}$  is constrained to be lower-triangular (Lopes & West, 2004). The diagonal elements of  $\boldsymbol{\beta}$  are constrained to be positive by assigning them standard half-normal priors with lower truncation point zero:  $\beta_{jj} \sim \mathcal{N}(0, 1)_{T(0, \cdot)}$ ,  $j = 1, 2, \dots, k$ , and the lower-diagonal elements are assigned standard normal priors. The residual variances are assigned inverse-gamma priors of the form  $\sigma_i^2 \sim \text{Inverse-Gamma}(\nu/2, \nu s^2/2)$ ,  $i = 1, 2, \dots, m$ , where  $\nu = 2.2$  and  $\nu s^2 = 0.1$  (for details, see Lopes & West, 2004).

The first step in our reanalysis is to specify the Stan model as the character string `model.code`. We can then fit the three models corresponding to  $k = 1$ ,

---

<sup>13</sup>Note that the model assumes that the observations are zero-centered.

<sup>14</sup>We use the original notation by Lopes and West (2004) who denoted the factor loadings matrix with a lower-case letter. In the remainder of the chapter, matrices are denoted by upper-case letters.

<sup>15</sup>Each series has been standardized with respect to its sample mean and standard deviation. These standardized data are included in the **bridgesampling** package.

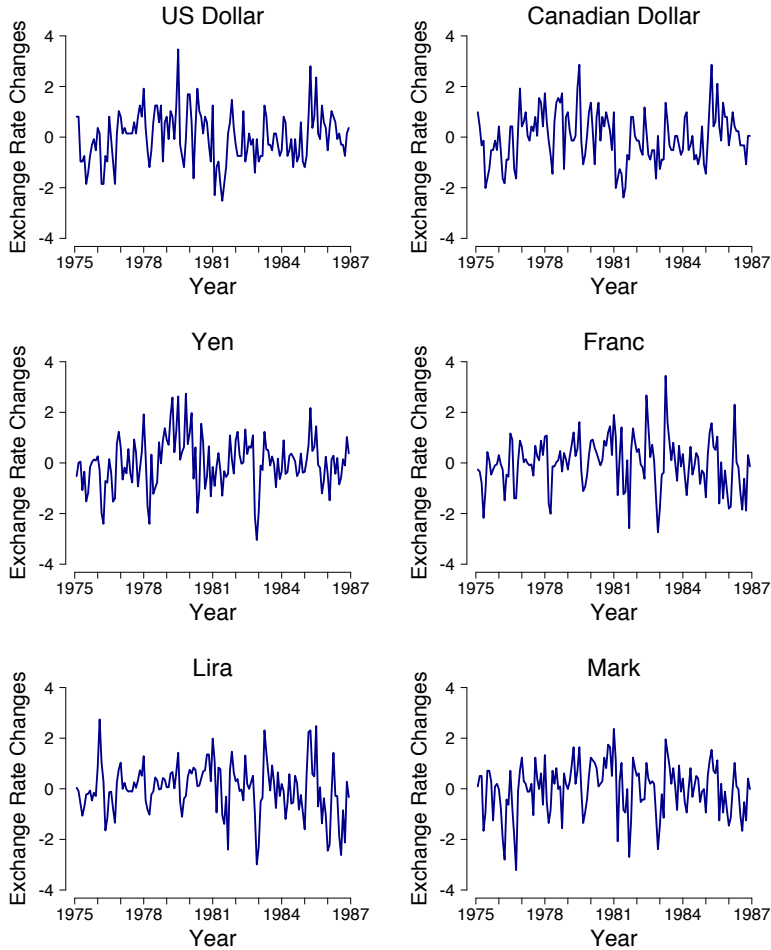


Figure 6.5: Changes in monthly international exchange rates for pounds sterling from January 1975 to December 1986 (West & Harrison, 1997, pp. 612–615). Currencies tracked are US Dollar (US), Canadian Dollar (CAN), Japanese Yen (JAP), French Franc (FRA), Italian Lira (ITA), and the (West) German Mark (GER). Each series has been standardized with respect to its sample mean and standard deviation. Figure reproduced from Lopes and West (2004). Figure available at <https://tinyurl.com/ybtdddyv> under CC license <https://creativecommons.org/licenses/by/2.0/>.

$k = 2$ , and  $k = 3$  latent factors and estimate the log marginal likelihoods using **bridgesampling** as follows:<sup>16</sup>

<sup>16</sup>The complete code can be found in the supplemental material, on the Open Science Framework (<https://osf.io/3yc8q/>), and is also available at [?ier](https://github.com/StanStan/StanStan). Note that we specify initial values

```
R> library("rstan")
R> library("bridgesampling")
R> data("ier")
R> cores <- 4
R> options(mc.cores = cores) # for parallel MCMC chains
R> model <- stan_model(model_code = model_code) # compile model
R> set.seed(1)
R> stanfit <- bridge <- vector("list", 3)
R> for (k in 1:3) {
+   stanfit[[k]] <- sampling(model,
+                             data = list(Y = ier, T = nrow(ier),
+                                           m = ncol(ier), k = k),
+                             iter = 11000, warmup = 1000, chains = 4,
+                             init = init_fun(nchains = 4, k = k,
+                                               m = ncol(ier)),
+                             cores = cores, seed = 1)
+   bridge[[k]] <- bridge_sampler(stanfit[[k]], method = "warp3",
+                                 repetitions = 10, cores = cores)
+ }
```

Note that in this example, we use the "warp3" method instead of the "normal" method. Furthermore, since the `error_measures` function cannot be used when the estimate has been obtained using `method = "warp3"` with `repetitions = 1`, we set `repetitions = 10` to obtain an empirical estimate of the estimation uncertainty (conditional on the posterior samples). We also select parallel computation by setting `cores = 4`. The `summary` method provides a convenient overview of the estimate and the estimation uncertainty. For instance, for the 2-factor model, we obtain as output:

```
R> summary(bridge[[2]])
```

```
Bridge sampling log marginal likelihood estimate
(method = "warp3", repetitions = 10):
```

```
-903.4522
```

```
Error Measures:
```

```
Min: -903.4565
```

```
Max: -903.4481
```

```
Interquartile Range: 0.002682305
```

---

using a custom `init_fun` function. This function may need to be changed for different applications. Furthermore, it is strongly advised to check that the chains have indeed converged since we sometimes encountered convergence issues with this model. Note that the results are dependent on the compiler and the optimization settings. Thus, even with identical seeds results can differ slightly from the ones reported here.

Table 6.2: Log marginal likelihood (logml) estimates for the  $k = 1$ ,  $k = 2$ , and  $k = 3$  factor model. The rightmost column displays the values based on bridge sampling reported in Lopes and West (2004).

Number of Factors	Median Logml	Min Logml	Max Logml	Lopes & West
$k = 1$	-1014.271	-1014.273	-1014.269	-1014.5
$k = 2$	-903.452	-903.457	-903.448	-903.7
$k = 3$	-905.271	-905.454	-905.138	$-\infty$

Note:

All error measures are based on 10 estimates.

Table 6.2 displays for each of the three factor models (i.e.,  $k = 1$ ,  $k = 2$ ,  $k = 3$ ) the median log marginal likelihood (logml) across repetitions, the minimum/maximum log marginal likelihood across repetitions, and the log marginal likelihood value reported in Lopes and West (2004) based on bridge sampling. Note that the negative infinity reported by Lopes and West (2004) might be due to a numerical problem. For the 1-factor model and the 2-factor model, the log marginal likelihoods obtained via **bridgesampling** are very similar to the ones reported in Lopes and West (2004). Furthermore, the narrow range of the estimates indicates that the estimation uncertainty is small (conditional on the posterior samples, as described above).

To examine the support for the three different models (i.e., different numbers of latent factors), we can use the `post_prob` function to compute posterior model probabilities. By default, the function assumes that all models are equally likely a priori; this can be adjusted using the `prior_prob` argument. Furthermore, the `model_names` argument can optionally be used to provide names for the models. Here we use the default of equal prior model probabilities and we obtain:

```
R> post_prob(bridge[[1]], bridge[[2]], bridge[[3]],
+           model_names = c("k = 1", "k = 2", "k = 3"))
```

```

           k = 1      k = 2      k = 3
[1,] 6.278942e-49 0.8435919 0.1564081
[2,] 6.309963e-49 0.8491811 0.1508189
[3,] 6.373407e-49 0.8554668 0.1445332
[4,] 6.511718e-49 0.8739641 0.1260359
[5,] 6.582895e-49 0.8805172 0.1194828
[6,] 6.384273e-49 0.8596401 0.1403599
[7,] 6.469723e-49 0.8736989 0.1263011
[8,] 6.403270e-49 0.8616183 0.1383817
[9,] 6.426132e-49 0.8635907 0.1364093
[10,] 6.417346e-49 0.8592737 0.1407263
```

Each row presents the posterior model probabilities based on one repetition of the bridge sampling procedure for all three models (i.e., each row sums to one). Hence,

there are as many rows as `repetitions`.<sup>17</sup> The 2-factor model receives most support from the observed data. This is in line with Lopes and West (2004), who also preferred the 2-factor model;<sup>18</sup> based on the factor loadings, they proposed the presence of a North American factor and a European Union factor.

## 6.5 Discussion

This chapter introduced **bridgesampling**, an R package for computing marginal likelihoods, Bayes factors, posterior model probabilities, and normalizing constants in general. We have demonstrated how researchers can use **bridgesampling** to conduct Bayesian model comparisons in a generic, user-friendly way: researchers need only provide posterior samples, a function that computes the log of the unnormalized posterior density, the data, and lower and upper bounds for the parameters. Furthermore, we have described the **Stan** interface which makes it even easier to obtain the marginal likelihood: researchers need only provide a **stanfit** object and the **bridgesampling** package will automatically produce an estimate of the log marginal likelihood.<sup>19</sup> In other words, the **bridgesampling** package makes it possible to obtain marginal likelihood estimates for any model that can be implemented in **Stan** (in a way that retains the constants). By combining the **Stan** state-of-the-art No-U-Turn sampler with **bridgesampling**, researchers are provided with a general purpose, easy-to-use computational solution to the challenging task of comparing complex Bayesian models.

As practical advice, we recommend to keep the following four points in mind when using the **bridgesampling** package (see also Gronau, Heathcote, & Matzke, 2020; Gronau, Wagenmakers, et al., 2019). First, one should always check the posterior samples carefully. A successful application of bridge sampling requires a sufficient number of representative samples from the posterior distribution. Thus, it is important to use efficient sampling algorithms and, in case of MCMC sampling, it is crucial that researchers confirm that the chains have converged to the joint posterior distribution. In addition, researchers need to make sure that the model does not contain any discrete parameters since those are currently not supported. This may sound more restrictive than it is. In practice the solution is to marginalize out the discrete parameters, something that is often possible. Note the similarity to **Stan** which also deals with discrete parameters by marginalizing them out (Stan Development Team, 2017, section 15). Furthermore, as demonstrated in the examples, for conducting model comparisons based on bridge sampling, the

---

<sup>17</sup>Note that the output of the `post_prob` function can be directly passed to the `boxplot` function which allows one to visualize the estimation uncertainty in the posterior model probabilities across repetitions.

<sup>18</sup>Note that Lopes and West (2004) report a posterior model probability of 1 for the 2-factor model. However, this estimate may be inflated by the infinite log marginal likelihood value for the 3-factor model.

<sup>19</sup>Similar to the **stanfit** method, the `bridge_sampler` method for **nimble** only requires the fitted object (of class `MCMC_refClass`) and extracts all necessary information for computing the marginal likelihood (including the function for computing the unnormalized log posterior density and the parameter bounds). However, at the time of writing we have not yet tested this method in the same intensity as the **stanfit** method. We will add a vignette describing the **nimble** interface in more detail when we have done so.

number of posterior samples often needs to be an order of magnitude larger than for estimation. This of course depends on a number of factors such as the complexity of the model of interest, the number of posterior samples that one usually uses for estimation, the posterior sampling algorithm used, and also the accuracy of the marginal likelihood estimate that one desires to achieve.

Second, one should always assess the uncertainty of the bridge sampling estimate. In case the uncertainty is deemed too high, one can attempt to achieve a higher precision by increasing the number of posterior samples or, in case `method = "normal"`, by using the more sophisticated `method = "warp3"` instead (see the third point below). Users of the **bridgesampling** package have different options for assessing the estimation uncertainty. In our opinion, the “gold standard” may be to obtain an empirical uncertainty assessment by repeating the bridge sampling procedure multiple times, each time using a fresh set of posterior samples. This approach allows users to assess the uncertainty directly for the quantity of interest. For instance, if the focus is on computing a Bayes factor, users may repeat the following steps: (a) obtain posterior samples for both models, (b) use the `bridge_sampler` function to estimate the log marginal likelihoods, (c) compute the Bayes factor using the `bf` function. The variability of these Bayes factor estimates across repetitions then provides an assessment of the uncertainty. For certain applications, this approach may be infeasible due to computational restrictions. If this is the case and `method = "normal"`, we recommend to use the approximate errors based on Frühwirth-Schnatter (2004) which are available through the `error_measures` function. As mentioned before, we have found these approximate errors to work well for `method = "normal"`, but not for `method = "warp3"` which is the reason why they are not available for the latter method. Alternatively, one can also assess the estimation uncertainty by setting the `repetitions` argument to an integer larger than one. This provides an assessment of the estimation uncertainty due to variability in the samples from the proposal distribution, but it should be kept in mind that this does not take into account variability in the posterior samples.

Third, one should consider whether using the more time-consuming Warp-III method may be beneficial. The accuracy of the estimate is governed not only by the number of samples, but also by the overlap between the posterior and the proposal distribution (e.g., Meng & Schilling, 2002; Meng & Wong, 1996). The **bridgesampling** package attempts to maximize this overlap by (a) focusing on one marginal likelihood at a time which allows one to use a convenient proposal distribution which closely resembles the posterior distribution, (b) using a proposal distribution which matches the mean vector and covariance matrix of the posterior samples (i.e., `method = "normal"`) or additionally also the skewness (i.e., `method = "warp3"`). Consequently, as mentioned before, `method = "warp3"` will always be as precise or more precise than `method = "normal"`; however, it also takes about twice as long. We have found that in many applications, `method = "normal"` works well, however, in case the posterior is skewed (crucially, this refers to the joint posterior of the quantities that have been transformed to the real line), `method = "warp3"` may be the better choice. When in doubt, we believe that it may be beneficial to also explore the Warp-III results – if this is computationally feasible – to see how much (if any) improvement in precision is achieved by taking

into account potential skewness. It should be kept in mind that, in case the posterior distribution exhibits multiple modes, the overlap of the two distributions may still be subject to improvement – even when using `method = "warp3"`. The development of efficient bridge sampling variants for these cases is subject to ongoing research (e.g., Frühwirth–Schnatter, 2004; L. Wang & Meng, 2016).

Forth, users should carefully think about the choice of prior distribution. Even though the **bridgesampling** package enables researchers to compute the marginal likelihoods in an almost black-box manner, this does not imply that the user can mindlessly exploit the package functionality to conduct Bayesian model comparisons. As is apparent from Equation 6.4, Bayesian model comparisons depend on the choice of the parameter prior distribution. Crucially, the prior distribution has a lasting influence on the results. Hence, meaningful Bayesian model comparisons require that researchers carefully consider their parameter prior distribution (e.g., Lee & Vanpaemel, 2018), engage in sensitivity analyses, or use default prior choices that have certain desirable properties such as model selection and information consistency (e.g., Bayarri et al., 2012; Jeffreys, 1961; Ly et al., 2016b).<sup>20</sup> Thus, the **bridgesampling** package removes the computational hurdle of obtaining the marginal likelihood, thereby allowing researchers to spend more time and effort on the specification of meaningful prior distributions.

It should also be kept in mind that there may be cases in which the bridge sampling procedure may not be the ideal choice for conducting Bayesian model comparisons. For instance, when the models are nested it might be faster and easier to use the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010). Another example is when the comparison of interest concerns a very large model space, and a separate bridge sampling based computation of marginal likelihoods may take too much time. In this scenario, Reversible Jump MCMC (Green, 1995) may be more appropriate. The downside of Reversible Jump MCMC is that it is usually problem-specific and cannot easily be applied in a generic fashion to different nested and non-nested model comparison problems (but see Gelling et al., 2017). The goal with the **bridgesampling** package, however, was exactly that: to provide users with a generic way of computing marginal likelihoods which can in principle be applied to any Bayesian model comparison problem.

In the future, we hope that it may be possible to add **bridgesampling** support for a number of R packages, such as the **MCMCglmm** package (Hadfield, 2010), the JAGS interface of the **mgcv** package (Wood, 2016), the **glmmBUGS** package (P. E. Brown & Zhou, 2018), or the **blavaan**<sup>21</sup> package (Merkle & Rosseel, 2018) so that users could conduct Bayesian model comparisons in a black box way similar to the **Stan** interface. For packages that use themselves **Stan** for fitting the models, adding **bridgesampling** support is relatively straightforward: the only potential change that would have to be implemented is to make sure that the models

---

<sup>20</sup>Note that in the first example (i.e., the Bayesian *t*-test) we have used prior distributions which lead to these desirable properties. However, in the second and third example, we simply used the prior distributions that have been used in the literature so that we could compare our results to the reported results.

<sup>21</sup>Note that the **blavaan** package already provides approximate marginal likelihoods for the models that are obtained via a Laplace approximation.

are coded such that all constants are retained (as explained in section 4). Once this is achieved, computing the relevant quantities via **bridgesampling** works as described in the **Stan** examples. For packages that do not use **Stan** to fit the models, the main difficulty is specifying the unnormalized posterior density function and the parameter bounds in an automatized way. This is also the reason why there is currently no black box interface to JAGS since, to the best of our knowledge, specifying these quantities in an automatized way is not trivial. Nevertheless, if this hurdle could be overcome, adding **bridgesampling** support would be straightforward.

In sum, the **bridgesampling** package provides a generic, accurate, easy-to-use, automatic, and fast way of computing marginal likelihoods and conducting Bayesian model comparisons. With the computational challenge all but overcome, researchers can spend more time and effort on addressing the conceptual challenge that comes with Bayesian model comparisons: specifying prior distributions that are either robust or meaningful.

Supplemental materials can be found at <https://www.jstatsoft.org/article/view/v092i10> and <https://osf.io/3yc8q/>.



## Part II

# Multi-Model Meta-Analysis



---

# Bayesian Mixture Modeling of Significant $P$ Values: A Meta-Analytic Method to Estimate the Degree of Contamination from $\mathcal{H}_0$

---

## Abstract

Publication bias and questionable research practices have long been known to corrupt the published record. One method to assess the extent of this corruption is to examine the meta-analytic collection of significant  $p$  values, the so-called  $p$ -curve (Simonsohn, Nelson, & Simmons, 2014a). Inspired by statistical research on false-discovery rates, we propose a Bayesian mixture model analysis of the  $p$ -curve. Our mixture model assumes that significant  $p$  values arise either from the null-hypothesis  $\mathcal{H}_0$  (when their distribution is uniform) or from the alternative hypothesis  $\mathcal{H}_1$  (when their distribution is accounted for by a simple parametric model). The mixture model estimates the proportion of significant results that originate from  $\mathcal{H}_0$ , but it also estimates the probability that each specific  $p$  value originates from  $\mathcal{H}_0$ . We apply our model to two examples. The first concerns the set of 587 significant  $p$  values for all  $t$ -tests published in the 2007 volumes of *Psychonomic Bulletin & Review* and the *Journal of Experimental Psychology: Learning, Memory, and Cognition*; the mixture model reveals that  $p$  values higher than about .005 are more likely to stem from  $\mathcal{H}_0$  than from

---

This chapter is published as Gronau, Q. F., Duizer, M., Bakker, M., & Wagenmakers, E.-J. (2017). Bayesian mixture modeling of significant  $p$  values: A meta-analytic method to estimate the degree of contamination from  $\mathcal{H}_0$ . *Journal of Experimental Psychology: General*, 146, 1223–1233. doi: <https://doi.org/10.1037/xge0000324>. A preprint is available at: <https://osf.io/mysbp/>

$\mathcal{H}_1$ . The second example concerns 159 significant  $p$  values from studies on social priming and 130 from yoked control studies. The results from the yoked controls confirm the findings from the first example, whereas the results from the social priming studies are difficult to interpret because they are sensitive to the prior specification. To maximize accessibility, we provide a web application that allows researchers to apply the mixture model to any set of significant  $p$  values.

## 7.1 Introduction

Psychological science is experiencing a crisis of confidence (e.g., Pashler & Wagenmakers, 2012). In response to this crisis, psychologists have offered new guidelines for journals (e.g., Nosek et al., 2015), started large-scale replication initiatives (e.g., Open Science Collaboration, 2015), promoted preregistration (e.g., Chambers, 2013, 2015; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), suggested different statistical reporting practices (e.g., Eich, 2014), and developed novel statistical techniques (e.g., Francis, 2013; Guan & Vandekerckhove, 2016; Simonsohn et al., 2014a; van Assen, van Aert, & Wicherts, 2015).

Among the various newly developed statistical techniques, the  $p$ -curve procedure is of special interest (Simonsohn et al., 2014a; Simonsohn, Nelson, & Simmons, 2014b). This procedure considers a collection of significant  $p$  values and asks whether their distribution contains “evidential value”. This question can be answered because of the fact that, under  $\mathcal{H}_0$ , the distribution of significant  $p$  values is uniform (e.g., Becker, 1991). Hence, if the observed distribution of significant  $p$  values is relatively flat, the most likely explanation for the findings is *publication bias* (e.g., Rosenthal, 1979; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). In addition, when most observed  $p$  values are near .05 this indicates that the findings maybe have been the result of significance chasing (i.e., “ $p$ -hacking”; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Simonsohn et al., 2014a). In the presence of a true effect, however, the distribution of  $p$  values is right-skewed such that low  $p$  values occur more often than high  $p$  values. The current  $p$ -curve analysis conducts a classical hypothesis test on the observed  $p$  values and concludes that their distribution contains “evidential value” when it is judged to be right-skewed.

The classical  $p$ -curve analysis is a promising tool to obtain an overall impression about the presence of true effects. Here we offer a novel and complementary Bayesian analysis of the  $p$ -curve that approaches the problem from a slightly different angle. Similar to an analysis of false-discovery rates, our Bayesian method assumes that the observed significant  $p$  values may have originated from  $\mathcal{H}_0$  or  $\mathcal{H}_1$ . The method then estimates the overall rate of contamination from  $\mathcal{H}_0$ ; in addition, the method estimates the probabilities that each specific  $p$  value originates from  $\mathcal{H}_0$ . These estimates can help assess, on a continuous scale, the extent to which an empirical phenomenon or a larger field is based on  $p$  values that are spurious. Below we first outline the method and then apply it to two concrete examples.

## 7.2 A Bayesian Mixture Model for Significant $P$ Values

We depart from the assumption that the observed distribution of  $p$  values is comprised of two different kinds of  $p$  values: a set of  $p$  values that originates from the null hypothesis  $\mathcal{H}_0$  and a set of  $p$  values that originates from the alternative hypothesis  $\mathcal{H}_1$  representing true effects. Hence, the observed  $p$  value distribution will be a mixture of these two kinds of  $p$  values; in practice, we do not know which of the observed  $p$  values stem from  $\mathcal{H}_0$  and which stem from  $\mathcal{H}_1$ .

However, using techniques from Bayesian mixture modeling (Frühwirth-Schnatter, 2006), we can estimate (1) the overall “ $\mathcal{H}_0$  assignment rate”, that is, the proportion of  $p$  values that stem from  $\mathcal{H}_0$ ; and (2) the probability that any single  $p$  value originates from  $\mathcal{H}_0$ . The mixture model assumption states that the observed  $p$ -curve is the result of a combination of two distributions: a uniform distribution associated with  $\mathcal{H}_0$  and a right-skewed distribution associated with  $\mathcal{H}_1$ . Thus, the probability density function of the observed  $p$  value distribution can be written as

$$f(p_i) = \phi f_{\mathcal{H}_0}(p_i) + (1 - \phi) f_{\mathcal{H}_1}(p_i), \quad (7.1)$$

where  $p_i$  denotes a specific observed  $p$  value, and  $\phi \in [0, 1]$  is a mixing parameter that reflects the estimated proportion of studies originating from  $\mathcal{H}_0$  (i.e., “ $\mathcal{H}_0$  assignment rate”). Values of  $\phi$  near 1 indicate that the collection of studies are heavily contaminated by  $\mathcal{H}_0$ .

### 7.2.1 The Generative Model

Figure 7.1 provides an illustration of the proposed mixture model for observed  $p$  values. The assumed data-generating process is displayed from top to bottom. Panel A shows that  $p$  values originating from  $\mathcal{H}_0$  follow a uniform distribution. In practice, we mostly observe significant  $p$  values and hence, our model focuses on the part highlighted in blue, that is,  $p$  values smaller than .05. For statistical convenience, we first probit-transform the  $p$  values (e.g., Efron, 2012; Tamhane & Shi, 2009). As shown in panel B, the uniform distribution of  $p$  values under  $\mathcal{H}_0$  corresponds to a standard normal distribution of probit-transformed  $p$  values (i.e.,  $\Phi^{-1}(p_i) \mid \mathcal{H}_0 \sim N(0, 1)$ ).

Panel C of Figure 7.1 shows that under  $\mathcal{H}_1$ , the distribution of  $p$  values is right-skewed. However, the exact distribution of  $p$  values under  $\mathcal{H}_1$  is more complex than that under  $\mathcal{H}_0$  as it depends on several factors such as sample size and the values of population parameters that are relevant for the test statistic at hand (e.g., Becker, 1991). Furthermore, a given collection of observed  $p$  values will be comprised of studies with different sample sizes, different test statistics, and, potentially, different true effects; that is, there exists an unknown distribution of true effects such that the collection of observed  $p$  values is inherently heterogeneous. We therefore need to address the fact that the distribution of  $p$  values under  $\mathcal{H}_1$  is itself a combination of potentially many different distributions. In this chapter, we use a simple parametric form for the probitized  $p$  values under  $\mathcal{H}_1$ , namely a normal

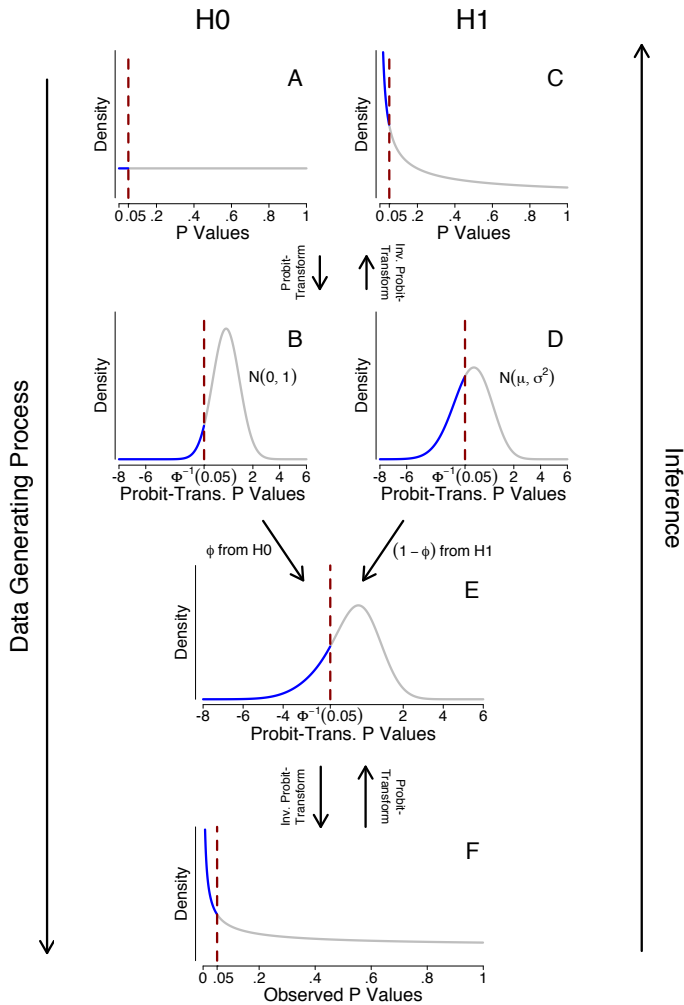


Figure 7.1: Illustration of the Bayesian mixture model for significant  $p$  values. The assumed data-generating process is displayed from top to bottom. Under  $\mathcal{H}_0$ ,  $p$  values are uniformly distributed (A) which corresponds to a standard normal distribution of the probit-transformed  $p$  values (B). Under  $\mathcal{H}_1$ , the distribution of  $p$  values is right-skewed (C) which we model on the probit scale using a normal distribution with unknown mean and standard deviation (D). The observed mixture distribution is obtained by taking a proportion  $\phi$  of  $p$  values from  $\mathcal{H}_0$ , the proportion  $(1 - \phi)$  from  $\mathcal{H}_1$  (E), and then applying the inverse probit-transformation (F). In practice, we start with the observed distribution of  $p$  values (F) and infer the model parameters by using Bayes' theorem to invert the generative model. Figure available at <http://tinyurl.com/zkxpx2> under CC license <https://creativecommons.org/licenses/by/2.0/>.

distribution with mean  $\mu$  and standard deviation  $\sigma$  (i.e.,  $\Phi^{-1}(p_i) \mid \mathcal{H}_1 \sim N(\mu, \sigma^2)$ ) which is shown in panel D of Figure 7.1.<sup>1</sup>

The next step in the data-generating process is that, after having specified the number of  $p$  values that we want to generate, we sample a proportion  $\phi$  of probit-transformed  $p$  values from  $\mathcal{H}_0$  and a proportion  $(1 - \phi)$  from  $\mathcal{H}_1$ . Combining the two samples yields a generated distribution of probit-transformed  $p$  values from the proposed mixture model. The final step is to apply the inverse probit-transformation which results in a generated distribution of  $p$  values.

In sum, in order to generate synthetic data from our mixture model one needs to determine the number of studies, the mixture proportion  $\phi$ , and the parameters of the normal distribution  $\mu$  and  $\sigma$  under  $\mathcal{H}_1$ . In practical applications, we do not know the parameters that govern the data-generating process. Instead, we only have a distribution of observed  $p$  values in hand and from this information we wish to infer quantities of interest. That is, instead of starting at the top of Figure 7.1 and working our way down (i.e., the data-generating perspective) we have to start at the bottom and move up (i.e., the inferential perspective). This allows us to decompose the observed  $p$  values into the ones that are likely to originate from  $\mathcal{H}_0$  and the ones who are likely to originate from  $\mathcal{H}_1$ .

## 7.2.2 Priors on the Model Parameters

In order to estimate the parameters of the mixture model we adopt a Bayesian approach (Diebolt & Robert, 1994; Frühwirth-Schnatter, 2006; Lee & Wagenmakers, 2013); this means that we specify our prior beliefs about the parameters of interest in the form of prior distributions and then update these by means of the observed data to yield posterior distributions. The posterior distributions reflect our beliefs about the parameters of interest after having seen the data.

For the  $\mathcal{H}_0$  assignment rate parameter  $\phi$ , we use a uniform prior on the interval  $[0, 1]$ . Furthermore, we need to specify priors for the mean and standard deviation of the normal distribution for the probit-transformed  $p$  values under  $\mathcal{H}_1$ . For the mean  $\mu$ , we use a truncated normal prior  $\mu \sim N(0, 1)_{T(-, 0)}$  (i.e., a folded standard normal that allows only negative values of  $\mu$ ). The truncation is imposed to reflect the fact that  $p$  values are expected to be smaller under  $\mathcal{H}_1$  than under  $\mathcal{H}_0$ ; the prior mean and standard deviation were chosen for simplicity and because they resulted in adequate performance across an extensive series of simulation studies. For the standard deviation  $\sigma$ , we use a uniform prior on the interval  $\sigma \in (0, 1)$ . The bounds on these parameters impose reasonable constraints on  $\mathcal{H}_1$ ; for instance, values of  $\sigma$  greater than 1 make the implausible prediction that  $p$  values near 1 are more common under  $\mathcal{H}_1$  than under  $\mathcal{H}_0$ .

As usual in Bayesian inference, the impact of the prior distributions lessens as sample size grows. In general, we recommend that researchers explore the sensitivity of the results to the prior choice for  $\mu$  (for instance, by changing the prior standard deviation from 1 to other plausible values). Researchers concerned about

<sup>1</sup>We also explored a non-parametric model which uses a flexible Bayesian procedure for density estimation (i.e., a *Dirichlet process mixture*). Unfortunately, the non-parametric model is harder to estimate and simulations suggest that it cannot easily be applied across sets of  $p$  values with different characteristics.

the performance of the model in repeated use may seek a Bayesian-frequentist compromise and calibrate the prior such that, for the sample size of interest, the mixture model yields good recovery of the contamination rate  $\phi$ .

### 7.3 Estimating the Model and Interpreting the Results

In order to estimate the model we use Markov chain Monte Carlo (MCMC) techniques implemented in the software program JAGS (Plummer, 2003) to draw samples from the posterior distributions of the model parameters (e.g., Gamerman & Lopes, 2006; Robert & Casella, 1999). After obtaining the posterior samples, we recommend the following three-step process.

**Step 1: Confirm Convergence.** Convergence can be confirmed visually (i.e., when the different Markov chains intermix) and by inspecting the  $\hat{R}$  statistic (Gelman & Rubin, 1992). Values of  $\hat{R}$  close to 1 indicate convergence, and values larger than 1.1 are often regarded as an indication of insufficient convergence. If the chains did not yet converge, it usually helps to increase the number of MCMC samples. Once convergence has been established it is safe to continue to the next step.

**Step 2: Confirm Quality of Fit.** Quality of fit can be assessed by plotting the observed  $p$  value quantiles against the quantiles of  $p$  values predicted by the model (i.e., a Q-Q plot). A perfect model fit results in a Q-Q plot that traces the main diagonal. When the quality of fit has been confirmed it is safe to continue to the next step.

**Step 3: Interpretation of Results.** The first quantity of interest is the overall  $\mathcal{H}_0$  contamination rate  $\phi$ ; the uncertainty about this parameter is reflected in a posterior distribution. This distribution can be summarized by a point (usually the mean, median, or mode) and by an interval; for a comprehensive assessment we recommend to inspect and report the entire posterior distribution. The second quantity of interest is, for each observed  $p$  value separately, the probability that it stems from  $\mathcal{H}_0$  – this is a single number and not a distribution.

To maximize accessibility of our procedure, we have developed a web application with an easy-to-use interface that allows researchers to apply the model to a set of significant  $p$  values without having to master a probabilistic programming language (<https://qfgronau.shinyapps.io/bmmssp/>). The supplemental material provides a detailed explanation of how to use our app. Furthermore, the supplemental material presents a set of simulation studies that highlight that the model is able to accurately estimate the quantities of interest under a relatively broad range of circumstances.<sup>2</sup> To illustrate the procedure and to show which conclusions can be drawn from the model output, we next present two example applications from the published literature.

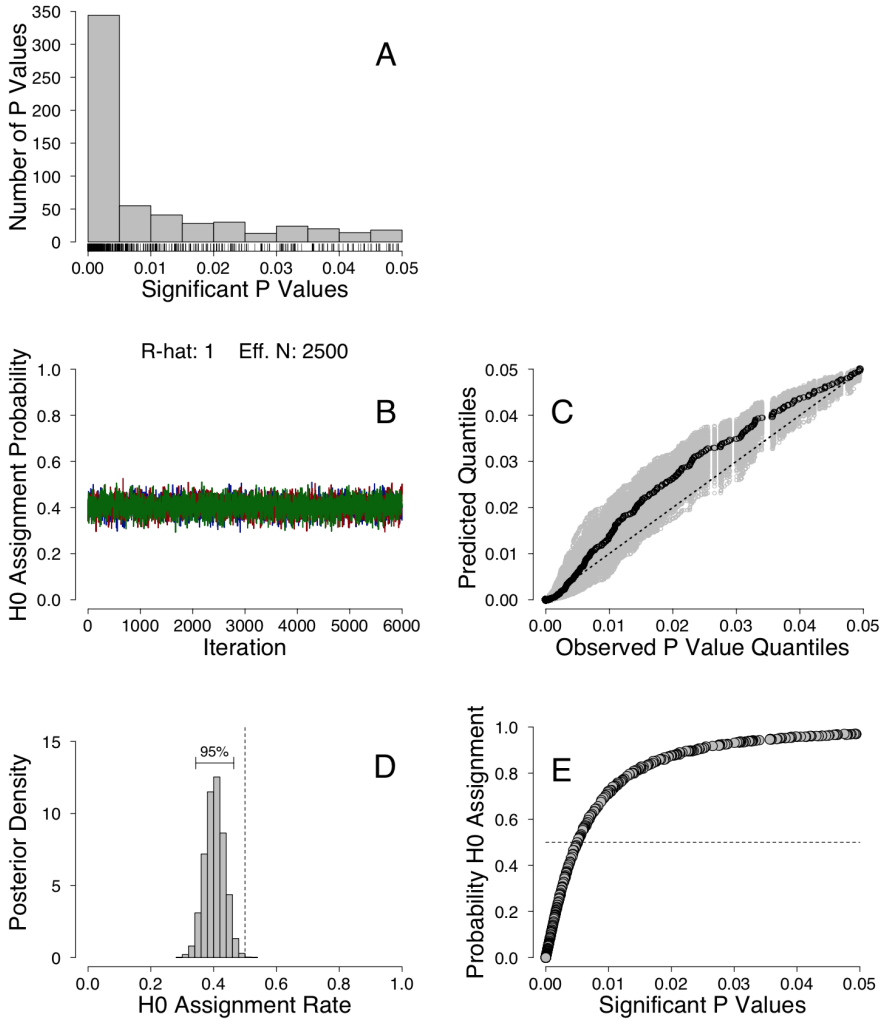


Figure 7.2: Application of the Bayesian mixture model to Example 1: 587  $t$ -test  $p$  values. Panel A: distribution of observed  $p$  values; panel B: traceplot of the MCMC chains for the  $\mathcal{H}_0$  assignment rate; panel C: Q-Q plot for comparing the observed  $p$  value distribution to the posterior predictive distribution; panel D: posterior distribution of the  $\mathcal{H}_0$  assignment rate; panel E: individual  $\mathcal{H}_0$  assignment probabilities. Figure available at <http://tinyurl.com/h8yx5h> under CC license <https://creativecommons.org/licenses/by/2.0/>.

## 7.4 Example 1: 587 $T$ -Test $P$ Values

For our first example we apply the model to a set of  $p$  values from Wetzels et al. (2011); these authors collected the results from all 855  $t$ -tests reported in the articles from the 2007 issues of *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Here we focus on the subset of 587  $p$  values that were significant.<sup>3</sup> It should be noted that these significant  $p$  values are inherently heterogeneous: they come from a wide range of empirical fields, and they were not screened for relevance. Thus, it is important to keep in mind that many of these  $p$  values may correspond to manipulation checks, and only a subset corresponds to the test of the key research hypothesis. That is, this is an undifferentiated set of  $p$  values which has not been selected in accordance with the guidelines proposed by Simonsohn et al. (2014a). Nevertheless, because of their heterogeneous nature, this set of  $p$  values provides a good test case for our model.

The results are presented in Figure 7.2. Panel A of Figure 7.2 shows the distribution of the 587 significant  $p$  values. The same distribution was inspected by Johnson (2013), who argued that the significant  $p$  values

“...presumably arise from two types of experiments: experiments in which a true effect was present and the alternative hypothesis was true, and experiments in which there was no effect present and the null hypothesis was true. For the latter experiments, the nominal distribution of  $P$  values is uniformly distributed on the range (0.0, 0.05) (...) The  $P$  values displayed in this plot thus represent a mixture of a uniform distribution and some other distribution. Even without resorting to complicated statistical methods to fit this mixture, the appearance of this histogram suggests that many, if not most, of the  $P$  values falling above 0.01 are approximately uniformly distributed. That is, most of the significant  $P$  values that fell in the range (0.01 – 0.05) probably represent  $P$  values that were computed from data in which the null hypothesis of no effect was true.”

Nevertheless, the overall distribution of  $p$  values is clearly right-skewed, and many  $p$  values are relatively low. The Markov chain Monte Carlo chains for the  $\mathcal{H}_0$  assignment rate  $\phi$  are shown in panel B and support the claim that the samples come from the posterior distribution. This is indicated by nicely intermixing chains and an  $\hat{R}$  value of 1.00.<sup>4</sup>

Panel C of Figure 7.2 shows the model fit by means of a Q-Q plot. The Q-Q plot allows a comparison between the distribution of observed  $p$  values and the distribution of posterior predictive  $p$  values, that is, the distribution of  $p$  values predicted by the model. Identical distributions yield a linear Q-Q plot

---

<sup>2</sup>The supplemental material, the example data sets that will be analyzed in the next sections, and all code that we used is available on the Open Science Framework: <https://osf.io/mysbp/>.

<sup>3</sup>We thank Valen Johnson for providing us with the 587 significant  $p$  values.

<sup>4</sup>This panel also shows the *effective samples size* which is an estimate of the number of independent samples obtained by applying a method (i.e., MCMC) that, by construction, produces samples that are not independent (i.e., autocorrelated).

with a slope of one. The black dots visualize the fit obtained by averaging across posterior samples, whereas the grey dots indicate the uncertainty in the Q-Q plot by displaying the results from individual draws from the posterior distribution. Although the Q-Q plot based on the averaged predicted distribution is not perfect, the uncertainty band suggests that the fit may be sufficiently acceptable to proceed to the interpretation stage.

Panel D of Figure 7.2 shows the posterior distribution of  $\phi$ , the  $\mathcal{H}_0$  assignment rate. This contamination rate is estimated to be near 0.4, and a Bayesian 95% highest density interval<sup>5</sup> ranges from 0.343 to 0.464, indicating a relatively high precision.

In addition to the estimation of the overall contamination rate, the Bayesian mixture model also allows us to estimate the probability that each individual  $p$  value is assigned to  $\mathcal{H}_0$ . These estimates are shown in panel E of Figure 7.2. The results indicate that for 41% of the observed  $p$  values the  $\mathcal{H}_0$  assignment probability is larger than .5; this means that, starting from a position of equipoise, for 41% of the observed significant  $p$  values it is more likely that they stem from  $\mathcal{H}_0$  than from  $\mathcal{H}_1$ . Similar to the qualitative conclusion drawn by Johnson (2013), the results suggest that  $p$  values between 0.01 – 0.05 are more likely to stem from  $\mathcal{H}_0$  than  $\mathcal{H}_1$ . Specifically,  $p$  values larger than about .005 are associated with a higher than 50%  $\mathcal{H}_0$  assignment rate.

To assess the robustness of the results to the prior choice for  $\mu$ , we examined how the results change as a function of the standard deviation for the truncated normal prior distribution for  $\mu$ ; in one analysis, we doubled the standard deviation to a value of 2; in another analysis, we halved the standard deviation to a value of 0.5. As detailed in the supplemental material, the results are robust to these changes.

## 7.5 Example 2: Social Priming Studies and Yoked Controls

For our second example we apply the model to a set of  $p$  values from social priming studies (e.g., Kahneman, 2011) and a matched set of  $p$  values from yoked control studies. To obtain the  $p$  values for the social priming studies we collected a large set of articles published by prominent researchers in the field of social priming. We used this selection method in order to preempt the critique that our results are biased by the inclusion of low-quality studies conducted, for instance, by novices or skeptics from unrelated fields. We followed the  $p$ -curve instructions from Simonsohn et al. (2014a) and distilled a single significant  $p$  value from each experiment. Every  $p$  value was evaluated by three raters; differences of opinion were rare and readily resolved by discussion. We believe the multi-rater method is advantageous as it furthers the use of a consistent selection policy and reduces the occurrence of erroneous selections. For our selection of  $p$  values, we did not record the interaction between the three raters; however, for future applications, it may be beneficial to document the selection process itself. An online table (<https://osf.io/344zz/>)

<sup>5</sup>A Bayesian 95% highest density interval is the shortest interval that captures 95% of the posterior mass.

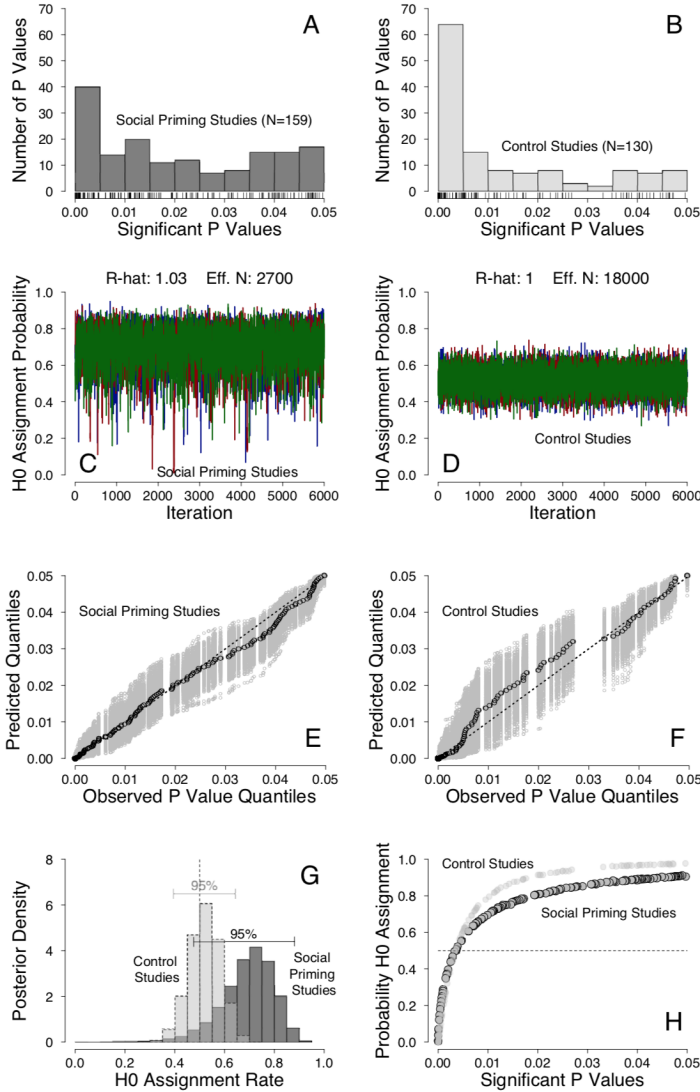


Figure 7.3: Application of the Bayesian mixture model to Example 2: social priming studies and yoked controls. First row: distributions of observed  $p$  values for the social priming studies (A) and the control studies (B); second row: traceplots of the MCMC chains for the  $\mathcal{H}_0$  assignment rate for the social priming studies (C) and control studies (D); third row: Q-Q plots for comparing the observed  $p$  value distribution to the posterior predictive distribution for the social priming studies (E) and control studies (F); panel G: posterior distributions of the  $\mathcal{H}_0$  assignment rate; panel H: individual  $\mathcal{H}_0$  assignment probabilities. Figure available at <http://tinyurl.com/gqj7c9e> under CC license <https://creativecommons.org/licenses/by/2.0/>.

identifies the selected  $p$  values by reporting the article, the experiment, the test statistic, and the  $p$  value. This information unambiguously identifies which  $p$  value we selected, making our analysis transparent and reproducible. Note that although we followed the guidelines by Simonsohn et al., we did not construct the  $p$  value disclosure table exactly in the form described in Simonsohn et al. (2014a). Although adding the specific reasons for the inclusion of each individual  $p$  value does not alter the statistical results in any way, we acknowledge that it is generally advisable to follow the Simonsohn et al. guidelines to the letter.

In addition, we sought to construct an appropriate comparison set of  $p$  values as a backdrop against which to evaluate the results for the social priming studies. This comparison set was constructed by selecting, for each social priming study under consideration, a yoked control study – that is, a study on a different topic and published in the same journal issue immediately after the social priming study. For each experiment in the yoked control studies, we distilled a single significant  $p$  value in the same manner as was done for the social priming studies. This procedure yielded a total of 159 significant social priming  $p$  values and 130 significant yoked control  $p$  values. Further details regarding the studies that were included are available at <https://osf.io/344zz/> (social priming studies) and <https://osf.io/4xgdz/> (control studies).

Figure 7.3 summarizes the results from applying the Bayesian mixture model. Panel A shows the distribution of  $p$  values for the social priming experiments, panel B the distribution of  $p$  values for the yoked controls. Although both distributions are right-skewed, the extent of this skew is much less pronounced than for the  $t$ -test  $p$  values from Example 1. Furthermore, the distribution of  $p$  values for the social priming studies shows less skew than that for the yoked control studies. Both distributions look relatively flat from .01 to .05.

The Markov chain Monte Carlo chains of the  $\mathcal{H}_0$  assignment rate  $\phi$  for the social priming  $p$  values are shown in panel C of the plot and support the claim that the samples come from the posterior distribution, indicated by nicely intermixing chains and an  $\hat{R}$  value of 1.03. The Markov chain Monte Carlo chains of the  $\mathcal{H}_0$  assignment rate  $\phi$  for the control  $p$  values are shown in panel D and suggest that these samples come from the posterior distribution as well ( $\hat{R} = 1.00$ ).

Panel E (social priming studies) and panel F (control studies) display the model fit by means of a Q-Q plot. As for the  $t$ -test example, the black dots provide a comparison of the observed  $p$  value distribution to the averaged predicted distribution and the grey dots represent the uncertainty. For both sets of  $p$  values the grey dots cover the dotted line that corresponds to a perfect fit, and hence we tentatively proceed to the stage of interpreting the parameter estimates.

Panel G of Figure 7.3 displays the posterior distributions of the  $\mathcal{H}_0$  assignment rate  $\phi$ . For the  $p$  values from the social priming studies, the degree of  $\mathcal{H}_0$  contamination appears to be substantial; the  $\mathcal{H}_0$  assignment rate has a 95% highest density interval that ranges from 0.475 to 0.880; for the yoked control  $p$  values, this interval ranges from 0.395 to 0.643.

Panel H of Figure 7.3 shows the  $\mathcal{H}_0$  assignment probabilities for the individual  $p$  values. These probabilities exceed 0.5 for 81% of the social priming  $p$  values and for 58% of the yoked control  $p$  values. Note that for the subset of  $p$  values between 0.01 – 0.05, the control studies have  $\mathcal{H}_0$  assignment rates that are actually

somewhat higher than those for the social priming studies. Nevertheless, for both control studies and social priming studies, the  $\mathcal{H}_0$  assignment probabilities for  $p$  values between 0.01 – 0.05 are high. As a side note, it is important to keep in mind that a high  $\mathcal{H}_0$  contamination rate does not necessarily imply that the underlying theories are false (see also Simonsohn et al., 2014a); however, it does suggest the need to change the experimental design and perhaps even the experimental paradigm.

To assess the robustness of the results to the prior choice for  $\mu$ , we conducted the same robustness check as in the previous example and examined how the results change as a function of the standard deviation for the truncated normal prior distribution for  $\mu$ ; in one analysis, we doubled the standard deviation to a value of 2; in another analysis, we halved the standard deviation to a value of 0.5. Results shown in the appendix suggest that for this example, a subset of the results is sensitive to the prior choice for the  $\mu$  parameter.

A visual inspection of the plots in the appendix suggests that the lack of robustness is particularly pronounced for the social priming studies. For these studies, the default prior setting had resulted in a 95% highest density interval for the  $\mathcal{H}_0$  assignment rate parameter  $\phi$  which ranged from 0.475 to 0.880. When the prior standard deviation for  $\mu$  is halved, the interval widens and ranges from 0.103 to 0.753; when it is doubled, the interval ranges from 0.564 to 0.906. For the yoked control studies, the changes are much less pronounced. For these studies, the default interval ranged from 0.395 to 0.643; when the prior standard deviation for  $\mu$  is halved, the interval ranges from 0.319 to 0.585; when the prior standard deviation for  $\mu$  is doubled, the interval ranges from 0.416 to 0.656.

In sum, for the yoked control studies the results are comparable to those obtained in the first example: the contamination rate is in the 40%-60% range, and significant  $p$  values larger than about .005 are more likely to be assigned to  $\mathcal{H}_0$  than to  $\mathcal{H}_1$ . For the social priming studies, the pattern is less clear. For two sets of prior distributions on  $\mu$  (i.e., the default and the one that doubles the standard deviation) the results suggest that the contamination rate is about 75%. However, when the standard deviation on  $\mu$  is halved, the posterior distribution on the contamination rate becomes very wide, and this lack of certainty is expressed through a  $\mathcal{H}_0$  assignment curve that is much less steep. In other words, the results for the social priming studies should be interpreted with extreme caution, and this underscores the importance of a sensitivity analysis. This naturally brings us to a discussion of the mixture model's limitations and challenges.

## 7.6 Challenges and Limitations

Although the mixture model is able to draw intuitive conclusions that are beyond the reach of existing methods, the procedure does come with three important caveats. First, estimating the parameters of the mixture model is an inherently difficult statistical problem. Because we are considering only the significant  $p$  values, all of the statistical action is in the tail of the distribution. The competing models  $\mathcal{H}_0$  and  $\mathcal{H}_1$  make relatively similar predictions with respect to this tail, and consequently a relatively large number of  $p$  values are required for the mixture

model to provide informative results.

It follows that one way to facilitate parameter estimation is to consider the complete set of  $p$  values and not just the significant ones. Until recently, the ubiquity of publication bias prevented this approach from yielding useful data; however, results from *Registered Reports* (Chambers, 2013; Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015) and *Registered Replication Reports* (e.g., Alogna et al., 2014; Cheung et al., 2016; Eerland et al., 2016; Wagenmakers, Beek, et al., 2016) are free from publication bias and unaffected by cherry-picking. In the future, data from these initiatives could be reanalyzed with a mixture model that considers all of the reported  $p$  values.

A second caveat is that, even when a reasonable number of  $p$  values are available, a change in the parameter priors might bring about a noticeably different result. We therefore recommend that researchers examine the robustness of the conclusions through a sensitivity analysis where the model is applied using various different standard deviations for the prior on the  $\mu$  parameter. In our experience, the results are even more strongly affected by the choice of the prior for the standard deviation  $\sigma$ . In particular, the model that restricts  $\sigma$  to range between zero and one can yield results different from the model that allows  $\sigma$  to span the entire positive part of the real line. However, in our opinion, the constraint that  $\sigma \in (0, 1)$  is reasonable and desirable; without this restriction, the model makes the implausible prediction that  $p$  values near 1 are more likely under  $\mathcal{H}_1$  than under  $\mathcal{H}_0$ . A complementary approach to choosing the prior distributions (especially for  $\mu$ ) is to use simulation studies to calibrate the priors for the sample size at hand, so as to achieve good recovery of the contamination rate.

The final caveat is that our approach uses a simple parametric form to account for the distribution of  $p$  values that stem from  $\mathcal{H}_1$ . Such simplicity comes with the risk of model-misspecification. Compared to a non-parametric model version that we explored in earlier work, the simple parametric version has the advantage that the model is easier to estimate; however, for specific sets of  $p$  values, the simple parametric distribution might not be able to accurately account for the complex distribution of  $p$  values originating from  $\mathcal{H}_1$ . This model-misspecification may be revealed by a non-acceptable model fit as reflected, for instance, by large deviations from the main diagonal in the Q-Q plot.

## 7.7 Concluding Comments

For studies that feature only a limited number of experiments, currently the sole arbiter of success is whether – for each experiment – the  $p$  value is lower than .05. This unfortunate state of affairs encourages publication bias, selective reporting, and questionable research practices (e.g., Barber, 1976). When studies are combined, however, the shape of the distribution of significant  $p$  values conveys additional information that allows one to estimate the degree of the bias. To this aim, a classical “ $p$ -curve” analysis method was recently proposed by Simonsohn et al. (2014a). Here we presented an alternative Bayesian analysis of the  $p$ -curve. Our Bayesian mixture model was inspired by a suggestion from Johnson (2013) and previous work on the control of false-discovery rates. The mixture model es-

estimates the extent to which the overall results have been contaminated by  $\mathcal{H}_0$ ; in addition, the method allows researchers to estimate how likely it is that a particular  $p$  value stems from  $\mathcal{H}_0$ . Note, however, that this estimate hinges on the context in which the particular  $p$  value was analyzed. This is true for mixture modeling in general: the estimated assignment probability for a certain observation depends on the values for the other observations.

Similar to the classical analysis method for  $p$ -curves, our model makes a number of assumptions. One assumption is that, under  $\mathcal{H}_0$ , the distribution of  $p$  values is uniform. In practice, this assumption may not hold; that is, particular forms of cherry-picking and questionable research practices may yield a  $p$ -curve that is right-skewed, thereby masquerading as the signature of a real effect. Hence, our model will mistakenly assign such  $p$  values to the mixture component corresponding to  $\mathcal{H}_1$ . Under this scheme, our mixture model contamination rate can be considered a lower bound on the true level of contamination from  $\mathcal{H}_0$ . However, there exist other forms of “ $p$ -hacking” and these may lead to left-skewed  $p$ -curves. In this case, our simulation studies – reported in the supplemental materials – suggest that the contamination rate is also underestimated. In general, if a literature is “ $p$ -hacked” to such an extent that it yields left-skewed  $p$ -curves, we recommend that our method should not be used.

We applied our mixture model to a set of significant  $p$  values from *Psychonomic Bulletin & Review* and the *Journal of Experimental Psychology: Learning, Memory, and Cognition* and also to a set of significant  $p$  values from social priming and yoked control studies. The examples highlighted the added inferential value of our model and caution against overinterpreting the evidential value of significant  $p$  values in the range from .01 – .05. In fact, fully consistent with the recommendation by Johnson (2013), our applications suggest that  $p$  values larger than .005 are more likely to be assigned to  $\mathcal{H}_0$  than to  $\mathcal{H}_1$ .

To maximize accessibility, we have provided an easy-to-use online application which allows researchers to apply our model in an intuitive way to any set of significant  $p$  values (<https://qfgronau.shinyapps.io/bmmssp/>). Furthermore, we provide the model code on the Open Science Framework (<https://osf.io/mysbp/>). This way we hope to encourage other researchers to apply the model within their field of interest.

The Supplemental Materials can be found at: <https://osf.io/mysbp/>.

## **7.A Prior Sensitivity Analysis for Example 2: Social Priming Studies and Yoked Controls**

A sensitivity analysis explored the effect of assigning different prior distributions to the  $\mu$  parameter. Figure 7.4 displays the results for a prior standard deviation of one half and Figure 7.5 shows the results for a prior standard deviation of two. The plots highlight that for this example, the results appear to be sensitive to the prior choice for the  $\mu$  parameter. See main text for details.

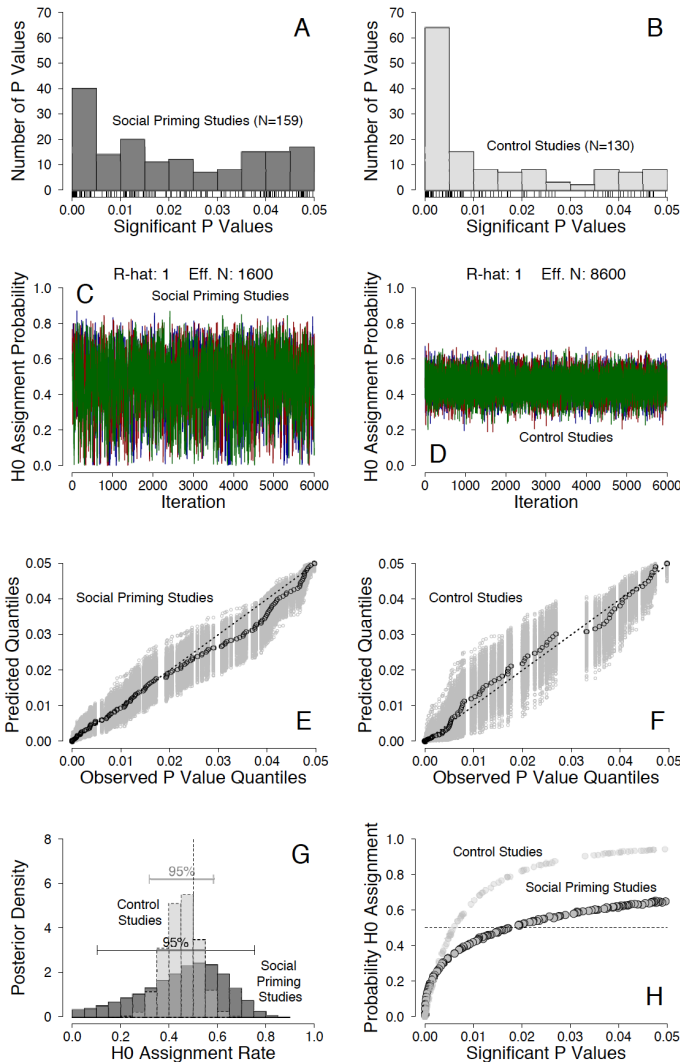


Figure 7.4: Sensitivity analysis for the application of the Bayesian mixture model to Example 2: social priming studies and yoked controls (prior standard deviation for  $\mu$  set to 0.5). First row: distributions of observed  $p$  values for the social priming studies (A) and the control studies (B); second row: traceplots of the MCMC chains for the  $\mathcal{H}_0$  assignment rate for the social priming studies (C) and control studies (D); third row: Q-Q plots for comparing the observed  $p$  value distribution to the posterior predictive distribution for the social priming studies (E) and control studies (F); panel G: posterior distributions of the  $\mathcal{H}_0$  assignment rate; panel H: individual  $\mathcal{H}_0$  assignment probabilities. Figure available at <http://tinyurl.com/jgyqn2g> under CC license <https://creativecommons.org/licenses/by/2.0/>.

## 7.A. Prior Sensitivity Analysis for Example 2: Social Priming Studies and Yoked Controls

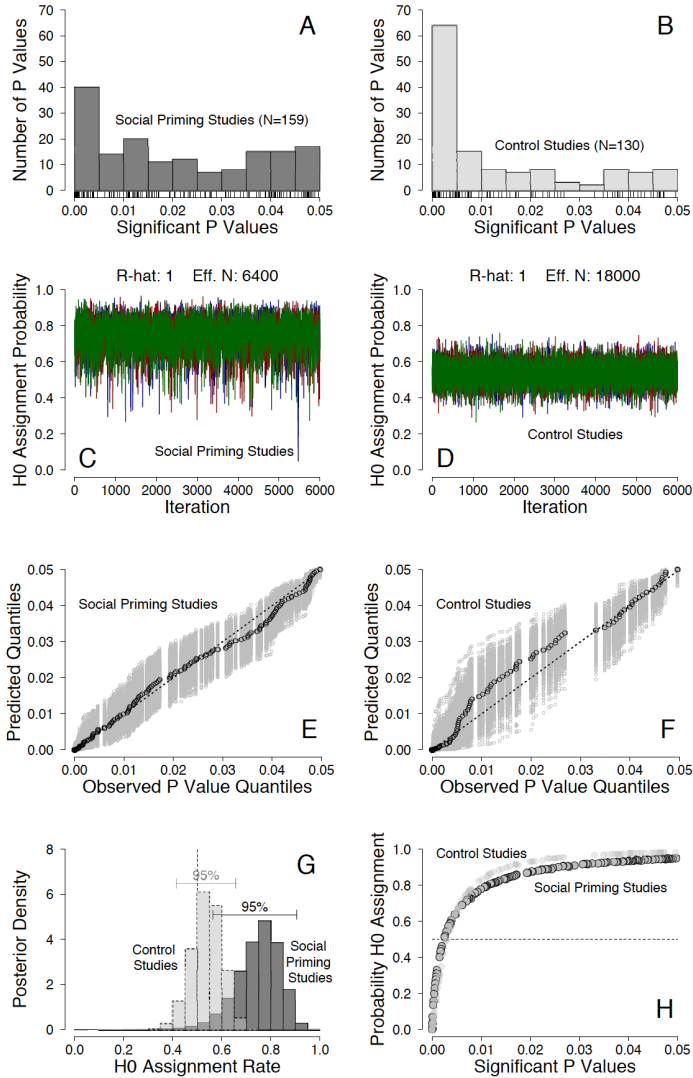


Figure 7.5: Sensitivity analysis for the application of the Bayesian mixture model to Example 2: social priming studies and yoked controls (prior standard deviation for  $\mu$  set to 2). First row: distributions of observed  $p$  values for the social priming studies (A) and the control studies (B); second row: traceplots of the MCMC chains for the  $\mathcal{H}_0$  assignment rate for the social priming studies (C) and control studies (D); third row: Q-Q plots for comparing the observed  $p$  value distribution to the posterior predictive distribution for the social priming studies (E) and control studies (F); panel G: posterior distributions of the  $\mathcal{H}_0$  assignment rate; panel H: individual  $\mathcal{H}_0$  assignment probabilities. Figure available at <http://tinyurl.com/huvlufn> under CC license <https://creativecommons.org/licenses/by/2.0/>.



---

# A Primer on Bayesian Model-Averaged Meta-Analysis

---

## Abstract

Meta-analysis is the predominant approach for quantitatively synthesizing a set of studies. If the studies themselves are of high quality, meta-analysis can provide valuable insights into the current scientific state of knowledge about a particular phenomenon. In psychological science, the most common approach is to conduct frequentist meta-analysis. In this primer, we discuss an alternative method, Bayesian model-averaged meta-analysis. This procedure combines the results of four Bayesian meta-analysis models: (1) fixed-effect null hypothesis, (2) fixed-effect alternative hypothesis, (3) random-effects null hypothesis, and (4) random-effects alternative hypothesis. These models are combined according to their plausibilities in light of the observed data to address the two key questions “Is the overall effect non-zero?” and “Is there between-study variability in effect size?”. Bayesian model-averaged meta-analysis therefore avoids the need to select either a fixed-effect or random-effects model and instead takes into account model uncertainty in a principled manner.

## 8.1 Introduction

Over the last decade, data collection in psychological science has become vastly more rigorous. Currently, experiments are often preregistered and the generally accepted best practice for investigating a particular effect is to conduct a many-labs Registered Report (e.g., Chambers, Munafò, & et al., 2013; Hagger et al., 2016; Klein et al., 2018; Landy et al., 2020; Wagenmakers, Beek, et al., 2016).

---

This chapter has been submitted for publication as Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2020). A primer on Bayesian model-averaged meta-analysis. Available as *PsyArXiv preprint*: <https://psyarxiv.com/97qup>

Although researchers now invest a lot of time and effort in preregistering their studies to ensure data of high quality, the way researchers analyze the resulting data has not changed markedly. Currently, the most popular analysis approach is still frequentist meta-analysis with  $p$ -values and confidence intervals (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Simons, Holcombe, & Spellman, 2014). Here we present a primer on an alternative method: Bayesian model-averaged meta-analysis (e.g., Gronau, van Erp, et al., 2017; Haaf, Hoogeveen, Berkhout, Gronau, & Wagenmakers, 2020; Hinne, Gronau, van den Bergh, & Wagenmakers, 2020; Hoogeveen, Wagenmakers, Kay, & Elk, 2018; Scheibehenne, Gronau, Jamil, & Wagenmakers, 2017; Vohs et al., under review). This method combines the results of Bayesian fixed-effect and Bayesian random-effects models according to the models' plausibilities in light of the data. Compared to the standard frequentist procedure, the Bayesian procedure affords researchers a number of pragmatic benefits (for a general introduction to Bayesian inference and its benefits, see the special issue in *Psychonomic Bulletin & Review*; Vandekerckhove, Rouder, & Kruschke, 2018). Specifically, the Bayesian procedure allows researchers to:

- assess the degree to which data make a claim more or less plausible. By quantifying evidence on a continuous scale, the Bayesian approach encourages more nuanced conclusions instead of all-or-none decisions. For instance, one may make statements of the form “compared to the effect-absent hypothesis, the data have made the effect-present hypothesis ten times more likely than it was before”.
- discriminate evidence of absence from absence of evidence. This enables researchers to disentangle whether there is evidence for the null hypothesis or whether the data are inconclusive. For instance, one may conclude that there is absence of evidence when the data support both the null hypothesis and the alternative hypothesis about equally. In meta-analysis, this scenario is most likely when the number of studies is small. Alternatively, one may conclude there is evidence of absence in case the data support the null hypothesis much more than the alternative hypothesis.
- update evidence and posterior distributions as experiments accumulate. This enables open-ended, sequential testing and estimation that is both efficient and ethical. For instance, if one planned to test 100 participants, but the evidence is already compelling after 50, one may stop data collection early. Similarly, researchers can update a Bayesian meta-analysis with data from new studies after the initial set has already been analyzed.
- make direct and intuitive statements concerning the plausibility of models and parameters. This enables a straightforward interpretation of the results. For instance, one may state that, based on the observed data, the alternative hypothesis receives probability .75 or that the probability is .50 that the effect size is between 0.1 and 0.3.
- include expert knowledge for more diagnostic tests. This enables the incorporation of expert knowledge not only in the design of a study, but also in the analysis of the resulting data. For instance, an expert may state that

the most likely effect size is 0.3, with 95% uncertainty interval ranging from 0.1 to 0.5. This can be incorporated in the analysis in form of an informed prior distribution for effect size. Robustness of the results can easily be checked by comparing the results to those obtained when using a default or less informative prior.

- model-average across fixed-effect and random-effects models which takes into account model uncertainty. This prevents overconfidence and allows for a graceful transition to more complicated models as data accumulate. For instance, when addressing the question whether the meta-analytic effect size is zero or not, model averaging allows one to take into account uncertainty with respect to whether there is heterogeneity in effect size across studies.

In this primer we provide an introduction to Bayesian model-averaged meta-analysis and we demonstrate the procedure using a concrete example from the literature. The goal of this primer is to (1) highlight the pragmatic benefits of a Bayesian model-averaged meta-analysis; (2) provide readers with the knowledge to correctly interpret the results of such an analysis; (3) demonstrate that applied researchers can straightforwardly conduct these analyses in practice using the R (R Core Team, 2019) package `metaBMA` (Heck, Gronau, & Wagenmakers, 2019) or JASP (JASP Team, 2020).

## 8.2 Bayesian Meta-Analysis

In Bayesian meta-analysis (e.g., Higgins, Thompson, & Spiegelhalter, 2009; Rouder, Haaf, Davis-Stober, & Hilgard, 2019; Rouder & Morey, 2011; T. C. Smith, Spiegelhalter, & Thomas, 1995; Sutton & Abrams, 2001), the most common approach is to use a *random-effects* model. Below, we first introduce the random-effects model and then outline hypotheses of interest about the model parameters.

### 8.2.1 The Random-Effects Model

In line with the frequentist meta-analysis procedure, Bayesian meta-analysis takes as input an observed effect size  $y_i$  and a corresponding standard error  $SE_i$ , for each study  $i = 1, 2, \dots, K$ . To accommodate studies with different dependent measures and designs, these effect sizes are typically standardized measures such as Cohen's  $d$  or Fisher's  $z$ . The random-effects model assumes that the observed effect size  $y_i$  is drawn from a normal distribution with mean equal to the latent true study effect  $\theta_i$  and standard deviation fixed to the observed  $SE_i$ . The latent study effects  $\theta_i$  are themselves drawn from a normal distribution, with mean given by the overall effect size  $\mu$  and standard deviation given by the between-study heterogeneity parameter  $\tau$ . This set-up is illustrated in Figure 8.1. The model parameters  $\mu$  and  $\tau$  are assigned prior distributions denoted by  $g(\cdot)$  and  $h(\cdot)$ , respectively (see Box 1 for recommendations on how to choose these prior distributions). In sum,

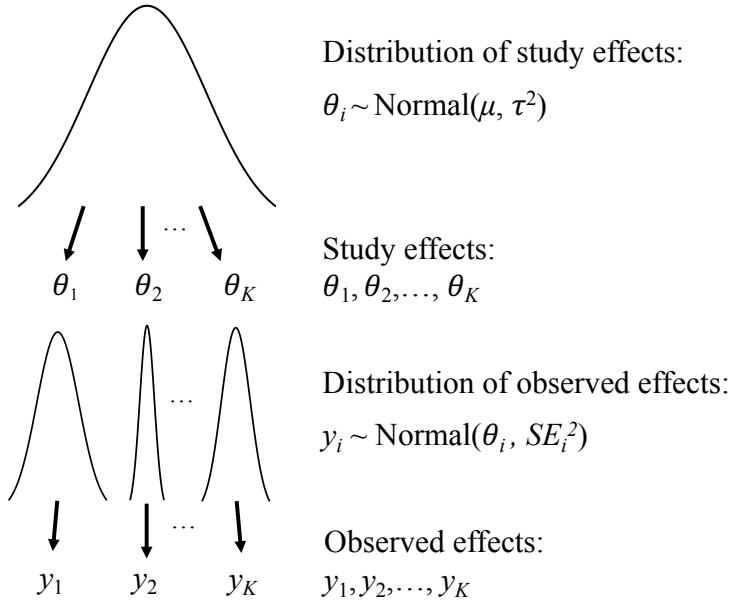


Figure 8.1: Meta-analytic random-effects model. The prior distributions for the overall effect size  $\mu$  and the between-study standard deviation  $\tau$  are not displayed. Available at <https://tinyurl.com/y7jgqyow> under CC license <https://creativecommons.org/licenses/by/2.0/>.

the model is specified as follows:

$$\begin{aligned}
 y_i &\sim \text{Normal}(\theta_i, SE_i^2) \\
 \theta_i &\sim \text{Normal}(\mu, \tau^2) \\
 \mu &\sim g(\cdot) \\
 \tau &\sim h(\cdot).
 \end{aligned}
 \tag{8.1}$$

Note that when the between-study standard deviation parameter  $\tau = 0$  the model implies that the effect for each study is identical and is equal to  $\mu$  (i.e., fixed-effect). In contrast, when  $\tau > 0$ , the model assumes that the latent true effect varies across studies (i.e., random-effects).

### 8.2.2 Limitations of the Random-Effects Model

Existing Bayesian meta-analysis procedures often focus on estimating the model-parameters  $\mu$  and  $\tau$  of the random-effects model (T. C. Smith et al., 1995; Stangl

## Box 1: Recommendations for Choosing the Parameter Prior Distributions

To apply the Bayesian model-averaged meta-analysis framework in practice, one needs to specify a prior distribution for the overall effect size  $\mu$  and the between-study standard deviation parameter  $\tau$ . Here we describe our approach to choosing these prior distributions when the considered effect size is a standardized mean difference (i.e., Cohen's  $d$  or Hedges'  $g$ ).<sup>a</sup> For the between-study standard deviation parameter  $\tau$ , we recommend an empirically informed prior distribution. This prior is based on the distribution of non-zero between-study standard deviation estimates for standardized mean difference effect sizes from meta-analyses reported in *Psychological Bulletin* in the years 1990–2013 (van Erp, Verhagen, Grasman, & Wagenmakers, 2017). Specifically, Gronau, van Erp, et al. (2017) approximated this empirical distribution by an Inverse-Gamma(1, 0.15) prior on  $\tau$  (see Figure 8.3). For the overall effect size parameter  $\mu$ , we recommend to consider both a “default” choice and an “informed” choice. By “default” we refer to a prior distribution that is (1) centered on zero, and (2) not overly narrow nor overly wide (Jeffreys, 1939; Lindley, 1957). We typically use a Cauchy prior with scale  $1/\sqrt{2} \approx 0.707$  (see Figure 8.3). This is the default choice for standardized mean differences in the **BayesFactor** package (Morey & Rouder, 2015). Nevertheless, other choices like a zero-centered normal prior also appear reasonable. By “informed” we refer to a prior distribution that is based on expert knowledge about the studied effect or based on a literature review. An informed prior is typically centered on a value different from zero to capture existing knowledge about effect size. Additionally, informed priors use expert knowledge to indicate the expected direction of an effect by truncating the prior distribution (e.g., practicing should increase memory performance). An example informed prior distribution is displayed in Figure 8.2. Considering both a “default” and “informed” prior for  $\mu$  serves as a robustness check: in case the results do not change qualitatively, the results are robust across different plausible prior choices. In case the results do change qualitatively, it needs to be accepted that the data may not be very informative and that the conclusion hinges on the prior specification. Another robustness check can be conducted by varying the width of the default prior on  $\mu$ .

<sup>a</sup>Other effect size measures are of course possible and can be easily analyzed using the referenced software. Nevertheless, the parameter prior distributions need to be adjusted for other effect size measures.

& Berry, 2000). Specifically, they focus on interpreting the posterior distribution and possibly summaries of the posterior distribution such as the mean, median, or 95% credible interval. However, simply fitting a random-effects model assumes that both  $\mu$  and  $\tau$  are non-zero – implying that there is an effect and heterogeneity in the effect across studies – and then focuses on estimating the size of  $\mu$  and  $\tau$ .

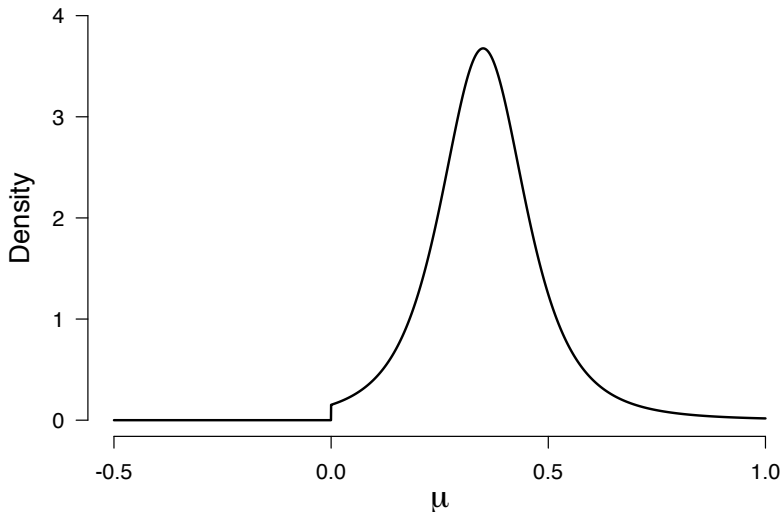


Figure 8.2: Example of an informed prior distribution for the overall effect size  $\mu$ : a  $t$  distribution with location 0.35, scale 0.102, and three degrees of freedom, truncated below at zero. This “Oosterwijk” prior (Gronau, Ly, & Wagenmakers, 2020) will be used later in the example. Available at <https://tinyurl.com/ycc965f2> under CC license <https://creativecommons.org/licenses/by/2.0/>.

Nevertheless, it has been argued that before one estimates a parameter, one should test whether there is anything to be estimated (i.e., testing whether a parameter is equal to zero should precede parameter estimation; Haaf, Ly, & Wagenmakers, 2019; Fisher, 1928, p. 274; Jeffreys, 1939, p. 345). Consequently, before estimating the parameters  $\mu$  and  $\tau$  one should address, in a principled manner, the two questions:

Q1: “Is the overall effect non-zero?”

Q2: “Is there between-study variability in effect size?”

Below we outline how to address these questions using Bayesian hypothesis testing in combination with Bayesian model averaging.<sup>1</sup> We have applied this framework to analyze power posing studies (Gronau, van Erp, et al., 2017), to investigate the effectiveness of descriptive social norms in facilitating ecological behavior (Scheibehenne et al., 2017), to test the compensatory control theory (Hoogeveen et al., 2018), to analyze facial feedback replication studies (Hinne et al., 2020), to

---

<sup>1</sup>Note that this framework does not preclude parameter estimation.

analyze how research results are influenced by subjective decisions that scientists make as they design studies (Landy et al., 2020), and to reanalyze the Many Labs 4 data (Haaf et al., 2020). Furthermore, we are currently applying this methodology to analyze a set of replication studies concerning the ego depletion effect (Vohs et al., under review).

### 8.2.3 Four Rival Hypotheses

Our Bayesian model-averaged meta-analysis framework considers four candidate hypotheses (e.g., Gronau, van Erp, et al., 2017; Scheibehenne et al., 2017).<sup>2</sup> These correspond to the four possibilities for fixing to zero either  $\mu$  or  $\tau$ , both, or neither:

1. the fixed-effect null hypothesis  $\mathcal{H}_0^f$ :  $\mu = 0$  ,  $\tau = 0$ ,
2. the fixed-effect alternative hypothesis  $\mathcal{H}_1^f$ :  $\mu \sim g(\cdot)$  ,  $\tau = 0$ ,
3. the random-effects null hypothesis  $\mathcal{H}_0^r$ :  $\mu = 0$ ,  $\tau \sim h(\cdot)$ ,
4. the random-effects alternative hypothesis  $\mathcal{H}_1^r$ :  $\mu \sim g(\cdot)$  ,  $\tau \sim h(\cdot)$ .

Figure 8.3 displays the differences in prior specification for the four hypotheses (each hypothesis corresponds to a separate row). Specifically, the first column displays the prior on the overall effect size  $\mu$  and the second column displays the prior on the between-study standard deviation  $\tau$ . For the hypotheses where the prior is not a point mass at zero, we have used the “default” prior recommendations from Box 1 (i.e., a zero-centered Cauchy prior with scale  $1/\sqrt{2}$  on  $\mu$  and an Inverse-Gamma(1, 0.15) prior on  $\tau$ ). The third column displays the implied joint prior on two hypothetical latent true study effects,  $\theta_i$  and  $\theta_j$ .<sup>3</sup> The fixed-effect null hypothesis  $\mathcal{H}_0^f$  fixes  $\mu$  and  $\tau$  to zero (Figure 8.3, row 1, column 1–2). Consequently, the true latent study effect is exactly zero for each study (Figure 8.3, row 1, column 3). The fixed-effect alternative hypothesis  $\mathcal{H}_1^f$  fixes  $\tau$  to zero (Figure 8.3, row 2, column 2) but allows  $\mu$  to differ from zero (i.e.,  $\mu$  is assigned a continuous prior distribution; Figure 8.3, row 2, column 1). Consequently, the latent true study effects can differ from zero. However, since  $\mathcal{H}_1^f$  does not specify any between-study variability (i.e.,  $\tau = 0$ ), all studies have the identical latent true effect size. Hence, the implied joint prior on two latent true study effects  $\theta_i$  and  $\theta_j$  assigns non-zero probability mass only to the diagonal line where  $\theta_i$  and  $\theta_j$  are identical (Figure 8.3, row 2, column 3). The random-effects null hypothesis  $\mathcal{H}_0^r$  fixes the overall effect size  $\mu$  to zero (Figure 8.3, row 3, column 1), but allows the between-study standard deviation  $\tau$  to differ from zero (i.e.,  $\tau$  is assigned a continuous prior distribution; Figure 8.3, row 3, column 2). Consequently, the latent true study effects may be different, but their distribution is centered on zero since the overall effect size  $\mu$  is fixed to zero (Figure 8.3, row 3, column 3). Finally, the random-effects alternative hypothesis  $\mathcal{H}_1^r$  allows both  $\mu$  and  $\tau$  to differ from zero (Figure 8.3, row 4, column 1–2). Consequently, each latent true study effect is unique. The latent true study

<sup>2</sup>The terms ‘hypothesis’ and ‘model’ are used interchangeably.

<sup>3</sup>Note that  $\theta_i$  and  $\theta_j$  correspond to two *latent* true study effects and do *not* refer to the observed effect sizes.

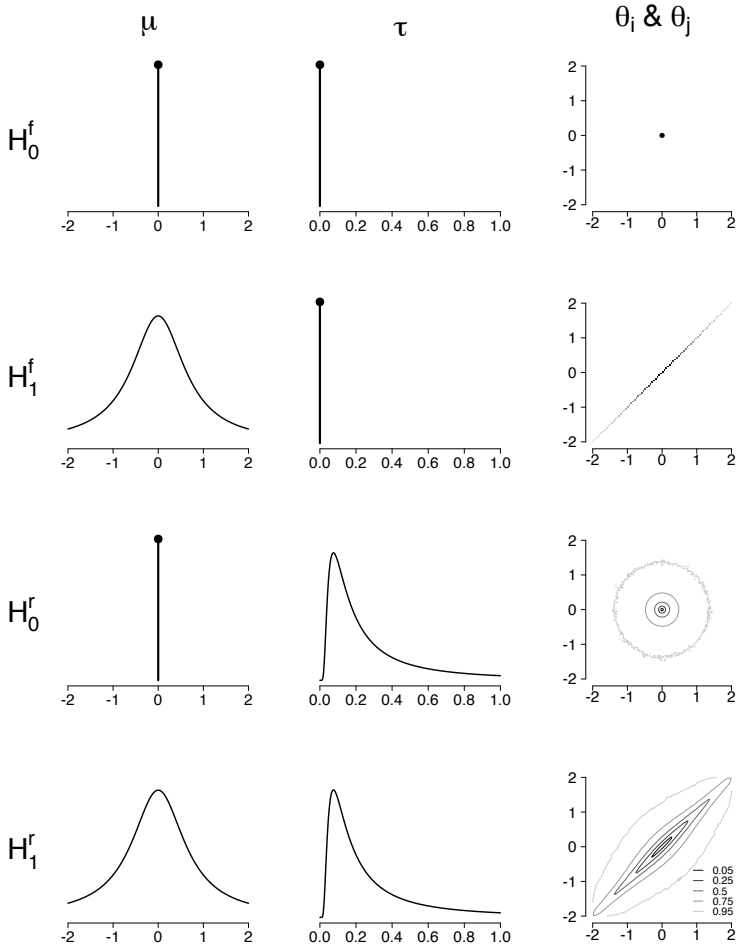


Figure 8.3: Parameter prior specifications for the four hypotheses of interest. Each row corresponds to one hypothesis (i.e.,  $\mathcal{H}_0^f$ ,  $\mathcal{H}_1^f$ ,  $\mathcal{H}_0^r$ , and  $\mathcal{H}_1^r$ ). The first column displays the prior distribution on the overall effect size  $\mu$  and the second column displays the prior distribution on the between-study standard deviation  $\tau$ . For the hypotheses where the prior is not a point mass at zero, we have used the “default” prior recommendations from Box 1 (i.e., a zero-centered Cauchy prior with scale  $1/\sqrt{2}$  on  $\mu$  and an Inverse-Gamma(1, 0.15) prior on  $\tau$ ). The third column displays the implied joint prior on two hypothetical latent true study effects,  $\theta_i$  and  $\theta_j$ . For the random-effects hypotheses the contours reflect 5%, 25%, 50%, 75%, and 95% of probability within the area. Available at <https://tinyurl.com/y98wqg5t> under CC license <https://creativecommons.org/licenses/by/2.0/>.

effects are correlated since their size depends on the specific values for  $\mu$  and  $\tau$ .

Hence, a priori, one latent true study effect being large implies that another one will likely also be large. The distribution of two hypothetical latent true study effects is still centered on zero since the prior on the overall effect  $\mu$  is centered on zero. However, the prior under  $\mathcal{H}_1^f$  spreads out its mass across a larger range of effect size values than the prior under  $\mathcal{H}_0^f$  since  $\mu$  is assigned a continuous prior that allows values other than zero.

### 8.2.4 Bayesian Hypothesis Testing

Each of the four rival hypotheses corresponds to one possible combination of the effect being present or absent and heterogeneity being present or absent. The goal is to assess the evidence for each of the four hypotheses by updating their plausibility in light of the observed data. Based on the shift in plausibility, one can then address Q1 and Q2 in a principled manner.

In the Bayesian framework, evidence for a model relative to another model is quantified using the Bayes factor (Etz & Wagenmakers, 2017; Jeffreys, 1935, 1961; Kass & Raftery, 1995; Wrinch & Jeffreys, 1921). For example, one may be interested in the evidence for the fixed-effect model with an effect versus the fixed-effect model with zero effect. The Bayes factor between these two models is

$$\underbrace{\text{BF}_{\mathcal{H}_1^f, \mathcal{H}_0^f}}_{\text{Bayes factor for effect}} = \frac{\underbrace{p(\text{data} \mid \mathcal{H}_1^f)}_{\text{Relative predictive accuracy}}}{\underbrace{p(\text{data} \mid \mathcal{H}_0^f)}_{\text{Relative predictive accuracy}}}, \quad (8.2)$$

where  $p(\text{data} \mid \mathcal{H})$  denotes how well a hypothesis  $\mathcal{H}$  predicted the data at hand. Therefore the Bayes factor may be interpreted as the *relative predictive accuracy* of two models (Rouder & Morey, 2019).

Here, we focus on an additional interpretation of the Bayes factor that comes from rearranging the terms of Bayes' rule. According to the additional interpretation the Bayes factor quantifies the change in beliefs about the hypotheses brought about by the data (i.e., the change from prior to posterior odds of two hypotheses):

$$\underbrace{\text{BF}_{\mathcal{H}_1^f, \mathcal{H}_0^f}}_{\text{Bayes factor for effect}} = \frac{\underbrace{p(\mathcal{H}_1^f \mid \text{data})}_{\text{Posterior odds for effect}}}{\underbrace{p(\mathcal{H}_0^f \mid \text{data})}_{\text{Posterior odds for effect}}} \bigg/ \frac{\underbrace{p(\mathcal{H}_1^f)}_{\text{Prior odds for effect}}}{\underbrace{p(\mathcal{H}_0^f)}_{\text{Prior odds for effect}}}. \quad (8.3)$$

In this equation,  $p(\mathcal{H}_1^f)$  denotes the prior probability of the fixed-effect alternative hypothesis  $\mathcal{H}_1^f$  and  $p(\mathcal{H}_1^f \mid \text{data})$  denotes the posterior probability of  $\mathcal{H}_1^f$  (i.e., after having updated one's knowledge based on observed data). Similarly,  $p(\mathcal{H}_0^f)$  denotes the prior probability of the fixed-effect null hypothesis  $\mathcal{H}_0^f$  and  $p(\mathcal{H}_0^f \mid \text{data})$  denotes the posterior probability of  $\mathcal{H}_0^f$ .<sup>4</sup>

To illustrate how to quantify change in beliefs using the Bayes factor we consider a hypothetical example. Figure 8.4 displays hypothetical prior and posterior

<sup>4</sup>Note that when comparing exactly two models, the prior probabilities do not affect the resulting Bayes factor as they cancel out (see Appendix).

probabilities for the four rival hypotheses. The top part of the plot shows prior probabilities of the hypotheses (i.e., plausibility before having seen any data), and by default all of them are set to .25. The bottom panel of Figure 8.4 displays hypothetical posterior probabilities of the hypotheses (i.e., plausibility after having updated one's knowledge based on observed data). In contrast to the prior probabilities, these are not equal anymore as the data have shifted one's beliefs.

We are now ready to calculate the Bayes factor from Equation 8.3. For the hypothetical example in Figure 8.4, the prior odds are given by  $.25/.25 = 1$  and the posterior odds are given by  $.40/.15 \approx 2.67$ . Consequently, the Bayes factor is  $BF_{\mathcal{H}_1^f, \mathcal{H}_0^f} \approx 2.67/1 = 2.67$  which indicates that – assuming a fixed-effect model – the data have made the effect-present hypothesis 2.7 times more likely than it was before, compared to the effect-absent hypothesis. In a similar fashion, one could compute  $BF_{\mathcal{H}_1^r, \mathcal{H}_0^r}$  to quantify the evidence for the effect being non-zero assuming random effects. The prior odds are again given by  $.25/.25 = 1$  and the posterior odds are given by  $.35/.10 = 3.5$ . Consequently, the Bayes factor is  $BF_{\mathcal{H}_1^r, \mathcal{H}_0^r} = 3.5/1 = 3.5$  which indicates that – assuming a random-effects model – the data have made the effect-present hypothesis 3.5 times more likely than it was before, compared to the effect-absent hypothesis.

To address the question whether or not there is heterogeneity in the effect across studies (Q2; i.e., test for fixed-effect or random-effects) one may compute  $BF_{\mathcal{H}_1^r, \mathcal{H}_1^f}$ . This Bayes factor compares the random-effects hypothesis to the fixed-effect hypothesis under the assumption that effect size  $\mu$  is non-zero. For the hypothetical example in Figure 8.4, the prior odds are given by  $.25/.25 = 1$  and the posterior odds are given by  $.35/.40 = 0.875$ . Consequently,  $BF_{\mathcal{H}_1^r, \mathcal{H}_1^f} = (.35/.40)/1 = 0.875$  or, equivalently,  $BF_{\mathcal{H}_1^f, \mathcal{H}_1^r} = 1/BF_{\mathcal{H}_1^r, \mathcal{H}_1^f} \approx 1.14$ . This Bayes factor indicates that – assuming that an effect is present – the data have made the heterogeneity-absent hypothesis about 1.14 times more likely than it was before, compared to the heterogeneity-present hypothesis.

### 8.2.5 Bayesian Model Averaging

For the fictional scenario above, one could conclude that the Bayes factor in favor of the effect-present hypothesis is either  $BF_{\mathcal{H}_1^r, \mathcal{H}_0^r} = 3.5$  (if there is heterogeneity in the effect) or  $BF_{\mathcal{H}_1^f, \mathcal{H}_0^f} \approx 2.67$  (if there is no heterogeneity). Furthermore, the data support both the random-effects alternative hypothesis and the fixed-effect alternative hypothesis about equally (i.e., assuming an effect,  $BF_{\mathcal{H}_1^f, \mathcal{H}_1^r} \approx 1.14$ ). Hence, considerable uncertainty remains with respect to whether a fixed-effect or a random-effects model is more appropriate. Instead of ignoring this uncertainty for final inference, one can take this uncertainty into account by considering all four hypotheses simultaneously according to their plausibility in light of the observed data. This procedure is known as Bayesian model averaging (e.g., Hinne et al., 2020; Hoeting et al., 1999).

To quantify the evidence for the effect being present while taking into account uncertainty with respect to choosing a fixed-effect or random-effects model, one can compute a model-averaged *inclusion Bayes factor*. This Bayes factor contrasts *all* hypotheses that allow the effect to be non-zero (i.e.,  $\mathcal{H}_1^f$  and  $\mathcal{H}_1^r$ ) to *all* hypotheses

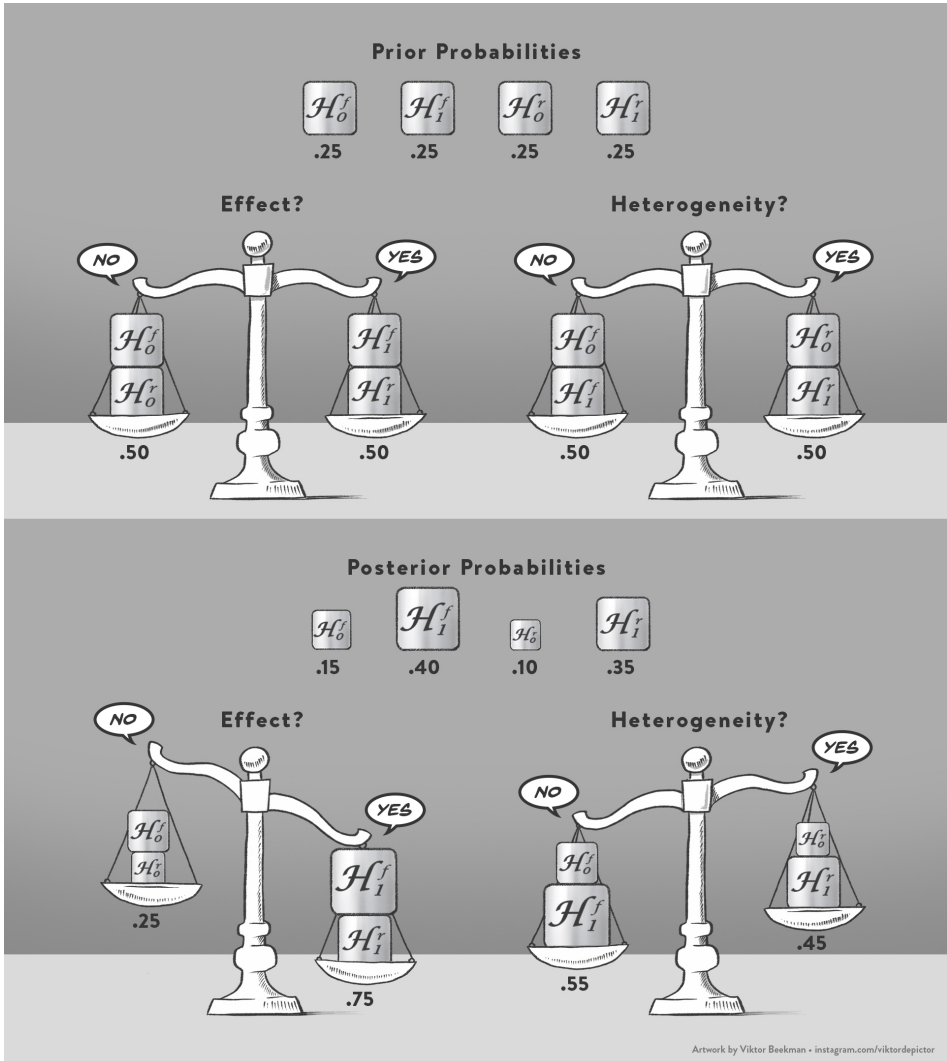


Figure 8.4: Prior probabilities of the hypotheses and computation of the model-averaged prior inclusion odds (top panel), and exemplary posterior probabilities and computation of the model-averaged posterior inclusion odds (bottom panel). Available at <https://www.bayesianspectacles.org/library/> under CC license <https://creativecommons.org/licenses/by/2.0/>.

that constrain the effect to be exactly zero (i.e.,  $\mathcal{H}_0^f$  and  $\mathcal{H}_0^r$ ) and thus fully takes into account model uncertainty with respect to choosing a fixed-effect or random-effects model.<sup>5</sup> Figure 8.4 illustrates how this model-averaged inclusion Bayes

<sup>5</sup>The term “inclusion” Bayes factor refers to the fact that it contrasts all hypotheses that

factor is computed. This Bayes factor, just as any Bayes factor, is given by the change from prior to posterior odds. However, this time, these are prior and posterior *inclusion* odds. The top panel of Figure 8.4 displays the prior probabilities of the hypotheses. By default all of them are set to .25. The left scale shows how to compute the prior inclusion odds for the presence of an effect. Specifically, the hypotheses that allow  $\mu$  to differ from zero (i.e.,  $\mathcal{H}_1^r$  and  $\mathcal{H}_1^f$ ) are contrasted with the hypotheses that fix  $\mu$  to zero (i.e.,  $\mathcal{H}_0^r$  and  $\mathcal{H}_0^f$ ). Since the combined prior probability of the hypotheses that allow  $\mu$  to differ from zero is .50 and the combined prior probability of the hypotheses that fix  $\mu$  to zero is also .50, the prior inclusion odds are equal to one.<sup>6</sup> The bottom panel of Figure 8.4 illustrates how to compute the posterior inclusion odds based on hypothetical posterior probabilities. In contrast to the prior probabilities, these are not equal anymore after having updated one's knowledge based on observed data. The left scale again compares the hypotheses that allow  $\mu$  to differ from zero with the hypotheses that fix  $\mu$  to zero. Based on the posterior probabilities, this comparison favors the hypotheses that allow  $\mu$  to be non-zero (combined posterior probability of .75) over the hypotheses that fix  $\mu$  to zero (combined posterior probability of .25). Consequently, the posterior inclusion odds are given by  $.75/.25 = 3$ . Finally, the model-averaged inclusion Bayes factor for an effect is obtained by dividing the posterior inclusion odds by the prior inclusion odds:<sup>7</sup>

$$\underbrace{\text{BF}_{10}}_{\text{Inclusion Bayes factor for effect}} = \frac{\underbrace{p(\mathcal{H}_1^f | \text{data}) + p(\mathcal{H}_1^r | \text{data})}_{\text{Posterior inclusion odds for effect}}}{\underbrace{p(\mathcal{H}_0^f | \text{data}) + p(\mathcal{H}_0^r | \text{data})}_{\text{Prior inclusion odds for effect}}} \bigg/ \frac{p(\mathcal{H}_1^f) + p(\mathcal{H}_1^r)}{p(\mathcal{H}_0^f) + p(\mathcal{H}_0^r)}. \quad (8.4)$$

In this example, dividing the posterior inclusion odds by the prior inclusion odds yields  $\text{BF}_{10} = 3/1 = 3$ . This Bayes factor indicates that compared to the effect-absent hypothesis, the data have made the effect-present hypothesis 3 times more likely than it was before.

In a similar fashion, one can compute a model-averaged inclusion Bayes factor to compare all hypotheses that allow the between-study standard deviation  $\tau$  to be non-zero (i.e.,  $\mathcal{H}_0^r$  and  $\mathcal{H}_1^r$ ) to all hypotheses that fix  $\tau$  to zero (i.e.,  $\mathcal{H}_0^f$  and  $\mathcal{H}_1^f$ ):

$$\underbrace{\text{BF}_{rf}}_{\text{Inclusion Bayes factor for heterogeneity}} = \frac{\underbrace{p(\mathcal{H}_0^r | \text{data}) + p(\mathcal{H}_1^r | \text{data})}_{\text{Posterior inclusion odds for heterogeneity}}}{\underbrace{p(\mathcal{H}_0^f | \text{data}) + p(\mathcal{H}_1^f | \text{data})}_{\text{Prior inclusion odds for heterogeneity}}} \bigg/ \frac{p(\mathcal{H}_0^r) + p(\mathcal{H}_1^r)}{p(\mathcal{H}_0^f) + p(\mathcal{H}_1^f)}. \quad (8.5)$$

---

*include*  $\mu$  as a free parameter to all hypotheses that do not include  $\mu$  as a free parameter but fix it to zero.

<sup>6</sup>Note that this may not be the case when the prior probabilities of the hypotheses are not set equal.

<sup>7</sup>Note that, in contrast to Bayes factors that compare only two models, inclusion Bayes factors that involve more than two models are affected by the setting of the prior probabilities as they do not cancel out (see Appendix).

The computation of this Bayes factor is also illustrated in Figure 8.4 (i.e., scales on the right). The prior inclusion odds for heterogeneity are equal to one, and the posterior inclusion odds are equal to  $.45/.55 \approx 0.82$ . Consequently,  $BF_{rf} = (.45/.55)/1 \approx 0.82$ , or expressed in favor of no heterogeneity,  $BF_{fr} \approx 1.22$ . This Bayes factor indicates that compared to the heterogeneity-present hypothesis, the data have made the heterogeneity-absent hypothesis about 1.22 times more likely than it was before.

One may also use model averaging in estimation to obtain a model-averaged posterior distribution for the parameters  $\mu$  and  $\tau$ . These model-averaged posterior distributions combine the posterior for each hypothesis by weighting them with the posterior probability of each hypothesis. There are two useful ways of obtaining model-averaged posteriors. First, one may combine the posterior for, say,  $\mu$  for all four hypotheses according to their posterior probabilities. Since two of the hypotheses fix  $\mu$  a priori to zero (i.e.,  $\mathcal{H}_0^f$  and  $\mathcal{H}_0^r$ ), the model-averaged posterior will be a mixture between a point-mass at zero and a continuous component. Second, one could choose to focus only on the hypotheses that do not fix the parameter to zero. This yields a model-averaged posterior without a spike at zero. Importantly, in this case one needs to be clear about the fact that this represents the model-averaged posterior under the assumption that the effect is non-zero. In the software that we use below (i.e., `metaBMA` and `JASP`), only the latter approach has currently been implemented (i.e., displaying the model-averaged posterior conditional on assuming that the effect is present).

### 8.3 Example: Testing the Self-Concept Maintenance Theory

According to the self-concept maintenance theory (Mazar, Amir, & Ariely, 2008), people will cheat to maximize self-profit, but only to the extent that they can still maintain a positive self-view. In their Experiment 1, Mazar et al. gave participants an incentive and opportunity to cheat. Before working on a problem-solving task, participants either recalled, as a moral reminder, the Ten Commandments, or, as a neutral condition, they recalled 10 books they had read in high school. In line with the self-concept maintenance hypothesis, participants in the moral reminder condition reported having solved fewer problems than those in the neutral condition which also reflected their actual performance better. Recently, a Registered Replication Report (Verschuere et al., 2018) attempted to replicate this finding. Here we focus on the primary meta-analysis that included data from 19 labs. Figure 8.5 displays the observed Cohen’s  $d$  effect size and corresponding 95% confidence interval for each lab.<sup>8</sup> Negative effect sizes are in line with the self-concept maintenance hypothesis (i.e., the self-concept maintenance theory predicts that participants in the Ten Commandments condition cheat less than participants in the neutral condition, not more) whereas positive effect sizes are opposite to what the theory predicts.

---

<sup>8</sup>We converted the raw effect sizes to standardized effect sizes (Cohen’s  $d$ ) with corresponding standard errors.

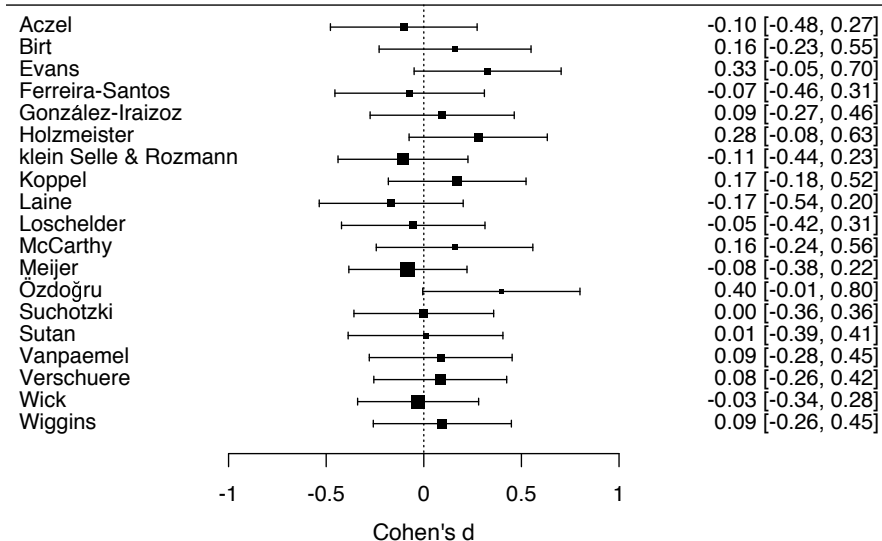


Figure 8.5: Observed effect sizes (Cohen's  $d$ ) with corresponding 95% confidence intervals for the Registered Replication Report by Verschuere et al. (2018). Only the 19 labs that were included in the primary analysis are displayed. Available at <https://tinyurl.com/ydad5k7p> under CC license <https://creativecommons.org/licenses/by/2.0/>.

For the primary analysis, Verschuere et al. reported a meta-analytic Cohen's  $d$  of 0.04 (95% CI = [-0.04, 0.12]).<sup>9</sup> Consequently, the effect was non-significant and in the opposite direction of the effect size in the original study. Furthermore, Verschuere et al. concluded that there was no heterogeneity across labs:  $\tau^2 = 0$ ,  $Q(18) = 13.16$ ,  $p = .78$ . Here we conduct a reanalysis using the Bayesian model-averaged meta-analysis approach.

### 8.3.1 Parameter Prior Settings

We use three different parameter prior specifications. These specifications differ only in the prior for  $\mu$  as the prior for  $\tau$  is always an Inverse-Gamma(1, 0.15) distribution. The first specification assigns  $\mu$  a default zero-centered Cauchy prior distribution with scale  $1/\sqrt{2}$ . This specification will be referred to as *Default*

<sup>9</sup>Note that Verschuere et al. attached a minus sign to this effect size to indicate that the effect goes in the direction opposite to that of the hypothesis.

Table 8.1: Prior and posterior probabilities of the four hypotheses of interest for the Verschuere et al. (2018) Registered Replication Report data. The posterior probabilities are displayed for three different prior settings for the effect size parameter  $\mu$ .

Hypothesis	$p(\mathcal{H})$	$p(\mathcal{H} \mid \text{data})$		
		Default (Two-Sided)	Default (One-Sided)	Informed (One-Sided)
$\mathcal{H}_0^f$	.25	.754	.823	.837
$\mathcal{H}_1^f$	.25	.087	.017	.004
$\mathcal{H}_0^r$	.25	.143	.156	.159
$\mathcal{H}_1^r$	.25	.016	.004	.001

(*Two-Sided*). The second specification is very similar, but truncates the default Cauchy prior distribution at zero in order to incorporate the directedness of the self-concept maintenance hypothesis (i.e., participants in the Ten Commandments condition are expected to cheat less than participants in the neutral condition, not more). This specification will be referred to as *Default (One-Sided)*. Finally, the third specification uses as an informed prior for  $\mu$  a  $t$  distribution that is centered on -0.35, with scale 0.102 and three degrees of freedom. This prior is also truncated at zero to preclude effect sizes in the direction opposite to what the hypothesis predicts. This “Oosterwijk” prior has been elicited for a reanalysis of a social psychology study (Gronau, Ly, & Wagenmakers, 2020), but we believe it is a reasonable prior for psychological studies more generally.<sup>10</sup> This specification will be referred to as *Informed (One-Sided)*.

## 8.3.2 Results

### 8.3.2.1 Hypotheses Posterior Probabilities

Table 8.1 displays the prior and posterior probabilities of the hypotheses for each of the three different prior specifications. The ordering of the posterior probabilities is identical for all three prior specifications: The fixed-effect null hypothesis  $\mathcal{H}_0^f$  receives most posterior probability, followed by the random-effects null hypothesis  $\mathcal{H}_0^r$ , the fixed-effect alternative hypothesis  $\mathcal{H}_1^f$ , and the random-effects alternative hypothesis  $\mathcal{H}_1^r$ .

### 8.3.2.2 Model-Averaged Bayes Factor for an Overall Effect

To address the question whether the meta-analytic effect is non-zero (i.e., Q1), we compute the model-averaged Bayes factor  $\text{BF}_{10}$  for each prior setting. This can be achieved solely based on the probabilities presented in Table 8.1. For the *Default (Two-Sided)* prior setting, the posterior inclusion odds for an effect are given by

<sup>10</sup>We flipped the sign of the location parameter to align with the way the data are coded (i.e., the theory predicts negative effect sizes).

$(.087 + .016)/(.754 + .143) \approx 0.115$ . Since the prior inclusion odds are equal to one, this number equals the model-averaged Bayes factor,  $BF_{10} \approx 0.115$ . Consequently,  $BF_{01} = 1/BF_{10} \approx 8.696$  indicating moderate evidence for the absence of an effect. For the *Default (One-Sided)* prior setting, the posterior inclusion odds for an effect are given by  $(.017 + .004)/(.823 + .156) \approx 0.021$ ; this number equals the model-averaged Bayes factor,  $BF_{10} \approx 0.021$ . Consequently,  $BF_{01} = 1/BF_{10} \approx 47.619$  indicating very strong evidence for the absence of an effect. For the *Informed (One-Sided)* prior setting, the posterior inclusion odds are calculated in the same fashion. The model-averaged Bayes factor is therefore  $BF_{10} \approx (.004 + .001)/(.837 + .159) \approx 0.005$ . Consequently,  $BF_{01} = 1/BF_{10} \approx 200$  indicating extreme evidence for the absence of an effect. In sum, for all prior settings, the model-averaged Bayes factor indicates evidence in favor of the null hypothesis of no effect. However, the degree of evidence differs across prior settings. The reason why the *Default (One-Sided)* and the *Informed (One-Sided)* prior setting yield more evidence for the absence of an effect is that, as reported by Verschuere et al., the meta-analytic effect goes in the direction opposite of what the theory predicts and these priors for  $\mu$  do not assign any mass to population effect size values that go in the opposite direction.

### 8.3.2.3 Model-Averaged Bayes Factor for Heterogeneity

To address the question whether there is heterogeneity in effect size across studies (i.e., Q2), we compute the model-averaged Bayes factor  $BF_{rf}$  for each prior setting. This can again be achieved solely based on the probabilities presented in Table 8.1. For the *Default (Two-Sided)* prior setting, the posterior inclusion odds for heterogeneity are given by  $(.143 + .016)/(.754 + .087) \approx 0.189$ . Since the prior inclusion odds are equal to one, this number equals the model-averaged Bayes factor,  $BF_{rf} \approx 0.189$ . Consequently,  $BF_{fr} = 1/BF_{rf} \approx 5.291$  indicating moderate evidence for the absence of heterogeneity. For the *Default (One-Sided)* prior setting, the posterior inclusion odds for heterogeneity are given by  $(.156 + .004)/(.823 + .017) \approx 0.190$ ; this number equals the model-averaged Bayes factor,  $BF_{rf} \approx 0.190$ . Consequently,  $BF_{fr} = 1/BF_{rf} \approx 5.263$  indicating moderate evidence for the absence of heterogeneity. For the *Informed (One-Sided)* prior setting the model-averaged Bayes factor is given by  $BF_{rf} \approx (.159 + .001)/(.837 + .004) \approx 0.190$ . Consequently,  $BF_{fr} = 1/BF_{rf} \approx 5.263$  indicating moderate evidence for the absence of heterogeneity. In sum, for all prior settings, the model-averaged Bayes factor indicates evidence in favor of the null hypothesis of no heterogeneity. The degree of evidence is very similar across prior settings, indicating moderate evidence for the absence of heterogeneity.

### 8.3.2.4 Sequential Analysis

For this particular example, studies were conducted at about the same time and we do not know the order in which they finished. However, in other cases the temporal order may be known and of interest. This is especially the case for meta-analyses combining studies from several decades where trends in the field may affect study design and results. Here we demonstrate how to conduct a sequential analysis that displays the evidence as studies accumulate. Since the presented approach

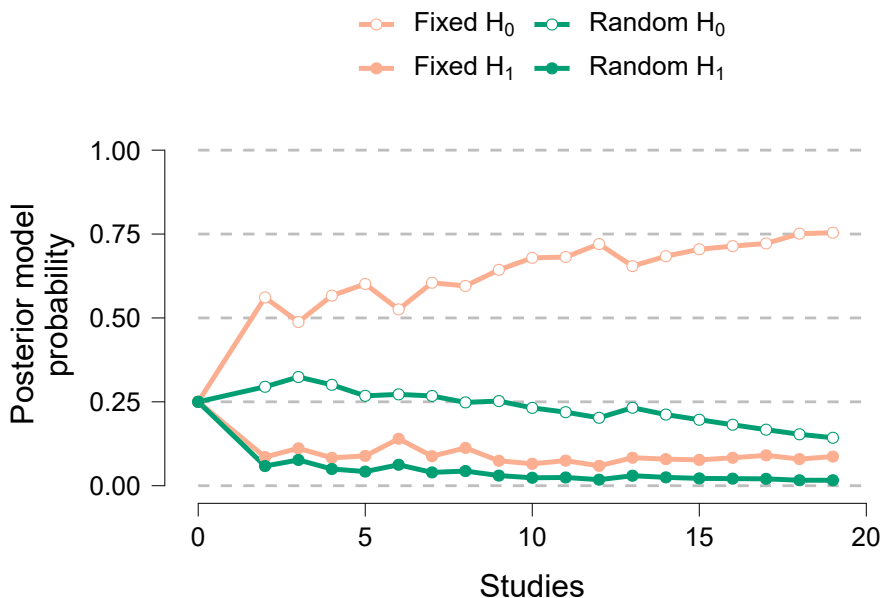


Figure 8.6: Sequential analysis. The posterior probability for each of the four hypotheses is displayed as a function of the number of studies included in the analysis. Figure from JASP ([jasp-stats.org](http://jasp-stats.org)).

is Bayesian, current knowledge can be updated by new evidence without having to worry about optional stopping (Rouder, 2014). To demonstrate the sequential analysis, we make the arbitrary assumption that the temporal order of the studies coincides with the alphabetical order of the last names of the labs' leading researchers. Furthermore, for demonstration purposes, we focus on one prior setting, *Default (Two-Sided)*. Figure 8.6 displays how the posterior probability for each of the four hypotheses changes as studies accumulate. Note that at the zero point of the  $x$ -axis, all hypotheses have “posterior” probability .25: without any data, the posterior probability equals the prior probability. Figure 8.6 highlights that the posterior probability for the fixed-effect null hypothesis  $\mathcal{H}_0^f$  increases as more studies become available. Compared to the prior probability all other hypotheses decrease in plausibility over time. Notably, both hypotheses that fix effect size  $\mu$  to zero ( $\mathcal{H}_0^f$  and  $\mathcal{H}_0^r$ ) have a higher posterior probability than the two hypotheses that allow  $\mu$  to differ from zero ( $\mathcal{H}_1^f$  and  $\mathcal{H}_1^r$ ). The lines end with the inclusion of study 19, and this point describes the current state of evidence. However, as more studies become available one could extend this analysis further and interpret the updated state of evidence (Berger & Wolpert, 1988; Rouder, 2014; Wagenmakers, Gronau, & Vandekerckhove, 2018).

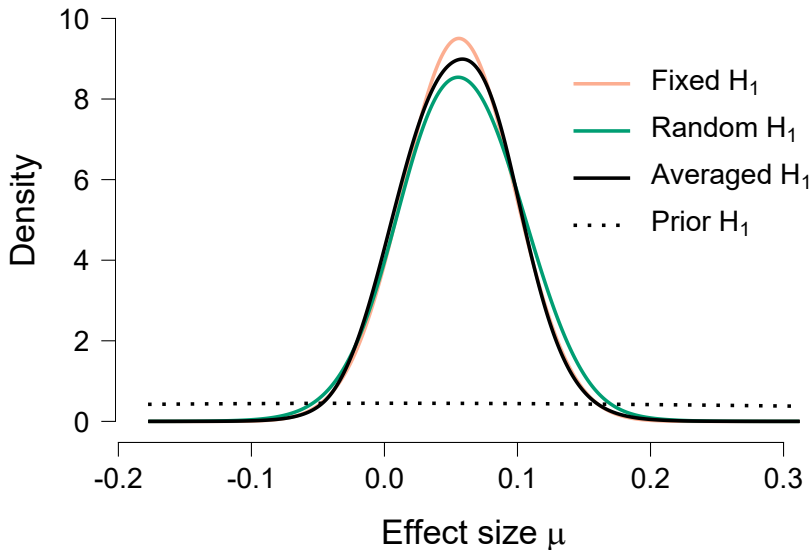


Figure 8.7: Posterior distribution for the effect size parameter  $\mu$ . The posterior is displayed for both hypotheses that do not fix  $\mu$  to zero. Additionally, the model-averaged posterior distribution is displayed. The prior distribution is shown as a dotted line. Figure from JASP ([jasp-stats.org](http://jasp-stats.org)).

### 8.3.2.5 Parameter Posterior Distribution

As shown above, all prior settings resulted in evidence against the self-concept maintenance theory. It could be argued that this makes estimation of the population effect size unnecessary – the data offer no reason to consider an estimate other than  $\mu = 0$ . Nevertheless, in practice, it may still be of interest to show how small or large the effect size is estimated under the assumption that the effect is non-zero. In general we believe that for parameter estimation, it is advisable to not use a truncated prior for the parameter of interest (van Doorn et al., in press). The reason is that, as in the present example, the effect may be in the direction opposite to what the hypothesis predicts. Whenever a prior is truncated to allow only effect sizes that align with the hypothesis, it is impossible to obtain a posterior that assigns probability mass to effect sizes in the opposite direction. As a consequence, a posterior distribution based on truncated priors may be misleading (in the present example, the truncated posterior would be left-skewed with almost all probability mass close to zero). Figure 8.7 displays the posterior distribution for  $\mu$  based on the *Default (Two-Sided)* prior setting. Posteriors are shown for both hypotheses that allow  $\mu$  to differ from zero ( $\mathcal{H}_1^f$  and  $\mathcal{H}_1^r$ ) and, additionally, the model-averaged posterior that is obtained by combining these two posteriors according to the plausibility of the hypotheses in light of the data. Figure 8.7 shows that, assuming  $\mu$  is not exactly equal to zero, it is likely to be small with

most posterior mass in the direction opposite to what the theory predicts. Furthermore, the posterior distributions under both hypotheses are very similar which results in a model-averaged posterior that is also very similar.

## 8.4 Discussion

In this primer we have discussed Bayesian model-averaged meta-analysis as a method for quantitatively synthesizing the results of a set of studies. This procedure affords researchers the well-known pragmatic benefits of a Bayesian method (Wagenmakers, Marsman, et al., 2018; Wagenmakers, Morey, & Lee, 2016). In addition, it allows researchers to take into account model uncertainty with respect to choosing a fixed-effect or random-effects model when addressing the two key questions “Is the overall effect non-zero?” (Q1) and “Is there between-study variability in effect size?” (Q2).

### 8.4.1 Effects of Prior Settings

There are two a priori settings to consider for a Bayesian model-averaged meta-analysis: the prior probabilities for the four models (i.e., prior model probabilities) and the prior distributions for the overall effect  $\mu$  and the study heterogeneity  $\tau$  (i.e., prior parameter distributions). We now discuss each setting in turn.

Concerning the prior model probabilities, in the Appendix we show how the results change as a function of how the prior probability is distributed across the four models. When comparing *two* models the choice of prior model probabilities does not affect the Bayes factor; however, this is no longer the case when more than two models are in play. In such scenarios, the model-averaged Bayes factors are generally sensitive to the choice of prior model probabilities. For unequal prior probabilities the posterior probabilities may change quite drastically. In our application to the data from Verschuere et al. (2018), however, the pattern of Bayes factors is relatively robust to reasonable changes in the prior model probabilities (see Appendix). Nevertheless, we recommend using uniform prior probability settings across the models if there are no clear theoretical reasons for different settings.

Concerning the prior distributions for the model parameters, concrete recommendations are provided in Box 1. We showed that in our application to the data from Verschuere et al. (2018), for some reasonably informed choices the pattern of evidence from the Bayes factors is comparable. The more informed a prior distribution is (e.g., choosing a one-sided prior distribution for the overall effect size) the faster evidence accumulates for or against this hypothesis. When in doubt about these settings, we recommend conducting a robustness analysis where researchers choose several reasonable prior settings and check how these choices affect the results. Note that in this primer, we focused on standardized mean difference effect sizes (i.e., Cohen’s  $d$  or Hedge’s  $g$ ) and provided recommendations for how to choose the prior distributions for this case. If the observed effect sizes are not standardized mean differences, one needs to adjust these prior distributions. Providing recommendations for other cases such as Fisher’s  $z$  and log odd ratios is left to future research.

### 8.4.2 Beyond Overall Effects

In addition to the key questions Q1 and Q2, researcher may often be interested in incorporating discrete and continuous moderators at the study-level. Although we did not discuss this possibility here, the `metaBMA` package does provide functionality for including moderators. Including moderators in the analysis is one way of accounting for the fact that different subsets of studies might have different latent effect sizes. Another possible way of incorporating and testing this assumption would be to change the distribution of the latent study effects. Instead of assuming a single continuous normal distribution of effect sizes one could assume a latent mixture of normal distributions and then test how many components are necessary to describe the distribution of latent study effects best (e.g., Moreau & Corballis, 2019).

An additional approach to a Bayesian meta-analysis is to focus on the entire distribution of study effects instead of the overall effect. For instance, Rouder et al. (2019) propose to test whether all studies in the meta-analytic sample show an effect in the same, expected direction, or whether some studies show an opposite effect. An appropriate model for this analysis is one where both the distribution of the overall effect and the distribution of individual study effects are truncated; the latter truncation is imposed in order to allow individual study effects in one direction only (upper level of Figure 8.1). This model can then be compared to the unconstrained alternative (i.e., the random-effects alternative). Similar tests have been proposed in the clinical literature where meta-analysis also serves the purpose to test whether one treatment is superior for one patient population, and another treatment is superior for another patient population (Gail & Simon, 1985). Such a “Does every study show an effect?” analysis is implemented in the `metaBMA` package.

As a final word of caution, we would like to stress that, in line with the adage “garbage in, garbage out”, no statistical analysis can provide high quality inference based on low quality data that might be the result of problematic study design, shortcomings of the implementation or sample, publication bias, significance-chasing, etc.; Bayesian model-averaged meta-analysis is no exception. For instance, one may use the procedure to analyze studies that have not been pre-registered, however, the conclusions might need to be interpreted with scepticism in case the quality of the included studies is questionable, or if the included studies represent a biased sample of all conducted studies in a field. In contrast, when the set of studies is of high quality, preregistered, and possibly even the result of a Registered (Replication) Report, we believe that Bayesian model-averaged meta-analysis can be a valuable tool that allows researchers to address key questions of interest in a principled manner.

R code and a JASP file for reproducing the analyses can be found at: <https://osf.io/npw5c/>.

## 8.A Changing the Prior Probabilities of the Hypotheses

When computing Bayes factors that compare two models such as  $\text{BF}_{\mathcal{H}_1^f, \mathcal{H}_0^f}$  (see Equation 8.2 and Equation 8.3) the prior probabilities of the hypotheses do not affect the resulting Bayes factor. For instance, when inserting the expressions for the posterior probabilities in Equation 8.3, the prior probabilities cancel out:

$$\text{BF}_{\mathcal{H}_1^f, \mathcal{H}_0^f} = \frac{p(\text{data} \mid \mathcal{H}_1^f) p(\mathcal{H}_1^f)}{p(\text{data} \mid \mathcal{H}_0^f) p(\mathcal{H}_0^f)} \bigg/ \frac{p(\mathcal{H}_1^f)}{p(\mathcal{H}_0^f)} = \frac{p(\text{data} \mid \mathcal{H}_1^f)}{p(\text{data} \mid \mathcal{H}_0^f)}. \quad (8.6)$$

In contrast, when computing *inclusion* Bayes factors that involve more than two models, the prior probabilities affect the resulting Bayes factors. For instance, when inserting the expressions for the posterior probabilities in Equation 8.4, the prior probabilities do not cancel out:<sup>11</sup>

$$\text{BF}_{10} = \frac{p(\text{data} \mid \mathcal{H}_1^f) p(\mathcal{H}_1^f) + p(\text{data} \mid \mathcal{H}_1^r) p(\mathcal{H}_1^r)}{p(\text{data} \mid \mathcal{H}_0^f) p(\mathcal{H}_0^f) + p(\text{data} \mid \mathcal{H}_0^r) p(\mathcal{H}_0^r)} \bigg/ \frac{p(\mathcal{H}_1^f) + p(\mathcal{H}_1^r)}{p(\mathcal{H}_0^f) + p(\mathcal{H}_0^r)}. \quad (8.7)$$

Here we demonstrate the effect of changing the prior probabilities of the hypotheses using the self-concept maintenance example. Specifically, we show how the posterior probabilities of the hypotheses and the inclusion Bayes factors change when:

1. increasing the prior probability of the winning hypothesis  $\mathcal{H}_0^f$  from .25 to .70;
2. increasing the prior probability of the worst hypothesis  $\mathcal{H}_1^r$  from .25 to .70.

The remaining prior probability .30 is distributed evenly across the other three hypotheses (i.e., each of the remaining hypotheses is assigned prior probability .10).

### 8.A.1 Increasing the Prior Probability of $\mathcal{H}_0^f$

#### 8.A.1.1 Hypotheses Posterior Probabilities

Table 8.2 displays the prior probabilities of the hypotheses and the posterior probabilities of the hypotheses for each of the three different prior specifications for  $\mu$ . Although the numbers changed, the ordering of the posterior probabilities is

---

<sup>11</sup>The prior probabilities do cancel out when the models that allow for an effect (i.e.,  $\mathcal{H}_1^f$  and  $\mathcal{H}_1^r$ ) are assigned equal prior probability  $c_1$  and the models that do not allow for an effect (i.e.,  $\mathcal{H}_0^f$  and  $\mathcal{H}_0^r$ ) are assigned equal prior probability  $c_2$ . Note that  $c_1$  and  $c_2$  can be different. However, in that case, the model-averaged Bayes factor for testing the presence of between-study heterogeneity  $\text{BF}_{rf}$  will be affected since the prior probabilities do not cancel out. Similarly, for  $\text{BF}_{rf}$ , the prior probabilities do cancel out when the models that allow for heterogeneity (i.e.,  $\mathcal{H}_0^f$  and  $\mathcal{H}_1^r$ ) are assigned equal prior probability  $c_3$  and the models that do not allow for heterogeneity (i.e.,  $\mathcal{H}_0^r$  and  $\mathcal{H}_1^f$ ) are assigned equal prior probability  $c_4$ . However, in that case, the model-averaged Bayes factor for testing the presence of an effect  $\text{BF}_{10}$  will be affected since the prior probabilities do not cancel out anymore.

Table 8.2: Prior and posterior probabilities of the four hypotheses of interest for the Verschuere et al. (2018) Registered Replication Report data. The posterior probabilities are displayed for three different prior settings for the effect size parameter  $\mu$ . Note that the prior probability of  $\mathcal{H}_0^f$  is set to .70.

Hypothesis	$p(\mathcal{H})$	$p(\mathcal{H} \mid \text{data})$		
		Default (Two-Sided)	Default (One-Sided)	Informed (One-Sided)
$\mathcal{H}_0^f$	.70	.955	.970	.973
$\mathcal{H}_1^f$	.10	.016	.003	.001
$\mathcal{H}_0^r$	.10	.026	.026	.026
$\mathcal{H}_1^r$	.10	.003	.001	.000

identical to the one obtained when using equal prior probabilities for all four hypotheses: For all prior specifications, the fixed-effect null hypothesis  $\mathcal{H}_0^f$  receives most posterior probability, followed by the random-effects null hypothesis  $\mathcal{H}_0^r$ , the fixed-effect alternative hypothesis  $\mathcal{H}_1^f$ , and the random-effects alternative hypothesis  $\mathcal{H}_1^r$ .

### 8.A.1.2 Model-Averaged Bayes Factor for an Overall Effect

For the *Default (Two-Sided)* prior setting,  $\text{BF}_{10} \approx 0.077$ . Consequently,  $\text{BF}_{01} \approx 12.987$  indicating strong evidence for the absence of an effect. Recall that equal prior probabilities for all four hypotheses yielded  $\text{BF}_{01} \approx 8.696$  indicating moderate evidence for the absence of an effect. For the *Default (One-Sided)* prior setting,  $\text{BF}_{10} \approx 0.016$ . Consequently,  $\text{BF}_{01} \approx 62.5$  indicating very strong evidence for the absence of an effect. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{01} \approx 47.619$  indicating also very strong evidence for the absence of an effect. For the *Informed (One-Sided)* prior setting,  $\text{BF}_{10} \approx 0.004$ . Consequently,  $\text{BF}_{01} \approx 250$  indicating extreme evidence for the absence of an effect. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{01} \approx 200$  indicating also extreme evidence for the absence of an effect. In sum, the inclusion Bayes factors based on the different setting of the prior probabilities of the four hypotheses (see Table 8.2) qualitatively agree with the ones obtained when using equal prior probabilities: there is evidence for the absence of an effect. However, they differ in the degree of evidence for the absence of an effect.

### 8.A.1.3 Model-Averaged Bayes Factor for Heterogeneity

For the *Default (Two-Sided)* prior setting,  $\text{BF}_{rf} \approx 0.119$ . Consequently,  $\text{BF}_{fr} \approx 8.403$  indicating moderate evidence for the absence of heterogeneity. Recall that equal prior probabilities for all four hypotheses yielded  $\text{BF}_{fr} \approx 5.291$  indicating also moderate evidence for the absence of heterogeneity. For the *Default (One-Sided)* prior setting,  $\text{BF}_{rf} \approx 0.111$ . Consequently,  $\text{BF}_{fr} \approx 9.009$  indicating moderate evidence for the absence of heterogeneity. Equal prior probabilities for all four

hypotheses yielded  $\text{BF}_{fr} \approx 5.263$  indicating also moderate evidence for the absence of heterogeneity. For the *Informed (One-Sided)* prior setting,  $\text{BF}_{rf} \approx 0.107$ . Consequently,  $\text{BF}_{fr} \approx 9.346$  indicating moderate evidence for the absence of heterogeneity. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{fr} \approx 5.263$  indicating also moderate evidence for the absence of heterogeneity. In sum, the inclusion Bayes factors based on the different setting of the prior probabilities of the four hypotheses (see Table 8.2) qualitatively agree with the ones obtained when using equal prior probabilities: there is evidence for the absence of heterogeneity. However, they differ in the degree of evidence for the absence of heterogeneity.

## 8.A.2 Increasing the Prior Probability of $\mathcal{H}_1^r$

### 8.A.2.1 Hypotheses Posterior Probabilities

Table 8.3 displays the prior probabilities of the hypotheses and the posterior probabilities of the hypotheses for each of the three different prior specifications for  $\mu$ . Although the numbers changed, the ordering of the posterior probabilities is similar to the one obtained when using equal prior probabilities for all four hypotheses: For all prior specifications, the fixed-effect null hypothesis  $\mathcal{H}_0^f$  receives most posterior probability, followed by the random-effects null hypothesis  $\mathcal{H}_0^r$ . However, now the fixed-effect alternative hypothesis  $\mathcal{H}_1^f$  receives less posterior probability than the random-effects alternative hypothesis  $\mathcal{H}_1^r$ .

### 8.A.2.2 Model-Averaged Bayes Factor for an Overall Effect

For the *Default (Two-Sided)* prior setting,  $\text{BF}_{10} \approx 0.056$ . Consequently,  $\text{BF}_{01} \approx 17.857$  indicating strong evidence for the absence of an effect. Recall that equal prior probabilities for all four hypotheses yielded  $\text{BF}_{01} \approx 8.696$  indicating moderate evidence for the absence of an effect. For the *Default (One-Sided)* prior setting,  $\text{BF}_{10} \approx 0.011$ . Consequently,  $\text{BF}_{01} \approx 90.909$  indicating very strong evidence for the absence of an effect. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{01} \approx 47.619$  indicating also very strong evidence for the absence of an effect. For the *Informed (One-Sided)* prior setting,  $\text{BF}_{10} \approx 0.003$ . Consequently,  $\text{BF}_{01} \approx 333.333$  indicating extreme evidence for the absence of an effect. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{01} \approx 200$  indicating also extreme evidence for the absence of an effect. In sum, the inclusion Bayes factors based on the different setting of the prior probabilities of the four hypotheses (see Table 8.3) qualitatively agree with the ones obtained when using equal prior probabilities: there is evidence for the absence of an effect. However, they differ in the degree of evidence for the absence of an effect.

### 8.A.2.3 Model-Averaged Bayes Factor for Heterogeneity

For the *Default (Two-Sided)* prior setting,  $\text{BF}_{rf} \approx 0.076$ . Consequently,  $\text{BF}_{fr} \approx 13.158$  indicating strong evidence for the absence of heterogeneity. Recall that equal prior probabilities for all four hypotheses yielded  $\text{BF}_{fr} \approx 5.291$  indicating moderate evidence for the absence of heterogeneity. For the *Default (One-Sided)* prior setting,  $\text{BF}_{rf} \approx 0.054$ . Consequently,  $\text{BF}_{fr} \approx 18.519$  indicating strong

Table 8.3: Prior and posterior probabilities of the four hypotheses of interest for the Verschuere et al. (2018) Registered Replication Report data. The posterior probabilities are displayed for three different prior settings for the effect size parameter  $\mu$ . Note that the prior probability of  $\mathcal{H}_1^r$  is set to .70.

Hypothesis	$p(\mathcal{H})$	$p(\mathcal{H} \mid \text{data})$		
		Default (Two-Sided)	Default (One-Sided)	Informed (One-Sided)
$\mathcal{H}_0^f$	.10	.687	.805	.833
$\mathcal{H}_1^f$	.10	.079	.017	.004
$\mathcal{H}_0^r$	.10	.130	.153	.158
$\mathcal{H}_1^r$	.70	.104	.026	.006

evidence for the absence of heterogeneity. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{fr} \approx 5.263$  indicating moderate evidence for the absence of heterogeneity. For the *Informed (One-Sided)* prior setting,  $\text{BF}_{rf} \approx 0.049$ . Consequently,  $\text{BF}_{fr} \approx 20.408$  indicating strong evidence for the absence of heterogeneity. Equal prior probabilities for all four hypotheses yielded  $\text{BF}_{fr} \approx 5.263$  indicating moderate evidence for the absence of heterogeneity. In sum, the inclusion Bayes factors based on the different setting of the prior probabilities of the four hypotheses (see Table 8.3) qualitatively agree with the ones obtained when using equal prior probabilities: there is evidence for the absence of heterogeneity. However, they differ in the degree of evidence for the absence of heterogeneity.

### 8.A.3 Summary

In sum, changing the prior probabilities of the hypotheses – as expected – has an effect on the posterior probabilities of the hypotheses. Furthermore, it also has an effect on the inclusion Bayes factors, that is, it has an effect on the degree of model-averaged evidence. However, in this particular example, using the particular changes to the prior probability that we used, it does not change the qualitative overall conclusions that there is evidence for the absence of an effect and that there is evidence for the absence of heterogeneity. In general we believe that unless there is strong prior knowledge that suggests to set the prior probabilities differently, it is prudent to set the prior probabilities of all four hypotheses uniformly to .25.

# A Bayesian Model-Averaged Meta-Analysis of the Power Pose Effect with Informed and Default Priors: The Case of Felt Power

---

## Abstract

Carney, Cuddy, and Yap (2010) found that – compared to participants who adopted constrictive body postures – participants who adopted expansive body postures reported feeling more powerful, showed an increase in testosterone and a decrease in cortisol, and displayed an increased tolerance for risk. However, these power pose effects have recently come under considerable scrutiny. Here we present a Bayesian meta-analysis of six pre-registered studies from a special issue, focusing on the effect of power posing on felt power. Our analysis improves on standard classical meta-analyses in several ways. First and foremost, we considered only preregistered studies, eliminating concerns about publication bias. Second, the Bayesian approach enables us to quantify evidence for both the alternative and the null hypothesis. Third, we use Bayesian model averaging to account for the uncertainty with respect to the choice for a fixed-effect model or a random-effect model. Fourth, based on a literature review we obtained an empirically informed prior distribution for the between-study heterogeneity of effect sizes. This empirically informed prior can serve as a default choice not only for the investigation of the power pose effect, but for effects in the field of psychology more generally. For effect size, we considered a default and an informed prior. Our meta-analysis yields very strong evidence for an effect of power

---

This chapter is published as Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, 2, 123–138. doi: <https://doi.org/10.1080/23743603.2017.1326760>. Also available as *PsyArXiv preprint*: <https://psyarxiv.com/9z8ch/>

posing on felt power. However, when the analysis is restricted to participants unfamiliar with the effect, the meta-analysis yields evidence that is only moderate.

## 9.1 Introduction

Could adopting a powerful body posture make us more powerful? Carney et al. (2010) found that participants who adopted expansive, high-power body postures (Figure 9.1, top row) as opposed to constrictive, low-power body postures (Figure 9.1, bottom row) reported feeling more powerful and in charge, showed an increase in testosterone and a decrease in cortisol, and displayed an increased tolerance for risk. The power pose effect has attracted a lot of attention, partly due to the anticipated consequences for day-to-day life suggesting that it might be possible to “fake it ‘til you make it”.

However, this power pose effect has recently come under scrutiny. When Ranehill et al. (2015) attempted to replicate the effect, they found – similar to the original study – that adopting high-power poses increased participants’ self-reported feelings of power; nevertheless, they did not find an effect on testosterone or cortisol nor on behavioral measures such as risk taking. Carney, Cuddy, and Yap (2015) pointed out a number of methodological differences that they believe might have been the cause for the diverging results. Recently, Garrison, Tang, and Schmeichel (2016) conducted a preregistered replication and extension of the power pose study, and they failed to identify an effect of power posing on risk taking behavior. Furthermore, in contrast to Ranehill et al. (2015), these authors did not find evidence for a power pose effect on subjective feelings of power.

In a special issue, seven preregistered studies investigated the effect of power posing under various circumstances (i.e., A. H. Bailey, LaFrance, & Dovidio, 2017; Bombari, Schmid Mast, & Pulfrey, 2017; Jackson, Nault, Smart Richman, LaBelle, & Rohleder, 2017; Keller, Johnson, & Harder, 2017; Klaschinski, Schröder-Abé, & Schnabel, 2017; Latu, Duffy, Pardal, & Alger, 2017; Ronay, Tybur, van Huijstee, & Morssinkhof, 2017). Here we present a meta-analysis of the effect of power posing on self-reported felt power, which was included as a dependent variable in six of the seven studies in the special issue.

Our analysis improves upon classical analyses in several ways. First, we only consider a set of preregistered studies which comes with the advantage that publication bias can be ruled out a priori (cf. the concept of a *prospective meta-analysis* in medicine). Second, the Bayesian approach enables us to quantify evidence for both the alternative hypothesis *and* for the null hypothesis; note that this evidence can be seamlessly updated as future studies on the effect become available. Third, Bayesian model averaging enables us to fully acknowledge uncertainty with respect to the choice of a fixed-effect or random-effect model; in the fixed-effect model, the effect is assumed to be identical across studies; in the random-effect model, the effect is assumed to vary across studies. Instead of adopting one model for inference and ignoring the other model entirely, we can weight the results of both models according to their posterior plausibilities. This yields a model-averaged measure of evidence and a model-averaged estimate for the meta-analytic effect

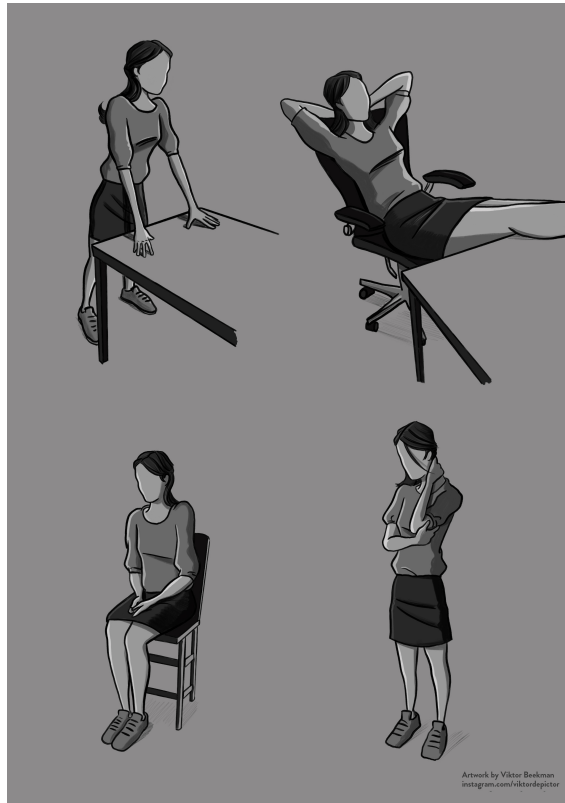


Figure 9.1: High-power poses (top row) and low-power poses (bottom row). CC-BY: Artwork by Viktor Beekman, commissioned by Eric-Jan Wagenmakers.

size. Fourth, the Bayesian approach enables us to incorporate existing knowledge into our analysis (e.g., Rhodes, Turner, & Higgins, 2015). Based on an extensive literature review of meta-analyses in the field of psychology, we obtained an informed prior distribution for the between-study heterogeneity. This informed prior distribution can serve as an informed default not only for the investigation of the power pose effect in the present meta-analysis, but for the field of psychology more generally. For effect size we also consider an informed prior distribution based on knowledge about effect sizes in the field of psychology. As a robustness check with respect to the prior choice we show that qualitatively similar results are obtained when we instead use a default prior for the effect size parameter.

The outline of this chapter is as follows: first, we explain the details of our analysis. Second, we present the results of an extensive literature review that allowed us to specify an informed prior distribution for the between-study heterogeneity. Third, we present the results of the model-averaged Bayesian meta-analysis for two different prior choices for effect size. Finally, we investigate whether the results change when only participants unaware of the power pose effect are included

in the analysis.

## 9.2 Method

In our meta-analysis, we focused on the dependent variable *felt power* which was measured in all replication studies in the special issue except for the study by Jackson et al., which was therefore not considered in the analysis. We investigated the question whether felt power was higher in the high-power condition than in the low-power condition.

### 9.2.1 Analysis of Individual Studies

When considering a single study, the power pose effect can be tested using a standard one-sided, independent-samples *t*-test. Hence, the first step in our analysis was to compute one-sided Bayesian *t*-tests (Gronau, Ly, & Wagenmakers, 2020; Ly et al., 2016b; Rouder et al., 2009). This allowed us (1) to estimate for each study the posterior distribution of the standardized effect size that represents our beliefs about the effect size after having observed the data of that study and (2) to quantify the evidence that each study provides in favor of the hypothesis that the power pose effect is positive ( $\mathcal{H}_+$ ) versus the null hypothesis that the effect is zero ( $\mathcal{H}_0$ ).

To quantify the evidence that the data provide for or against  $\mathcal{H}_+$  we computed the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995) which is the predictive updating factor that quantifies how much the data have changed the relative plausibility of the competing models. The Bayes factor has an intuitive interpretation: when  $\text{BF}_{+0} = 10$  this indicates that the data are ten times more likely under  $\mathcal{H}_+$  than under  $\mathcal{H}_0$ ; when  $\text{BF}_{+0} = 1/5$  this indicates that the data are five times more likely under  $\mathcal{H}_0$  than under  $\mathcal{H}_+$ .

### 9.2.2 Meta-Analysis

The next step in our analysis was to combine the studies with the help of a Bayesian meta-analysis (e.g., Marsman et al., 2017) to obtain an estimate of the overall effect size and to quantify the evidence for an effect that takes into account all studies simultaneously. In a classical meta-analysis the analyst has to make a choice between a fixed-effect and a random-effect model. A fixed-effect model makes the assumption that there is one underlying effect size so that the true effect in each study is identical; differences in the observed effect sizes are solely due to normally distributed sampling error. This can be formalized as follows: we assume that  $y_i \sim \mathcal{N}(\delta_{\text{fixed}}, SE_i^2)$ , where  $y_i$ ,  $i = 1, 2, \dots, n$  denotes the observed effect size in the  $i$ th of  $n$  studies,  $SE_i$  denotes the corresponding standard error which is commonly assumed to be known, and  $\delta_{\text{fixed}}$  corresponds to the common true effect size.

In contrast, a random-effect model allows for idiosyncratic study effects, that is, we no longer impose the constraint that there exists one common true effect size for all studies. The random study effects are usually assumed to follow a normal distribution with a mean equal to the overall effect size that we are interested

in and a standard deviation that corresponds to the between-study heterogeneity. Note that analogously to the fixed-effect model, the model still incorporates random sampling error so that the observed effect size for a given study is not necessarily identical to the true effect size for that study. These assumptions yield a model with a hierarchical structure which can be formalized as follows: let  $\delta_{\text{random}}$  denote the mean of the normal distribution of the study effects (i.e., the quantity that we are interested in),  $\tau$  denote the standard deviation of that normal distribution (i.e., between-study heterogeneity), and  $\theta_i$  denote the true study effect for the  $i$ th study. Then,  $\theta_i \sim \mathcal{N}(\delta_{\text{random}}, \tau^2)$  and  $y_i | \theta_i \sim \mathcal{N}(\theta_i, SE_i^2)$ . The structure of the model allows one to analytically integrate out the random study effects so that the model can equivalently be written as  $y_i \sim \mathcal{N}(\delta_{\text{random}}, \tau^2 + SE_i^2)$  which can be more convenient from a computational perspective.

### 9.2.3 Bayesian Model Averaging

The choice of a fixed-effect or random-effect model commonly relies on a test for heterogeneity or on *a priori* considerations. Final inference is then based on either the fixed-effect or random-effect model. When the number of studies is small, this choice may be difficult; and in certain cases, the choice may be consequential. The Bayesian approach, however, allows a compromise solution: instead of selecting either a fixed-effect or random-effect model, we can use Bayesian model averaging (e.g., Haldane, 1932; Hoeting et al., 1999) and retain all models for final inference. Conclusions are then based on a combination of all models where the results of each model are taken into account according to the model's plausibility in light of the observed data. Concretely, Bayesian model averaging allows us to obtain a model-averaged estimate for the meta-analytic effect size (Sutton & Abrams, 2001) and to quantify the overall evidence for an effect that considers both the fixed-effect and random-effect model (Scheibehenne et al., 2017).

With respect to hypothesis testing, for the current analysis we entertained four models of interest, shown in Table 9.1: (1) the fixed-effect model  $\mathcal{H}_+$ ; (2) the fixed-effect model  $\mathcal{H}_0$  (i.e.,  $\delta_{\text{fixed}} = 0$ ); (3) the random-effect model  $\mathcal{H}_+$ ; (4) the random-effect model  $\mathcal{H}_0$  (i.e.,  $\delta_{\text{random}} = 0$ ). The fixed-effect meta-analytic Bayes factor was obtained by comparing case (1) to case (2); the random-effect meta-analytic Bayes factor pitched case (3) against case (4). To compute the model-averaged Bayes factor, we contrasted the summed posterior model probabilities (i.e., the probability of a model given the data) for cases (1) and (3) against the summed posterior model probabilities for cases (2) and (4). This assumes that all four models are equally likely *a priori*, a common assumption in model averaging scenarios. In case the prior model probabilities were not identical, the ratio of the summed posterior model probabilities for cases (1) and (3) over (2) and (4) would need to be divided by a ratio obtained in a similar fashion but this time based on the prior model probabilities.

With respect to parameter estimation, we computed a model-averaged effect size estimate based on the four model versions described above, except that we no longer imposed the constraint that the effect size has to be positive. In other words, consistent with standard practice, we imposed a directional constraint for testing but not for estimation (cf. Jeffreys, 1961, who also used different priors

9. A BAYESIAN MODEL-AVERAGED META-ANALYSIS OF THE POWER POSE EFFECT WITH INFORMED AND DEFAULT PRIORS: THE CASE OF FELT POWER

Table 9.1: The four meta-analysis models included in the Bayesian model averaging for hypothesis testing.

Hypotheses	Fixed-Effect Meta-Analysis	Random-Effect Meta-Analysis
$\mathcal{H}_0$ : No effect	Fixed overall effect size $\delta_{\text{fixed}} = 0$	Mean overall effect size $\delta_{\text{random}} = 0$ Study heterogeneity $\tau$ Study effect size $\theta_i$ ( $i = 1, 2, \dots, n$ )
$\mathcal{H}_+$ : Positive effect	Fixed overall effect size $\delta_{\text{fixed}}$	Mean overall effect size $\delta_{\text{random}}$ Study heterogeneity $\tau$ Study effect size $\theta_i$ ( $i = 1, 2, \dots, n$ )

for estimation and testing). This reflects the fact that the estimation framework is generally more exploratory in nature, and this mindset is inconsistent with the use of hard boundaries. The combined estimate was obtained by combining the estimates of models (1) and (3) – but without the order-constraints – according to their posterior model probabilities. To conduct the model-averaged Bayesian meta-analysis, we used the R package `metaBMA` (Heck et al., 2019).

9.2.4 Prior Distributions

In the Bayesian approach, model parameters are assigned prior distributions that reflect the knowledge, uncertainty, or beliefs for the parameters before seeing the data. Using Bayes’ theorem, these prior distributions are then updated by the data to yield posterior distributions, which reflect the uncertainty for the parameters after the data have been observed. Consequently, in order to conduct our Bayesian analyses, prior distributions were required for all model parameters.

For the standardized effect size, we considered two different prior choices. First, we used what has now become the default choice in the field of psychology, that is, a zero-centered Cauchy distribution with scale parameter equal to  $1/\sqrt{2}$  (Morey & Rouder, 2015). Second, we considered the informed prior distribution reported in Gronau, Ly, and Wagenmakers (2020): a  $t$  distribution with location 0.350, scale 0.102, and three degrees of freedom, which is displayed in Figure 9.2. This prior distribution was elicited from Dr. Oosterwijk, a social psychologist at the University of Amsterdam, for a reanalysis of the Registered Replication Report on the facial feedback hypothesis (Wagenmakers, Beek, et al., 2016). We believe this prior distribution is generally plausible for a wide range of small-to-medium effects in social psychology (i.e., for effects whose presence needs to be ascertained by statistical analysis). One could elicit a “power pose prior”, but we believe the resulting distribution would be highly similar to the Oosterwijk prior, and therefore yield highly similar inferences. Researchers interested in using a specific “power pose prior” are invited to explore this option using the R code provided online (<https://osf.io/r2cds/>).

For the one-sided hypothesis tests, the priors were truncated at zero, that is, the model encoded the a priori assumption that negative effect sizes are impossible. For estimating the effect size, however, we removed this truncation. The informed and default priors are depicted in Figure 9.2. The informed prior expresses the

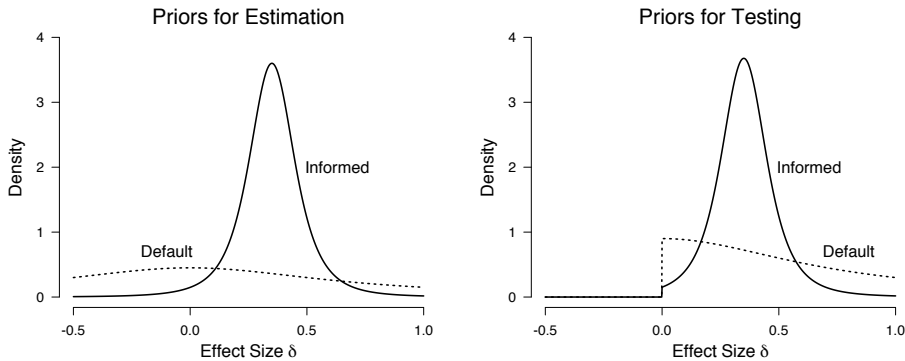


Figure 9.2: Depiction of the default and informed prior distribution for the standardized effect size. The default prior is a Cauchy distribution with scale  $1/\sqrt{2}$ , the informed prior is a  $t$  distribution with location 0.350, scale 0.102, and three degrees of freedom. Figure available at <http://tinyurl.com/j9dthma> under CC license <https://creativecommons.org/licenses/by/2.0/>.

belief that the effect size is positive but most likely small to medium in size. The default prior on the other hand is more spread out (i.e., less informative) and it is centered on zero. Figure 9.2 also illustrates how the priors were truncated at zero for testing whereas for estimation, this truncation was removed.

In addition to the prior distribution for the effect size, the Bayesian meta-analysis required a prior distribution for the between-study heterogeneity. Here we chose an informed prior distribution for the between-study standard deviation  $\tau$ . This informed prior was based on all available between-study heterogeneity estimates for mean-difference effect sizes in meta-analyses reported in Psychological Bulletin in the years 1990 to 2013 (van Erp et al., 2017, <https://osf.io/preprints/psyarxiv/myu9c>). The distribution of these 162 estimates is shown in Figure 9.3. Note that we have excluded between-study heterogeneity estimates that were exactly equal to zero, as the prior should reflect knowledge conditional on the assumption that the random-effect model is true; between-study heterogeneity estimates of exactly zero, however, suggest that the fixed-effect model was more appropriate. The distribution of the estimates in Figure 9.3 suggests that (1) the between-study standard deviations in the field of psychology range from 0 to 1 and (2) there are more small estimates than large ones. These two features are captured by an Inverse-Gamma(1, 0.15) distribution (depicted in Figure 9.3 as a solid line).<sup>1</sup> Note, however, that this prior distribution does not completely rule out the possibility that between-study heterogeneity is larger than 1; the distribution merely assigns values larger than 1 a relatively small prior credibility. This inverse-

<sup>1</sup>For computational convenience, it is common practice to assign an inverse-gamma prior to the variance instead of to the standard deviation. Here we use the inverse-gamma as a convenient summary for the empirical distribution of the between-study heterogeneity estimates.

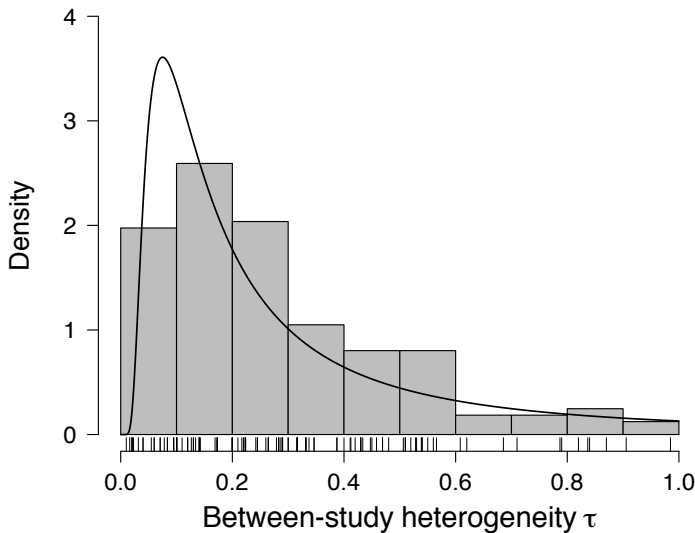


Figure 9.3: Distribution of the non-zero between-study standard deviations from meta-analyses reported in *Psychological Bulletin* (1990-2013; van Erp et al., 2017). The informed Inverse-Gamma(1, 0.15) prior distribution is displayed on top. Figure available at <http://tinyurl.com/lwfa9rd> under CC license <https://creativecommons.org/licenses/by/2.0/>.

gamma distribution resembles the one obtained when maximum-likelihood methods are used to fit an inverse-gamma distribution to the between-study heterogeneity estimates. However, in our opinion, the maximum-likelihood inverse-gamma distribution slightly overemphasizes small between-study heterogeneity values. In the appendix, we present the results obtained under two alternative prior choices for between-study heterogeneity: (1) the maximum-likelihood inverse-gamma distribution; and (2) a Beta(1, 2) prior distribution. The results are robust across all of these prior choices.

Having specified the models and prior distributions, we needed to compute the probability of the data given each model under consideration. This was achieved by integrating out the model parameters with respect to their prior distributions. For the models for which this was not possible analytically, we evaluated this quantity using numerical integration as implemented in the R package *metaBMA* (Heck et al., 2019). R code for reproducing all analyses can be found on the Open Science Framework: <https://osf.io/r2cds/>.<sup>2</sup>

---

<sup>2</sup>The R code also allows one to explore alternative prior choices easily.

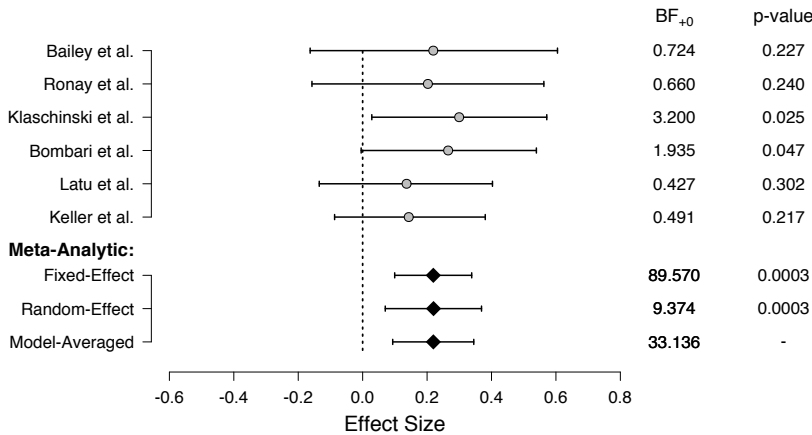


Figure 9.4: Bayesian model-averaged meta-analysis using the default Cauchy prior with scale  $1/\sqrt{2}$  for the standardized effect size. The dots and diamonds correspond to the median of the posterior distribution for the effect size; the lines correspond to the 95% highest density intervals. The one-sided Bayes factors are displayed on the right, flanked by classical two-sided  $p$ -values. Figure available at <http://tinyurl.com/kz2jpw> under CC license <https://creativecommons.org/licenses/by/2.0/>.

## 9.3 Results

### 9.3.1 Analysis of Reported Studies: Default Prior on Effect Size

Figure 9.4 displays the results of the Bayesian analysis using the default effect size prior for the studies as reported in the special issue. Note that most studies did not exclude participants who were familiar with the effect, for instance, from viewing the TED talk about power posing, which is currently the second most popular TED talk of all time ([https://www.ted.com/playlists/171/the\\_most\\_popular\\_talks\\_of\\_all](https://www.ted.com/playlists/171/the_most_popular_talks_of_all)). This analysis is based on a total of 1071 participants. Below, we investigate how the results change when considering only those participants who indicated not to know the power pose effect. The upper part of Figure 9.4 displays the results of the Bayesian  $t$ -tests. The left-part of the figure displays for each study the median of the posterior distribution for the effect size (grey dots) and a 95% highest density interval (HDI; i.e., the shortest interval that captures 95% of the posterior mass). The right part of the figure shows the one-sided default Bayes factors in favor of  $\mathcal{H}_+$  and, for comparison, the (two-sided)  $p$ -values obtained from classical independent samples  $t$ -tests. Based on the posterior distributions, it appears that there might be a positive effect. However, this is hard to assess since the 95% highest density intervals are relatively wide. All Bayes factors except one

are between  $1/3$  and  $3$  indicating that there is not much evidence for  $\mathcal{H}_+$  or  $\mathcal{H}_0$ . Hence, when considering the individual studies separately, we cannot draw strong conclusions about whether there is an effect or not.

Each study alone does not provide much evidence in favor of either hypothesis; however, a Bayesian meta-analysis allows us to obtain an impression of the overall evidence obtained when considering all studies simultaneously. The lower part of Figure 9.4 displays the result of the Bayesian meta-analysis using the default Cauchy prior with scale  $1/\sqrt{2}$  for the meta-analytic effect size. The black diamonds display the median of the posterior distribution of the meta-analytic effect size for the fixed-effect, random-effect, and model-averaged analysis, and the lines correspond to the 95% highest density intervals. The model-averaged posterior distribution is obtained by combining the estimates of the fixed-effect and the random-effect model according to their plausibility in light of the data. The lower right part of Figure 9.4 shows the meta-analytic one-sided Bayes factors and, for the fixed-effect and the random-effect model, the two-sided  $p$ -value obtained by conducting classical meta-analyses. The meta-analytic fixed-effect Bayes factor equals  $\text{BF}_{+0} = 89.6$ , indicating very strong evidence in favor of an effect of power posing on felt power. The meta-analytic random-effect Bayes factor is less extreme but still indicates evidence for an effect:  $\text{BF}_{+0} = 9.4$ . The observed data support a fixed-effect model more than a random-effect model: the Bayes factor that compares case (1), fixed-effect  $\mathcal{H}_+$ , to case (3), random-effect  $\mathcal{H}_+$ , (not displayed) indicates that the data are 4.0 times more likely under the fixed-effect model than under the random-effect model. This is reflected in the model-averaged result: the meta-analytic model-averaged Bayes factor equals  $\text{BF}_{+0} = 33.1$  indicating very strong evidence in favor of an effect of power posing on felt power. The median of the model-averaged meta-analytic effect size is equal to 0.22 [95% HDI: 0.09, 0.34].

To sum up, the Bayesian meta-analytic results based on the default prior for the effect size provide very strong evidence in favor of the hypothesis that power posing leads to an increase in felt power.

### 9.3.2 Analysis of Reported Studies: Informed Prior on Effect Size

Next, we consider the results based on the informed  $t$  prior distribution for the effect size with location 0.350, scale 0.102, and three degrees of freedom (cf. Figure 9.2). The results are displayed in Figure 9.5. The effect size posterior distributions for the individual studies clearly show the influence of the informed prior distribution: the posteriors are narrower and slightly shifted towards the location of the informed prior. The individual study one-sided informed Bayes factors are larger than the default ones. This can be explained by interpreting the Bayes factor as an assessment tool of the predictive success of two competing hypotheses. The informed alternative hypothesis makes much riskier predictions than the default alternative hypothesis; however, these risky predictions are rewarded because the observed effect sizes fall within the range of values predicted by the informed hypothesis. Hence, since the predictions match the observed data, the informed hypothesis yields more evidence for the presence of the power pose

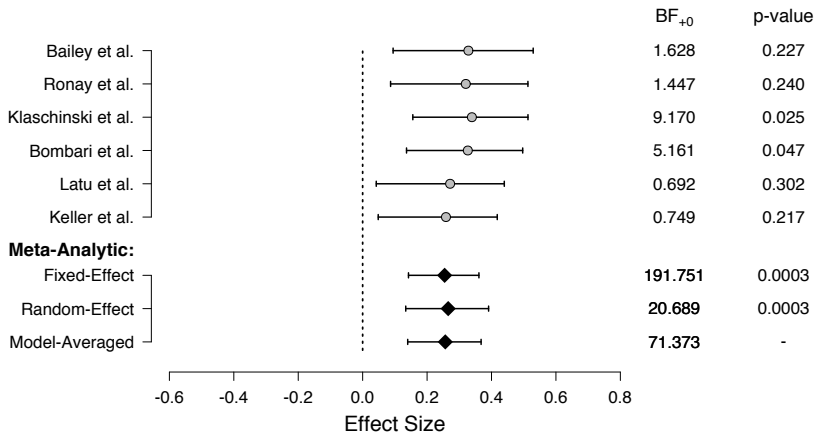


Figure 9.5: Bayesian model-averaged meta-analysis using the informed  $t$  prior with location 0.350, scale 0.102, and three degrees of freedom for the standardized effect size (depicted in Figure 9.2). The dots and diamonds correspond to the median of the posterior distribution for the effect size; the lines correspond to the 95% highest density intervals. The one-sided Bayes factors are displayed on the right, flanked by classical two-sided  $p$ -values. Figure available at <http://tinyurl.com/n8mwfsv> under CC license <https://creativecommons.org/licenses/by/2.0/>.

effect as compared to an alternative hypothesis that specifies a default prior for the effect size. Nevertheless, only two of the study-specific Bayes factors provide moderate evidence for an effect, whereas the other four provide only anecdotal evidence for  $\mathcal{H}_+$  or  $\mathcal{H}_0$ .

The informed meta-analytic fixed-effect Bayes factor is  $BF_{+0} = 191.8$  indicating extreme evidence in favor of an effect of power posing on felt power. The informed meta-analytic random-effect Bayes factor is less extreme but still indicates strong evidence for an effect:  $BF_{+0} = 20.7$ . As for the default prior, the observed data support a fixed-effect model more than a random-effect model, the Bayes factor that compares case (1), fixed-effect  $\mathcal{H}_+$ , to case (3), random-effect  $\mathcal{H}_+$ , (not displayed) indicates that the data are 3.9 times more likely under the fixed-effect model than under the random-effect model (not displayed). The informed meta-analytic model-averaged Bayes factor is equal to  $BF_{+0} = 71.4$  indicating very strong evidence in favor of an effect of power posing on felt power. The median of the model-averaged meta-analytic effect size is similar to the default one and is equal to 0.26 [95% HDI: 0.14, 0.37].

To sum up, the Bayesian meta-analytic results based on the informed prior for the effect size provide very strong evidence in favor of the hypothesis that power posing leads to an increase in felt power. The informed analysis yields more evidence for an effect as compared to the default analysis indicating that the

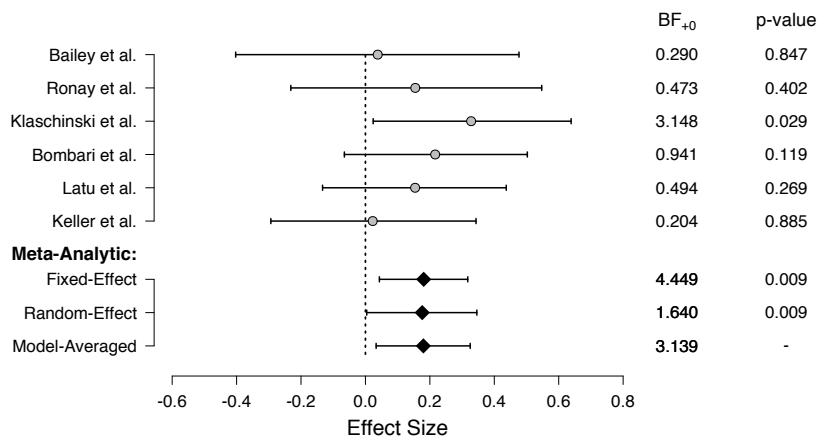


Figure 9.6: Bayesian model-averaged meta-analysis for the subset of participants unfamiliar with the effect using the default Cauchy prior with scale  $1/\sqrt{2}$  for the standardized effect size. The dots and diamonds correspond to the median of the posterior distribution for the effect size; the lines correspond to the 95% highest density intervals. The one-sided Bayes factors are displayed on the right, flanked by classical two-sided  $p$ -values. Figure available at <http://tinyurl.com/kmfcnhz> under CC license <https://creativecommons.org/licenses/by/2.0/>.

successful predictions of the informed hypothesis are rewarded.

9.3.3 Moderator Analysis: Knowledge of the Effect (Default Prior on Effect Size)

Next we investigate whether and how the results change when considering only participants who indicated to be unaware of the power posing effect. Hence, participants who could guess the goal of the study or were familiar with the power pose TED talk were excluded in all studies under consideration, leaving a total of 809 participants. Figure 9.6 displays the results of the Bayesian analysis using the default effect size prior. Compared to Figure 9.4, the posterior distributions are shifted towards smaller values and the 95% highest density intervals are relatively wide (due to the reduced sample size). Three Bayes factors are between  $1/3$  and  $3$  indicating that there is little evidence for  $\mathcal{H}_+$  or  $\mathcal{H}_0$ , one Bayes factor indicates moderate evidence for the alternative hypothesis, and two Bayes factors indicate moderate evidence for the null hypothesis. Hence, similar to the previous analysis, when considering the individual studies separately, we cannot draw strong conclusions about whether or not there is an effect.

The lower part of Figure 9.6 displays the result of the Bayesian meta-analysis using the default Cauchy prior with scale  $1/\sqrt{2}$ . The meta-analytic fixed-effect

Bayes factor equals  $\text{BF}_{+0} = 4.4$  indicating moderate evidence in favor of an effect of power posing on felt power. The meta-analytic random-effect Bayes factor equals  $\text{BF}_{+0} = 1.6$  indicating only anecdotal evidence for the alternative hypothesis. The observed data support a fixed-effect model more than a random-effect model: the Bayes factor that compares case (1), fixed-effect  $\mathcal{H}_+$ , to case (3), random-effect  $\mathcal{H}_+$ , (not displayed) indicates that the data are 3.1 times more likely under the fixed-effect model than under the random-effect model. This is reflected in the model-averaged result: the meta-analytic model-averaged Bayes factor is equal to  $\text{BF}_{+0} = 3.1$  indicating moderate evidence in favor of an effect of power posing on felt power. The median of the model-averaged meta-analytic effect size is equal to 0.18 [95% HDI: 0.03, 0.33].

To sum up, when considering only participants who were unaware of the effect and using the default effect size prior, we obtain only moderate evidence for an effect of power posing on felt power. This is in contrast to the results of the previous analysis in which participants who were familiar with the effect were mostly not excluded.

### 9.3.4 Moderator Analysis: Knowledge of the Effect (Informed Prior on Effect Size)

Next we consider the results based on the informed  $t$  prior distribution for effect size with location 0.350, scale 0.102, and three degrees of freedom (depicted in Figure 9.2) when taking into account only participants unfamiliar with the effect. The results are displayed in Figure 9.7. As before, the effect size posterior distributions for the individual studies clearly show the influence of the informed prior distribution: the posteriors are narrower and slightly shifted towards the location of the informed prior. Again, the individual study one-sided informed Bayes factors are larger than the default ones. Nevertheless, only one Bayes factor provides moderate evidence for an effect, four provide anecdotal evidence for the alternative or the null hypothesis, and one provides moderate evidence for the null.

The informed meta-analytic fixed-effect Bayes factor equals  $\text{BF}_{+0} = 6.8$ , indicating moderate evidence in favor of an effect of power posing on felt power. The informed meta-analytic random-effect Bayes factor is  $\text{BF}_{+0} = 2.6$ , indicating anecdotal evidence for an effect. As for the default prior, the observed data support a fixed-effect model more than a random-effect model, the Bayes factor that compares case (1), fixed-effect  $\mathcal{H}_+$ , to case (3), random-effect  $\mathcal{H}_+$ , (not displayed) indicates that the data are 3.0 times more likely under the fixed-effect model than under the random-effect model. The informed meta-analytic model-averaged Bayes factor is equal to  $\text{BF}_{+0} = 4.9$  indicating moderate evidence in favor of an effect of power posing on felt power. The median of the model-averaged meta-analytic effect size is equal to 0.23 [95% HDI: 0.10, 0.36].

To sum up, when considering only participants who were unaware of the effect, the results were robust with respect to using the informed or the default prior for the effect size. In both analyses, we found only moderate evidence in favor of the hypothesis that power posing leads to an increase in felt power.

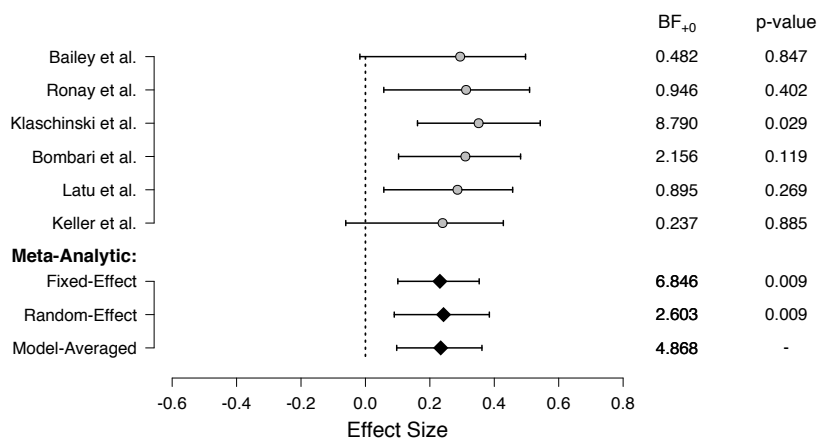


Figure 9.7: Bayesian model-averaged meta-analysis for the subset of participants unfamiliar with the effect using the informed  $t$  prior with location 0.350, scale 0.102, and three degrees of freedom for the standardized effect size. The dots and diamonds correspond to the median of the posterior distribution for the effect size; the lines correspond to the 95% highest density intervals. The one-sided Bayes factors are displayed on the right, flanked by classical two-sided  $p$ -values. Figure available at <http://tinyurl.com/n7r4huj> under CC license <https://creativecommons.org/licenses/by/2.0/>.

## 9.4 Discussion

Six preregistered studies in a special issue were subjected to a Bayesian meta-analysis of the effect of power posing on self-reported felt power. The Bayesian approach enabled us to fully acknowledge uncertainty with respect to the choice of a fixed-effect or a random-effect model, and allowed us to incorporate prior information about between-study heterogeneity and plausible effect sizes in the field of psychology. The informed prior distribution for between-study heterogeneity was based on an extensive literature review, and we believe it may serve as an informed default in the field of psychology more generally (cf. Rhodes et al., 2015, for a similar approach in medicine).

When considering the studies as reported (i.e., most studies did not exclude participants who were familiar with the effect), we obtained very strong evidence that adopting high-power poses increases subjective feelings of power; this was the case for both the analysis based on a default prior and an informed prior for the effect size. However, when considering only participants unfamiliar with the effect, we obtained only moderate evidence for an effect for both the default and informed effect size prior analysis. This suggests that knowledge of the effect might play a role with respect to the size of the effect of power posing on felt power, although

a formal assessment of this possibility requires a different statistical analysis (e.g., Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011), the development of which is beyond the scope of this chapter. Future studies might investigate this potential moderating effect and explore the extent to which the felt power effect is a demand characteristic. Note that the Bayesian approach allows us to seamlessly update the evidence as more studies become available (e.g., Scheibehenne et al., 2017).

Our meta-analysis focused on the effect of power posing on feelings of subjective power and did not consider behavioral or hormonal measures. Nevertheless, we would like to emphasize that given a set of preregistered studies that include the behavioral and hormonal measures of interest, our methodology can readily be applied to quantify evidence in a coherent Bayesian way for those measures as well.

R scripts for reproducing the analyses presented in this chapter are available at <https://osf.io/r2cds/>.

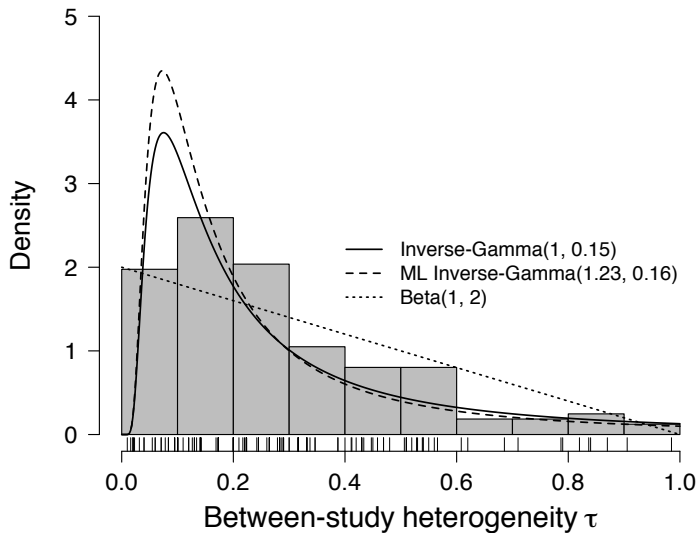


Figure 9.8: Distribution of the non-zero between-study standard deviations from meta-analyses reported in Psychological Bulletin (1990-2013; van Erp et al., 2017). The informed Inverse-Gamma(1,0.15) prior distribution is displayed on top as a solid line, the maximum-likelihood inverse-gamma distribution is depicted as a dashed line, and the Beta(1,2) distribution is depicted as a dotted line. Figure available at <http://tinyurl.com/k6yyz6b> under CC license <https://creativecommons.org/licenses/by/2.0/>.

9.A Robustness Check: Different Priors for the Between-Study Heterogeneity

Here we investigate whether and how the analyses results change under different priors for the between-study heterogeneity. Specifically, we explore two alternative prior choices to the Inverse-Gamma(1,0.15) prior: (1) the maximum-likelihood inverse-gamma distribution (depicted as a dashed line in Figure 9.8); and (2) a Beta(1,2) prior distribution (depicted as a dotted line in Figure 9.8). Table 9.2 displays the results for the reported data and Table 9.3 displays the results for the data of the subset of participants who were unfamiliar with the power pose effect: for all three prior choices for the between-study heterogeneity the results are highly similar.

Table 9.2: Meta-analytic Bayes factors ( $BF_{+0}$ ) for different prior choices for the between-study heterogeneity (reported data).

	<b>Inverse-Gamma(1, 0.15)</b>	<b>ML Inverse-Gamma</b>	<b>Beta(1, 2)</b>
meta-analytic fixed-effect Bayes factor	89.6	89.6	89.6
informed meta- analytic fixed- effect Bayes factor	191.8	191.8	191.8
meta-analytic random-effect Bayes factor	9.4	10.0	9.2
informed meta- analytic random- effect Bayes factor	20.7	22.0	20.2
meta-analytic model-averaged Bayes factor	33.1	32.1	35.1
informed meta- analytic model- averaged Bayes factor	71.4	69.1	75.5

Table 9.3: Meta-analytic Bayes factors ( $BF_{+0}$ ) for different prior choices for the between-study heterogeneity (unfamiliar participants).

	Inverse-Gamma(1, 0.15)	ML Inverse-Gamma	Beta(1, 2)
meta-analytic fixed-effect Bayes factor	4.4	4.4	4.4
informed meta- analytic fixed- effect Bayes factor	6.8	6.8	6.8
meta-analytic random-effect Bayes factor	1.6	1.7	1.7
informed meta- analytic random- effect Bayes factor	2.6	2.7	2.7
meta-analytic model-averaged Bayes factor	3.1	3.1	3.3
informed meta- analytic model- averaged Bayes factor	4.9	4.8	5.1

**Part III**

**Hypothesis Testing**



# Bayesian Evidence Accumulation in Experimental Mathematics: A Case Study of Four Irrational Numbers

---

## Abstract

Many questions in experimental mathematics are fundamentally inductive in nature. Here we demonstrate how Bayesian inference – the logic of partial beliefs – can be used to quantify the evidence that finite data provide in favor of a general law. As a concrete example we focus on the general law which posits that certain fundamental constants (i.e., the irrational numbers  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ ) are normal; specifically, we consider the more restricted hypothesis that each digit in the constant's decimal expansion occurs equally often. Our analysis indicates that for each of the four constants, the evidence in favor of the general law is overwhelming. We argue that the Bayesian paradigm is particularly apt for applications in experimental mathematics, a field in which the plausibility of a general law is in need of constant revision in light of data sets whose size is increasing continually and indefinitely.

## 10.1 Introduction

Experimental mathematics focuses on data and computation in order to address and discover mathematical questions that have so far escaped formal proof (D. H. Bailey & Borwein, 2009). In many cases, this means that mathematical conjectures are examined by studying their consequences for a large range of data;

---

This chapter is published as Gronau, Q. F., & Wagenmakers, E.-J. (2018). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*, 27, 277–286. doi: <https://doi.org/10.1080/10586458.2016.1256006>. Also available as *arXiv preprint*: <https://arxiv.org/abs/1602.03423>

every time a consequence is confirmed this increases one's confidence in the veracity of the conjecture. Complete confidence in the truth or falsehood of a conjecture can only be achieved with the help of a rigorous mathematical proof. Nevertheless, in between absolute truth and falsehood there exist partial beliefs, the intensity of which can be quantified using the rules of probability calculus (Borel, 1965; Ramsey, 1926).

Thus, an important role in experimental mathematics is played by heuristic reasoning and induction. Even in pure mathematics, inductive processes facilitate novel development:

“every mathematician with some experience uses readily and effectively the same method that Euler used which is basically the following: To examine a theorem  $T$ , we deduce from it some easily verifiable consequences  $C_1, C_2, C_3, \dots$ . If one of these consequences is found to be false, theorem  $T$  is refuted and the question is decided. But if all the consequences  $C_1, C_2, C_3, \dots$  happen to be valid, we are led after a more or less lengthy sequence of verifications to an ‘inductive’ conviction of the validity of theorem  $T$ . We attain a degree of belief so strong that it seems superfluous to make any ulterior verifications.” (Polya, 1941, pp. 455-456)

Here we illustrate how to formalize the process of induction for a venerable problem in experimental mathematics: we will quantify degree of belief in the statement that particular irrational numbers (i.e.,  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ ) are normal, or, more specifically, that the 10 digits of their decimal expansions occur equally often. This illustration does not address the more complicated question of whether all sequences of digits occur equally often: the sequence studied here is of length 1. Nevertheless, the simplified problem highlights the favorable properties of the general method and can be extended to more complicated scenarios.

To foreshadow the conclusion, our study shows that there is overwhelming evidence in favor of the general law that all digits in the decimal expansion of  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  occur equally often. Our statistical analysis improves on standard frequentist inference in several major ways that we elaborate upon below.

## 10.2 Bayes Factors to Quantify Evidence for General Laws

In experimental mathematics, the topic of interest often concerns the possible existence of a general law. This law – sometimes termed the null hypothesis  $\mathcal{H}_0$  – specifies an invariance (e.g.,  $\pi$  is normal) that imposes some sort of restriction on the data (e.g., the digits of the decimal expansion of  $\pi$  occur equally often). The negation of the general law – sometimes termed the alternative hypothesis  $\mathcal{H}_1$  – relaxes the restriction imposed by the general law.

In order to quantify the evidence that the data provide for or against a general law, Jeffreys (1961) developed a formal system of statistical inference whose

centerpiece is the following equation (Wrinch & Jeffreys, 1921, p. 387):

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \text{data})}{p(\mathcal{H}_1 \mid \text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}}_{\text{Bayes factor BF}_{01}}. \quad (10.1)$$

Jeffreys's work focused on the Bayes factor, which is the change from prior to posterior model odds brought about by the data. The Bayes factor also quantifies the relatively predictive adequacy of the models under consideration, and the log of the Bayes factor is the weight of evidence provided by the data (Kass & Raftery, 1995). When  $\text{BF}_{01} = 10$  this indicates that the data are 10 times more likely under  $\mathcal{H}_0$  than under  $\mathcal{H}_1$ ; when  $\text{BF}_{01} = .2$  this indicates that the data are 5 times more likely under  $\mathcal{H}_1$  than under  $\mathcal{H}_0$ .

Let  $\mathcal{H}_0$  be specified by a series of nuisance parameters  $\zeta$  and, crucially, a parameter of interest that is fixed at a specific value,  $\theta = \theta_0$ . Then  $\mathcal{H}_1$  is specified using similar nuisance parameters  $\zeta$ , but in addition  $\mathcal{H}_1$  releases the restriction on  $\theta$ . In order to obtain the Bayes factor one needs to integrate out the model parameters as follows:

$$\text{BF}_{01} = \frac{\int_{\mathcal{Z}} p(\text{data} \mid \theta_0, \zeta, \mathcal{H}_0) p(\zeta \mid \theta_0, \mathcal{H}_0) d\zeta}{\int_{\Theta} \int_{\mathcal{Z}} p(\text{data} \mid \theta, \zeta, \mathcal{H}_1) p(\theta, \zeta \mid \mathcal{H}_1) d\zeta d\theta}. \quad (10.2)$$

Equation 10.2 reveals several properties of Bayes factor inference that distinguish it from frequentist inference using  $p$  values. First, the Bayes factor contrasts two hypotheses, the general law and its negation. Consequently, it is possible to quantify evidence in favor of the general law (i.e., whenever  $\text{BF}_{01} > 1$ ). As we will see below, one of our tests for the first 100 million digits of  $\pi$  produces  $\text{BF}_{01} = 1.86 \times 10^{30}$ , which is overwhelming evidence in favor of the law that the digits of the decimal expansion of  $\pi$  occur equally often; in contrast, a non-significant  $p$  value can only suggest a failure to reject  $\mathcal{H}_0$  (e.g., Frey, 2009). Moreover, as we will demonstrate below, the evidential meaning of a  $p$  value changes with sample size (Lindley, 1957). This is particularly problematic for the study of the behavior of decimal expansions, since there can be as many as 10 trillion digits under consideration.

Second, the Bayes factor respects the probability calculus and allows coherent updating of beliefs; specifically, consider two batches of data,  $y_1$  and  $y_2$ . Then,  $\text{BF}_{01}(y_1, y_2) = \text{BF}_{01}(y_1) \times \text{BF}_{01}(y_2 \mid y_1)$ : the Bayes factor for the joint data set can be decomposed as the product of the Bayes factor for the first batch multiplied by the Bayes factor for the second batch, conditional on the information obtained from the first data set. Consequently – and in contrast to  $p$  value inference – Bayes factors can be seamlessly updated as new data arrive, indefinitely and without a well-defined sampling plan (Berger & Berry, 1988a, 1988b). This property is particularly relevant for the study of normality of fundamental constants, since new computational and mathematical developments continually increase the length of the decimal expansion (Wrench Jr, 1960).

### 10.3 The Normality of Irrational Numbers

A real number  $x$  is normal in base  $b$  if all of the digit sequences in its base  $b$  expansion occur equally often (e.g., Borel, 1909); consequently, each string of  $t$  consecutive digits has limiting frequency  $b^{-t}$ . In our example, we consider the decimal expansion and focus on strings of length 1. Hence, normality entails that each digit occurs with limiting frequency  $1/10$ .

The conjecture that certain fundamental constants – irrational numbers such as  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  – are normal has attracted much scientific scrutiny (e.g., D. H. Bailey & Borwein, 2009; D. H. Bailey & Crandall, 2001; Borwein, Bailey, & Bailey, 2004). Aside from theoretical interest and practical application, the enduring fascination with this topic may be due in part to the paradoxical result that the digits sequences are perfectly predictable yet apparently appear random:

“Plenty of arrangements in which design had a hand [...] would be quite indistinguishable in their results from those in which no design whatever could be traced. Perhaps the most striking case in point here is to be found in the arrangement of the digits in one of the natural arithmetical constants, such as  $\pi$  or  $e$ , or in a table of logarithms. If we look to the process of production of these digits, no extremer instance can be found of what we mean by the antithesis of randomness: every figure has its necessarily pre-ordained position, and a moment’s flagging of intention would defeat the whole purpose of the calculator. And yet, if we look to results only, no better instance can be found than one of these rows of digits if it were intended to illustrate what we practically understand by a chance arrangement of a number of objects. Each digit occurs approximately equally often, and this tendency develops [sic] as we advance further [...] In fact, if we were to take the whole row of hitherto calculated figures, cut off the first five as familiar to us all, and contemplate the rest, no one would have the slightest reason to suppose that these had not come out as the results of a die with ten equal faces.” (Venn, 1888, p. 111)

But are constants such as  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  truly normal? Intuitive arguments suggest that normality must be the rule (Venn, 1888, pp. 111-115) but so far the problem has eluded a rigorous mathematical proof. In lieu of such a proof, research in experimental mathematics has developed a wide range of tests to assess whether or not the hypothesis of normality can be rejected (e.g., D. H. Bailey et al., 2012; Frey, 2009; Ganz, 2014; Jaditz, 2000; Marsaglia, 2005; Tu & Fischbach, 2005, p. 281), some of which involve visual methods of data presentation (e.g., Aragón Artacho, Bailey, Borwein, & Borwein, 2012; Venn, 1888, p. 118). In line with Venn’s conjecture, most tests conclude that for the constants under investigation, the hypothesis of normality cannot be rejected.

However, to the best of our knowledge only one study has tried to quantify the strength of inductive support in favor of normality (i.e., D. H. Bailey et al., 2012). Below we outline a multinomial Bayes factor test of equivalence that allows one to quantify the evidence in favor of the general law that each digit occurs equally often.

## 10.4 A Bayes Factor Multinomial Test for Normality

The general law or null hypothesis  $\mathcal{H}_0$  states that  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  are normal. Here we consider the more restricted law that each digit in the decimal expansion occurs equally often (i.e., we focus on series of length 1 only). Hence,  $\mathcal{H}_0$  stipulates that  $\theta_{0j} = \frac{1}{10} \forall j \in \{0, 1, \dots, 9\}$ , where  $j$  indexes the digits.

Next we need to specify our expectations under  $\mathcal{H}_1$ , that is, our beliefs about the distribution of digit occurrences under the assumption that the general law does not hold, and before having seen actual data. We explore two alternative models. The first model assigns the digit probabilities  $\theta_j$  an uninformative Dirichlet prior  $D(\mathbf{a} = 1)$ ; under this alternative hypothesis  $\mathcal{H}_1^{\mathbf{a}=1}$ , all combinations of digit probabilities are equally likely a priori. In other words, the predictions of  $\mathcal{H}_1^{\mathbf{a}=1}$  are relatively imprecise. The second model assigns the digit probabilities  $\theta_j$  an informative Dirichlet prior  $D(\mathbf{a} = 50)$ ; under this alternative hypothesis  $\mathcal{H}_1^{\mathbf{a}=50}$ , the predictions of  $\mathcal{H}_1^{\mathbf{a}=50}$  are relatively precise, and similar to those made by  $\mathcal{H}_0$ . In effect, the predictions from  $\mathcal{H}_1^{\mathbf{a}=50}$  are the same as those made by a model that is initialized with an uninformative Dirichlet prior  $D(\mathbf{a} = 1)$  which is then updated based on 49 hypothetical occurrences for each of the ten digits, that is, a hypothetical sequence of a total of 490 digits that corresponds perfectly with  $\mathcal{H}_0$ .

Thus, model  $\mathcal{H}_1^{\mathbf{a}=1}$  yields predictions that are relatively imprecise, whereas model  $\mathcal{H}_1^{\mathbf{a}=50}$  yields predictions that are relatively precise. The Bayes factor for  $\mathcal{H}_0$  versus  $\mathcal{H}_1$  is an indication of relative predictive adequacy, and by constructing two very different versions of  $\mathcal{H}_1$  – one predictively dissimilar to  $\mathcal{H}_0$ , one predictively similar – our analysis captures a wide range of plausible outcomes (e.g., Spiegelhalter, Freedman, & Parmar, 1994).

With  $\mathcal{H}_0$  and  $\mathcal{H}_1$  specified, the Bayes factor for the multinomial test of equivalence (O’Hagan & Forster, 2004, p. 350) is given by

$$\begin{aligned} \text{BF}_{01} &= \frac{B(\mathbf{a})}{B(\mathbf{a} + \mathbf{n})} \prod_{j=0}^9 \theta_{0j}^{n_j} \\ &= \frac{B(\mathbf{a})}{B(\mathbf{a} + \mathbf{n})} \prod_{j=0}^9 10^{-n_j}, \end{aligned} \quad (10.3)$$

where  $\mathbf{a}$  and  $\mathbf{n}$  are vectors of length ten (i.e., the number of different digits); the elements of  $\mathbf{n}$  contain the number of occurrences for each of the ten digits. Finally,  $B(\cdot)$  is a generalization of the beta distribution (O’Hagan & Forster, 2004, p. 341):

$$B(\mathbf{a}) = \frac{\prod_{j=0}^9 \Gamma(a_j)}{\Gamma\left(\sum_{j=0}^9 a_j\right)}, \quad (10.4)$$

where  $\Gamma(t)$  is the gamma function defined as  $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ . For computational convenience we use the natural logarithm of the Bayes factor:

$$\log \text{BF}_{01} = \log B(\mathbf{a}) - \log B(\mathbf{a} + \mathbf{n}) - N \log 10, \quad (10.5)$$

where  $N$  is the total number of observed digits.

### 10.4.1 Example 1: The Case of $\pi$

In our first example we compute multinomial Bayes factors for the digits of  $\pi$ . We compute the Bayes factor sequentially, as a function of an increasing number of available digits, with an upper bound of 100 million. Figure 10.1 displays the results in steps of 1,000 digits. The Bayes factor that contrasts  $\mathcal{H}_0$  versus  $\mathcal{H}_1^{a=1}$  is indicated by the black line, and it shows that the evidence increasingly supports the general law. After all 100 million digits have been taken into account, the observed data are  $1.86 \times 10^{30}$  times more likely to occur under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{a=1}$ . The extent of this support is overwhelming. The red line indicates the maximum Bayes factor, that is, the Bayes factor that is obtained in case the digits were to occur equally often – that is, hypothetical data perfectly consistent with  $\mathcal{H}_0$ .

The dark grey area in Figure 10.1 indicates where a frequentist  $p$  value hypothesis test would fail to reject the null hypothesis. This area was determined in two steps. First, we considered the hypothetical distribution of counts across the ten digit categories and constructed a threshold data set for which  $\mathcal{H}_0$  has a 5% chance of producing outcomes that are at least as extreme. Second, this threshold data set was used to compute a Bayes factor, and this threshold Bayes factor is plotted in Figure 10.1 as the lower bound of the dark grey area.

In order to construct the threshold data set, the number of counts in each digit category was obtained as follows. In this multinomial scenario there are nine degrees of freedom. Without loss of generality, the number of counts in the first eight of ten categories may be set equal to the expected frequency of  $\frac{N}{10}$ :  $n_0, n_1, \dots, n_7 = \frac{N}{10}$ . Consequently, the first eight summands of the  $\chi^2$ -test formula are equal to zero. Furthermore,  $\sum_{j=0}^9 n_j = N$ , so that if  $n_8$  is known,  $n_9$  is determined by  $n_9 = \frac{2}{10}N - n_8$ . We then obtain the number of counts in the ninth category  $n_8$  by solving the following quadratic equation for  $n_8$ :

$$\chi_{95\%}^2 = \frac{(n_8 - \frac{N}{10})^2}{N/10} + \frac{((\frac{2}{10}N - n_8) - \frac{N}{10})^2}{N/10}, \quad (10.6)$$

where  $\chi_{95\%}^2$  denotes the 95-th percentile of the  $\chi^2$  distribution with nine degrees of freedom.

Figure 10.1 shows that the height of the dark grey area's lower bound increases with  $N$ . This means that it is possible to encounter a data set for which the Bayes factor indicates overwhelming evidence in favor of  $\mathcal{H}_0$ , whereas the fixed- $\alpha$  frequentist hypothesis test suggests that  $\mathcal{H}_0$  ought to be rejected. In this way Figure 10.1 provides a visual illustration of the Jeffreys-Lindley paradox (Jeffreys, 1961; Lindley, 1957), a paradox that will turn out to be especially relevant for the later analysis of  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ .

A qualitative similar pattern of results is apparent when we consider the grey line in Figure 10.1: the Bayes factor that contrasts  $\mathcal{H}_0$  versus  $\mathcal{H}_1^{a=50}$ . Because this model makes predictions that are relatively similar to those of  $\mathcal{H}_0$ , the data are less diagnostic than before. Nevertheless, the evidence increasingly supports the general law. After all 100 million digits are observed, the observed data are  $\text{BF}_{01} = 1.92 \times 10^{22}$  times more likely to occur under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{a=50}$ . The extent of this support remains overwhelming.

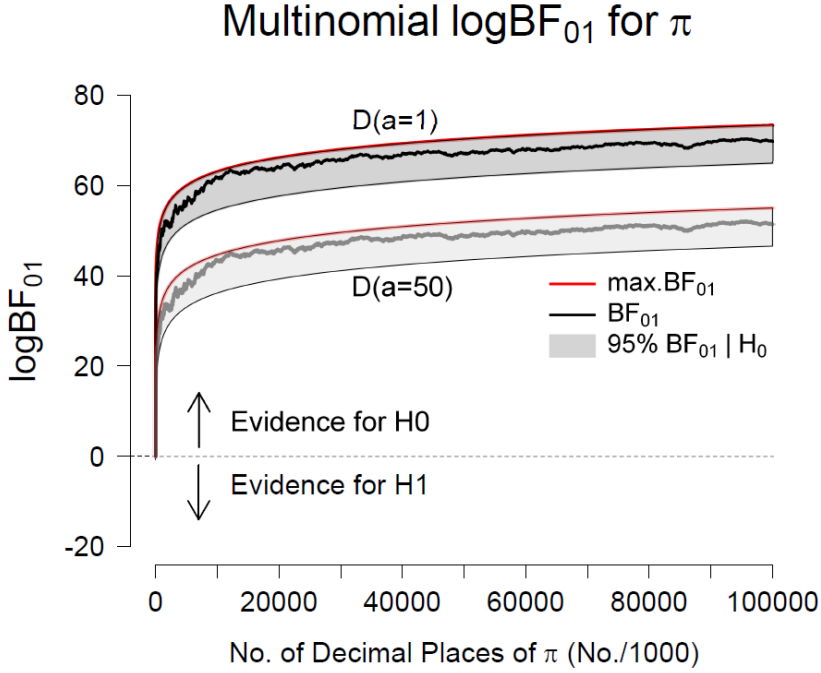


Figure 10.1: Sequential Bayes factors in favor of equal occurrence probabilities based on the first 100 million digits of  $\pi$ . The results in the top part of the panel correspond to an uninformative  $D(\mathbf{a} = 1)$  prior for the alternative hypothesis; the results in the lower part of the panel correspond to the use of an informative  $D(\mathbf{a} = 50)$  prior. The red lines indicate the maximum possible evidence for  $\mathcal{H}_0$ , and the grey areas indicate where 95% of the Bayes factors would fall if  $\mathcal{H}_0$  were true. After 100 million digits, the final Bayes factor under a  $D(\mathbf{a} = 1)$  prior is  $\text{BF}_{01} = 1.86 \times 10^{30}$  ( $\log \text{BF}_{01} = 69.70$ ); under a  $D(\mathbf{a} = 50)$  prior, the final Bayes factor equals  $\text{BF}_{01} = 1.92 \times 10^{22}$  ( $\log \text{BF}_{01} = 51.31$ ). Figure available at <http://tinyurl.com/zelm4o4> under CC license <https://creativecommons.org/licenses/by/2.0/>.

For completeness, we also computed Bayes factors based on the first trillion decimal digits of  $\pi$  as reported in D. H. Bailey and Borwein (2009, p. 11) (not shown). As expected from the upward evidential trajectories in Figure 10.1, increasing the sequence length strengthens the support in favor of the general law: based on one trillion decimal digits, the  $D(\mathbf{a} = 1)$  prior for  $\mathcal{H}_1$  yields  $\text{BF}_{01} = 3.65 \times 10^{46}$  ( $\log \text{BF}_{01} = 107.29$ )<sup>1</sup>, and the  $D(\mathbf{a} = 50)$  prior yields  $\text{BF}_{01} = 4.07 \times 10^{38}$  ( $\log \text{BF}_{01} = 88.90$ ).

Finally, consider the fact that the two evidential trajectories – one for a comparison against  $\mathcal{H}_1^{\mathbf{a}=1}$ , one for a comparison against  $\mathcal{H}_1^{\mathbf{a}=50}$  – have a similar shape

<sup>1</sup>Such an excessive degree of evidence in favor of a general law may well constitute a world record.

and appear to differ only by a constant factor. This pattern is not a coincidence, and it follows from the nature of sequential updating for Bayes factors (Jeffreys, 1961, p. 334). Recall that there exist two mathematically equivalent ways to update the Bayes factor when new data  $y_2$  appear. The first method is to compute a single new Bayes factor using all of the available observations,  $\text{BF}(y = y_1, y_2)$ ; the second method is to compute a Bayes factor only for the new data, but based on the posterior distribution that is the result of having encountered the previous data – this Bayes factor,  $\text{BF}(y_2 | y_1)$  is then multiplied by the Bayes factor for the old data,  $\text{BF}(y_1)$  to yield the updated Bayes factor  $\text{BF}(y = y_1, y_2)$ .

Now let  $y_1$  denote a starting sequence of digits large enough so that the joint posterior distribution for the  $\theta_j$ 's under  $\mathcal{H}_1^{a=1}$  is relatively similar to that under  $\mathcal{H}_1^{a=50}$  (i.e., when the data are said to have overwhelmed the prior). From that point onward, the change in the Bayes factor as a result of new data  $y_2$ ,  $\text{BF}(y_2 | y_1)$ , will be virtually identical for both instantiations of  $\mathcal{H}_1$ . Hence, following an initial phase of posterior convergence, the subsequent evidential updates are almost completely independent of the prior distribution on the model parameters.<sup>2</sup>

Equation 10.1 shows that the Bayes factor quantifies the change in belief brought about by the data; as a first derivative of belief (expressed on the log scale), it achieves independence of the prior model log odds. In turn, Figure 10.1 illustrates that the change in the log Bayes factor – the second derivative of belief – achieves independence of the prior distribution on the model parameters, albeit only in the limit of large samples.

The next three cases concern a study of the irrational numbers  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ ; the analysis and conclusion for these cases echo the ones for the case of  $\pi$ .

### 10.4.2 Example 2: The Case of $e$

In our second example we compute multinomial Bayes factors for the digits of the base of the natural logarithm: Euler's number  $e$ . Proceeding in similar fashion as for the case of  $\pi$ , Figure 10.2 shows the evidential trajectories (in steps of 1,000 digits) for the first 100 million digits of  $e$ .<sup>3</sup> As was the case for  $\pi$ , the upward trajectories signal an increasing degree of support in favor of the general law. After all 100 million digits have been taken into account, the observed data are  $2.61 \times 10^{30}$  times more likely to occur under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{a=1}$ , and  $2.69 \times 10^{22}$  times more likely under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{a=50}$ . Again, the extent of this support is overwhelming.

Note that, as for the case of  $\pi$ , the two evidential trajectories – one for a comparison against  $\mathcal{H}_1^{a=1}$ , one for a comparison against  $\mathcal{H}_1^{a=50}$  – have a similar shape and appear to differ only by a constant factor. In contrast to the case of  $\pi$ , however, the Jeffreys-Lindley paradox is more than just a theoretical possibility: Figure 10.2 shows that the evidential trajectories move outside the grey area when

---

<sup>2</sup>That is, after a sufficient number of observations, the trajectories of the log Bayes factors for the different priors for  $\mathcal{H}_1$  are equal, only shifted by a constant. In fact, regardless of the irrational number under consideration, this constant – which corresponds to the difference in  $\log(\text{BF}_{01}^{a=1})$  and  $\log(\text{BF}_{01}^{a=50})$  – approaches 18.39 (for a derivation see the appendix).

<sup>3</sup>Data were obtained using the `pifast` software (<http://numbers.computation.free.fr/Constants/PiProgram/pifast.html>).

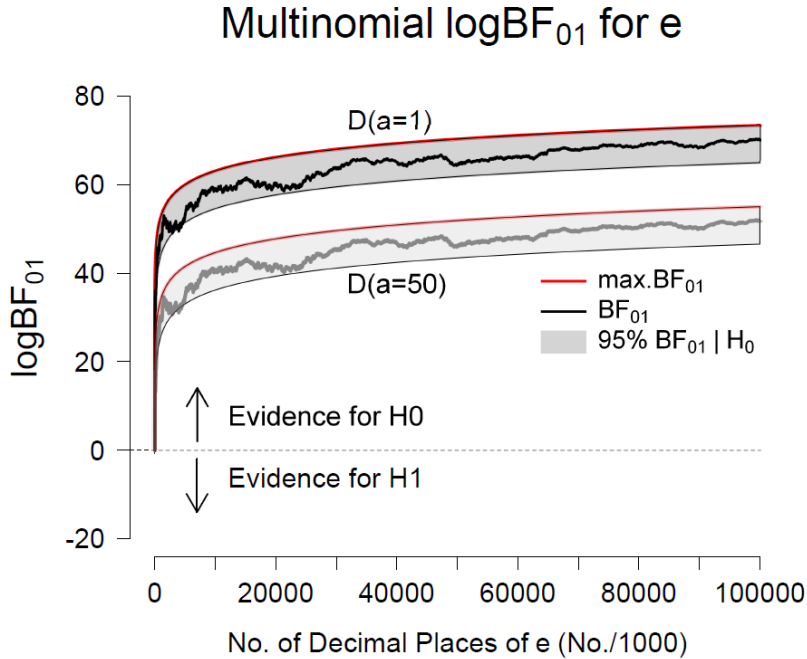


Figure 10.2: Sequential Bayes factors in favor of equal occurrence probabilities based on the first 100 million digits of  $e$ . The results in the top part of the panel correspond to an uninformative  $D(\mathbf{a} = 1)$  prior for the alternative hypothesis; the results in the lower part of the panel correspond to the use of an informative  $D(\mathbf{a} = 50)$  prior. The red lines indicate the maximum possible evidence for  $\mathcal{H}_0$ , and the grey areas indicate where 95% of the Bayes factors would fall if  $\mathcal{H}_0$  were true. After 100 million digits, the final Bayes factor under a  $D(\mathbf{a} = 1)$  prior is  $\text{BF}_{01} = 2.61 \times 10^{30}$  ( $\log \text{BF}_{01} = 70.04$ ); under a  $D(\mathbf{a} = 50)$  prior, the final Bayes factor equals  $\text{BF}_{01} = 2.69 \times 10^{22}$  ( $\log \text{BF}_{01} = 51.65$ ). Figure available at <http://tinyurl.com/h3wenqo> under CC license <https://creativecommons.org/licenses/by/2.0/>.

the total digit count is between 82,100 and 254,000, meaning that for those digit counts the frequentist hypothesis test (with a fixed  $\alpha$ -level of .05) suggests that  $\mathcal{H}_0$  ought to be rejected. For the same data, both Bayes factors indicate compelling evidence in favor of  $\mathcal{H}_0$ .<sup>4</sup>

<sup>4</sup>A frequentist statistician may object that this is a sequential design whose proper analysis demands a correction of the  $\alpha$  level. However, the same data may well occur in a fixed sample size design. In addition, the frequentist correction of  $\alpha$  levels is undefined when the digit count increases indefinitely.

### 10.4.3 Example 3: The Case of $\sqrt{2}$

In our third example we compute multinomial Bayes factors for the digits of  $\sqrt{2}$ . Proceeding in similar fashion as above, Figure 10.3 shows the evidential trajectories (in steps of 1,000 digits) for the first 100 million digits of  $\sqrt{2}$ .<sup>5</sup> As was the case for  $\pi$  and  $e$ , upward evidential trajectories reveal an increasing degree of support in favor of the general law. After all 100 million digits have been taken into account, the observed data are  $7.29 \times 10^{30}$  times more likely to occur under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{\alpha=1}$ , and  $7.52 \times 10^{22}$  times more likely under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{\alpha=50}$ . As before, the extent of this support is overwhelming.

As Figure 10.3 shows, the analysis of  $\sqrt{2}$  provides yet another demonstration of the Jeffreys-Lindley paradox: when the total digit count ranges between 1 million and 2 million, and between 20 and 40 million (especially close to 40 million), a frequentist analysis occasionally rejects  $\mathcal{H}_0$  at an  $\alpha$ -level of .05 (i.e., the evidential trajectories temporarily leave the grey area) whereas, for the same data, both Bayes factors indicate compelling evidence in favor of  $\mathcal{H}_0$ .

### 10.4.4 Example 4: The Case of $\ln 2$

In our fourth and final example we compute multinomial Bayes factors for the digits of  $\ln 2$ . Figure 10.4 shows the evidential trajectories (in steps of 1,000 digits) for the first 100 million digits of  $\ln 2$ .<sup>6</sup> As was the case for  $\pi$ ,  $e$ , and  $\sqrt{2}$ , upward trajectories reflect the increasing degree of support in favor of the general law. After all 100 million digits have been taken into account, the observed data are  $7.58 \times 10^{29}$  times more likely to occur under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{\alpha=1}$ , and  $7.81 \times 10^{21}$  times more likely under  $\mathcal{H}_0$  than under  $\mathcal{H}_1^{\alpha=50}$ . As Figure 10.4 shows, the analysis of  $\ln 2$  provides again a demonstration of the Jeffreys-Lindley paradox: the evidential trajectories leave the grey area multiple times indicating that a frequentist analysis rejects  $\mathcal{H}_0$  at an  $\alpha$ -level of .05 whereas, for the same data, both Bayes factors indicate compelling evidence in favor of  $\mathcal{H}_0$ .

## 10.5 Alternative Analysis

The analyses presented so far used two different Dirichlet distributions as a prior for the parameter vector under the alternative hypothesis  $\mathcal{H}_1$ . In this way, we demonstrated that the results do not change qualitatively when considering an uninformed or an informed Dirichlet prior distribution. A Dirichlet distribution is commonly used as a prior distribution for the parameter vector of a multinomial likelihood since it conveniently leads to an analytical solution for the Bayes factor.

However, one might ask whether the results are sensitive to the particular choice of the *family* of prior distributions used to specify the alternative hypothesis  $\mathcal{H}_1$ , that is the family of Dirichlet distributions. To highlight the robustness of our conclusion, we present the results of an analysis that is based on a more flexible

---

<sup>5</sup>Data were obtained using the `pifast` software (<http://numbers.computation.free.fr/Constants/PiProgram/pifast.html>).

<sup>6</sup>Data were obtained using the `pifast` software (<http://numbers.computation.free.fr/Constants/PiProgram/pifast.html>).

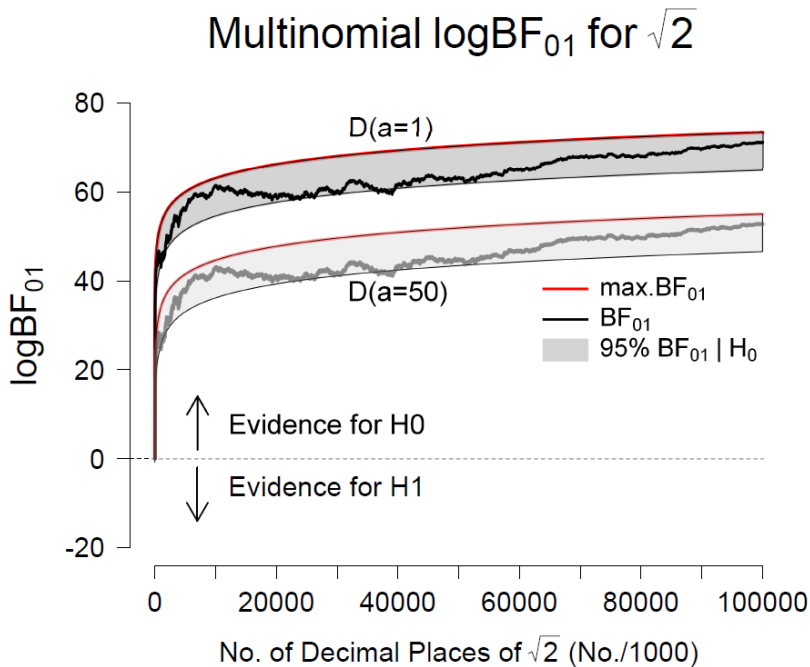


Figure 10.3: Sequential Bayes factors in favor of equal occurrence probabilities based on the first 100 million digits of  $\sqrt{2}$ . The results in the top part of the panel correspond to an uninformative  $D(\mathbf{a} = 1)$  prior for the alternative hypothesis; the results in the lower part of the panel correspond to the use of an informative  $D(\mathbf{a} = 50)$  prior. The red lines indicate the maximum possible evidence for  $\mathcal{H}_0$ , and the grey areas indicate where 95% of the Bayes factors would fall if  $\mathcal{H}_0$  were true. After 100 million digits, the final Bayes factor under a  $D(\mathbf{a} = 1)$  prior is  $\text{BF}_{01} = 7.29 \times 10^{30}$  ( $\log \text{BF}_{01} = 71.06$ ); under a  $D(\mathbf{a} = 50)$  prior, the final Bayes factor equals  $\text{BF}_{01} = 7.52 \times 10^{22}$  ( $\log \text{BF}_{01} = 52.67$ ). Figure available at <http://tinyurl.com/jgwu523> under CC license <https://creativecommons.org/licenses/by/2.0/>.

prior distribution than the Dirichlet distribution, namely a two component mixture of Dirichlet distributions. Mixture distributions have the property that the shape of the density is extremely flexible and can easily account for skewness, excess kurtosis, and even multi-modality (Frühwirth-Schnatter, 2006) which makes them an ideal candidate for testing the sensitivity to a wide range of prior distributions. As Dalal and Hall (1983) showed, in fact *any* prior distribution may be arbitrarily closely approximated by a suitable mixture of *conjugate* prior distributions (i.e., prior distributions that, combined with a certain likelihood, lead to a posterior distribution that is in the same family of distributions as the prior distribution).<sup>7</sup>

<sup>7</sup>Of course, in some cases this may require a very “rich” mixture, that is, a mixture prior with many components.

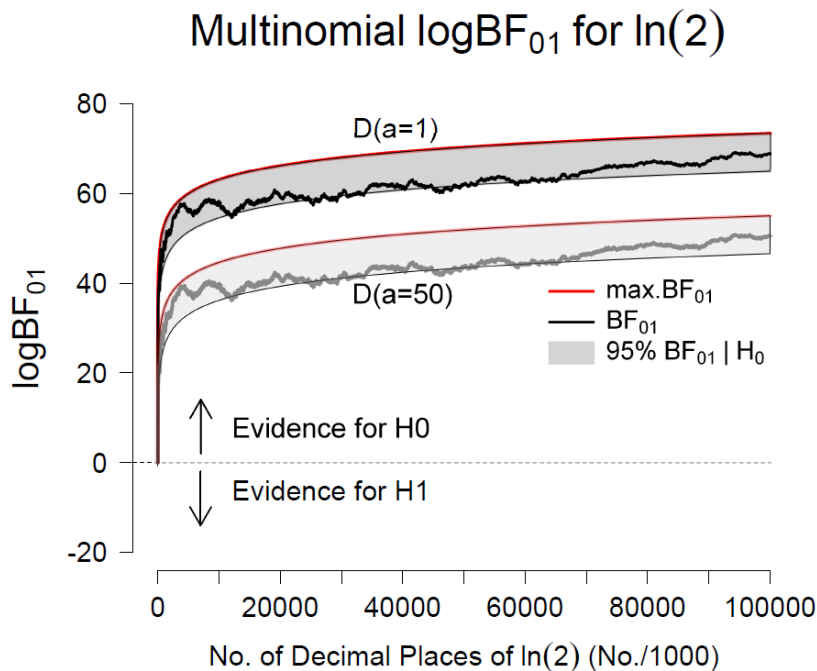


Figure 10.4: Sequential Bayes factors in favor of equal occurrence probabilities based on the first 100 million digits of  $\ln 2$ . The results in the top part of the panel correspond to an uninformative  $D(\mathbf{a}=1)$  prior for the alternative hypothesis; the results in the lower part of the panel correspond to the use of an informative  $D(\mathbf{a}=50)$  prior. The red lines indicate the maximum possible evidence for  $\mathcal{H}_0$ , and the grey areas indicate where 95% of the Bayes factors would fall if  $\mathcal{H}_0$  were true. After 100 million digits, the final Bayes factor under a  $D(\mathbf{a}=1)$  prior is  $\text{BF}_{01} = 7.58 \times 10^{29}$  ( $\log \text{BF}_{01} = 68.80$ ); under a  $D(\mathbf{a}=50)$  prior, the final Bayes factor equals  $\text{BF}_{01} = 7.81 \times 10^{21}$  ( $\log \text{BF}_{01} = 50.41$ ). Figure available at <http://tinyurl.com/jqdyd3w> under CC license <https://creativecommons.org/licenses/by/2.0/>.

As an example, we considered a two component mixture of a  $D(\mathbf{a}_1 = 5)$  Dirichlet distribution which assigns more mass to probability vectors that have components that are similar to each other (i.e., similar digit probabilities) and a  $D(\mathbf{a}_2 = 1/5)$  Dirichlet distribution which assigns more mass to the corners of the simplex (i.e., one digit probability dominates) where the mixing weight was equal to  $w = 0.5$ .<sup>8</sup> It is easily shown that also under this prior choice, the Bayes factor is available analytically. Recall that the Bayes factor is defined as  $\text{BF}_{01} = \frac{p(\text{data}|\mathcal{H}_0)}{p(\text{data}|\mathcal{H}_1)}$ .  $p(\text{data} | \mathcal{H}_0)$  is obtained by inserting  $\theta_{0j} = \frac{1}{10} \forall j \in \{0, 1, \dots, 9\}$  into the multino-

<sup>8</sup>R code that allows one to explore how the results change for a different choice of a two component Dirichlet mixture prior is available on the Open Science Framework under <https://osf.io/cmn2z/>.

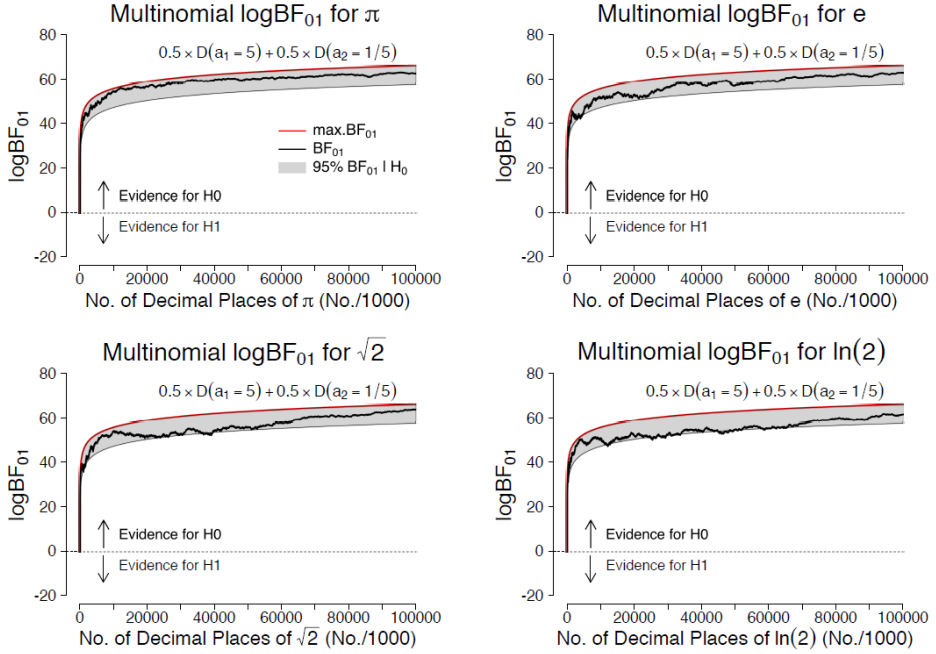


Figure 10.5: Sequential Bayes factors in favor of equal occurrence probabilities based on the first 100 million digits of  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ . The results correspond to the use of a two component mixture prior of a  $D(\mathbf{a}_1 = 5)$  and  $D(\mathbf{a}_2 = 1/5)$  Dirichlet distribution where the mixing weight was equal to  $w = 0.5$ . The red lines indicate the maximum possible evidence for  $\mathcal{H}_0$ , and the grey areas indicate where 95% of the Bayes factors would fall if  $\mathcal{H}_0$  were true. Figure available at <http://tinyurl.com/hw4gmlr> under CC license <https://creativecommons.org/licenses/by/2.0/>.

mial likelihood. In order to obtain  $p(\text{data} \mid \mathcal{H}_1)$ , we use the fact that any mixture of conjugate prior distributions is itself conjugate, that is, leads to a posterior distribution that is again a mixture of the same family of distributions, only with updated parameters (Dalal & Hall, 1983). Hence, since the Dirichlet distribution is conjugate to the multinomial likelihood, the posterior distribution when using a mixture of Dirichlet distributions as a prior is again a mixture of Dirichlet distributions (with updated parameters). This implies that we know the normalizing constant of the posterior distribution under the alternative hypothesis  $\mathcal{H}_1$  which

## 10. BAYESIAN EVIDENCE ACCUMULATION IN EXPERIMENTAL MATHEMATICS: A CASE STUDY OF FOUR IRRATIONAL NUMBERS

is equivalent to  $p(\text{data} \mid \mathcal{H}_1)$ . Hence, we can calculate the Bayes factor as follows:

$$\begin{aligned}
 \text{BF}_{01} &= \frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)} \\
 &= \frac{\frac{N!}{n_0!n_1!\dots n_9!} \prod_{j=0}^9 \theta_{0j}^{n_j}}{\int_{\Theta} \frac{N!}{n_0!n_1!\dots n_9!} \prod_{j=0}^9 \theta_j^{n_j} (w \frac{1}{B(\mathbf{a}_1)} \prod_{j=0}^9 \theta_j^{a_{1j}-1} + (1-w) \frac{1}{B(\mathbf{a}_2)} \prod_{j=0}^9 \theta_j^{a_{2j}-1}) d\theta} \\
 &= \frac{\prod_{j=0}^9 \theta_{0j}^{n_j}}{w \frac{1}{B(\mathbf{a}_1)} \int_{\Theta} \prod_{j=0}^9 \theta_j^{a_{1j}+n_j-1} d\theta + (1-w) \frac{1}{B(\mathbf{a}_2)} \int_{\Theta} \prod_{j=0}^9 \theta_j^{a_{2j}+n_j-1} d\theta} \\
 &= \frac{\prod_{j=0}^9 \theta_{0j}^{n_j}}{w \frac{B(\mathbf{a}_1+\mathbf{n})}{B(\mathbf{a}_1)} + (1-w) \frac{B(\mathbf{a}_2+\mathbf{n})}{B(\mathbf{a}_2)}}.
 \end{aligned} \tag{10.7}$$

Figure 10.5 displays the results for the 100 million digits of the four irrational numbers that are based on the two component mixture prior described above. For  $\pi$ , the final Bayes factor equals  $1.41 \times 10^{27}$ ; for  $e$ , the final Bayes factor equals  $1.97 \times 10^{27}$ ; for  $\sqrt{2}$ , the final Bayes factor equals  $5.52 \times 10^{27}$ ; for  $\ln 2$  the final Bayes factor equals  $5.73 \times 10^{26}$ .

The results based on the mixture prior are very similar to the previous ones, that is, we again obtain overwhelming support in favor of the assumption that all digits occur equally often; hence, we conclude that inference appears to be relatively robust to the particular choice of prior distribution that is used.

## 10.6 Discussion and Conclusion

With the help of four examples we illustrated how Bayesian inference can be used to quantify evidence in favor of a general law (Jeffreys, 1961). Specifically, we examined the degree to which the data support the conjecture that the digits in the decimal expansion of  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  occur equally often. Our main analysis featured two prior distributions used to instantiate models as alternatives to the general law: the alternative model  $\mathcal{H}_1^{\alpha=50}$  resembled the general law, whereas the alternative model  $\mathcal{H}_1^{\alpha=1}$  did not. An infinite number of plausible alternatives and associated inferences lie in between these two extremes. Regardless of whether the comparison involved  $\mathcal{H}_1^{\alpha=50}$  or  $\mathcal{H}_1^{\alpha=1}$ , the evidence was always compelling and the sequential analysis produced evidential trajectories that reflected increasing support in favor of the general law. Future data can update the evidence and extend these trajectories indefinitely.

Figures 10.1–10.4 clearly show the different outcomes for  $\mathcal{H}_1^{\alpha=50}$  versus  $\mathcal{H}_1^{\alpha=1}$ . This dependence on the model specification is sometimes felt to be a weakness of the Bayesian approach, as the specification of the prior distribution for the model parameters is not always straightforward or objective. However, the dependence on the prior distribution is also a strength, as it allows the researcher to insert relevant information into the model to devise a test that more closely represents the underlying theory. Does it make sense to assign the model parameters a Dirichlet

$D(\mathbf{a} = 50)$  prior? It is easy to use existing knowledge about the distribution of trillions of digits for  $\pi$  to argue that this Dirichlet distribution is overly wide and hence inappropriate; however, this conclusion confuses prior knowledge with posterior knowledge – as the name implies, the prior distribution should reflect our opinion before and not after the data have been observed.

In the present work we tried to alleviate concerns about the sensitivity to the prior specification in three ways. First, for our main analysis, we used a sandwich approach in which we examined the results for two very different prior distributions, thereby capturing a wide range of outcomes for alternative specifications (e.g., Spiegelhalter et al., 1994). Second, we considered a different, very flexible family of alternative prior distributions (i.e., a two component mixture of Dirichlet distributions) and we demonstrated that the results do not change qualitatively – the evidence in favor of the general law remains overwhelming. Third, we have shown that the second derivative of belief – the change in the Bayes factor as a result of new data – becomes insensitive to the prior specification as  $N$  grows large. Here, the evidential trajectories all suggest that the evidence for the general law increases as more digits become available. Figure 10.6 displays the results for  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  side-by-side and emphasizes that for all four irrational numbers that we investigated, we obtain similar overwhelming support for the general law which states that all digits occur equally often – this is the case for all three prior distributions that we considered.

A remaining concern is that our Dirichlet prior on  $\mathcal{H}_1^{a=50}$  may be overly wide and therefore bias the test in favor of the general law. To assess the validity of this concern we conducted a simulation study in which the normality assumption was violated: one digit was given an occurrence probability of .11, whereas each of the remaining digits were given occurrence probabilities of .89/9. Figure 10.7 shows that for all 1,000 simulated data sets, the evidential trajectories indicate increasing evidence against the general law. After 1 million digits, the average Bayes factor in favor of the alternative hypothesis is  $\text{BF}_{10} = 1.19 \times 10^{214}$  ( $\log \text{BF}_{10} = 492.93$ ) under the  $D(\mathbf{a} = 1)$  prior and  $\text{BF}_{10} = 8.88 \times 10^{221}$  ( $\log \text{BF}_{10} = 511.05$ ) under the  $D(\mathbf{a} = 50)$  prior. Thus, with our instantiations of  $\mathcal{H}_1$  the Bayes factor is able to provide overwhelming evidence against the general law when it is false.

One of the main challenges for Bayesian inference in the study of normality for fundamental constants is to extend the simple multinomial approach presented here to account for longer digit sequences. As the digit series grows large, the number of multinomial categories also grows while the number of unique sequences decreases. Ultimately, this means that even with trillions of digits, a test for normality may lack the data for a diagnostic test. Nevertheless, alternative models of randomness can be entertained and given a Bayesian implementation – once this is done, the principles outlined by Jeffreys can be used to quantify the evidence for or against the general law.

The Supplemental Materials can be found at: <https://osf.io/5ysiu/>.

# 10. BAYESIAN EVIDENCE ACCUMULATION IN EXPERIMENTAL MATHEMATICS: A CASE STUDY OF FOUR IRRATIONAL NUMBERS

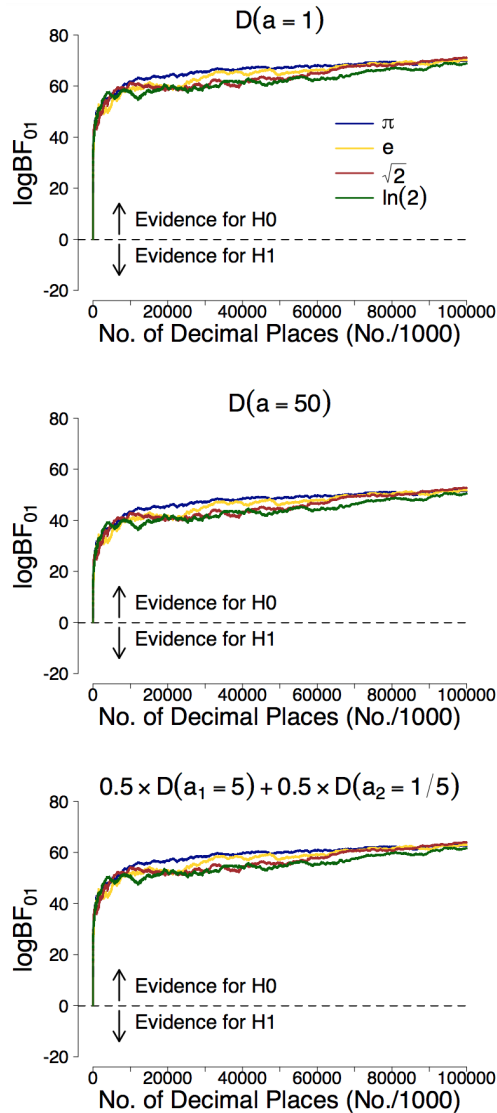


Figure 10.6: Sequential Bayes factors in favor of equal occurrence probabilities based on the first 100 million digits of  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$ . The results in the upper panel correspond to the use of an uninformative  $D(a = 1)$  prior for the alternative hypothesis; the results in the middle panel correspond to the use of an informative  $D(a = 50)$  prior; the results in the lower panel correspond to the use of a two component mixture prior of a  $D(a_1 = 5)$  and  $D(a_2 = 1/5)$  Dirichlet distribution where the mixing weight was equal to  $w = 0.5$ . Figure available at <http://tinyurl.com/hhut8dp> under CC license <https://creativecommons.org/licenses/by/2.0/>.

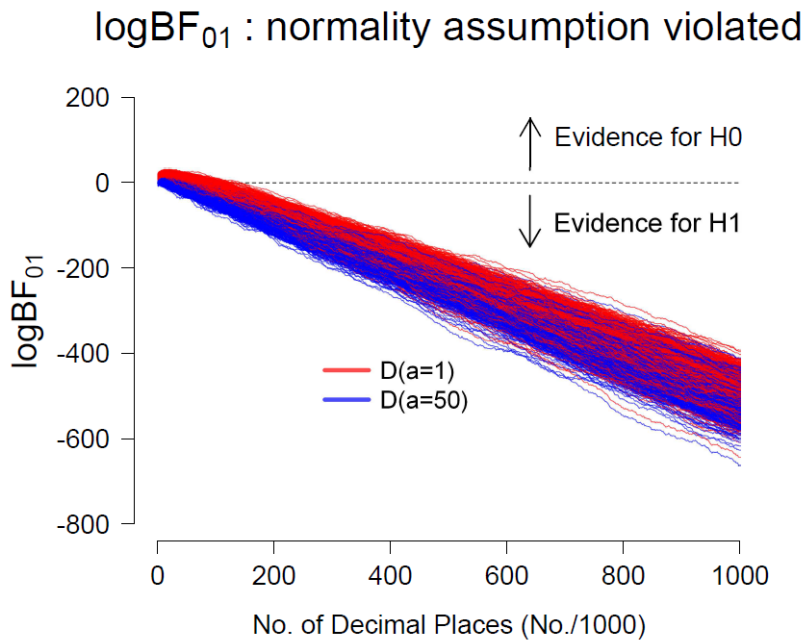


Figure 10.7: Sequential Bayes factors in favor of equal occurrence probabilities for 1,000 simulated data sets of 1 million digits each. In every data set, one digit was given an occurrence probability of .11 whereas each of the other digits occurred with probability .89/9. The evidential trajectories indicate increasingly strong evidence against the general law. Figure available at <http://tinyurl.com/j4qk2ht> under CC license <https://creativecommons.org/licenses/by/2.0/>.

## 10.A Limit of the Difference Between the Log Bayes Factors

The Savage-Dickey density ratio (Dickey & Lientz, 1970; Wetzels, Grasman, & Wagenmakers, 2010) representations of the Bayes factors are:

$$\text{BF}_{01}^{a=1} = \frac{p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 1)}{p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 1)}$$

$$\text{BF}_{01}^{a=50} = \frac{p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 50)}{p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 50)},$$

where  $\boldsymbol{\theta}_0$  is a vector of length ten with all elements being equal to  $\frac{1}{10}$ . Hence, the log Bayes factors are given by:

$$\log \text{BF}_{01}^{a=1} = \log(p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 1)) - \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 1))$$

$$\log \text{BF}_{01}^{a=50} = \log(p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 50)) - \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 50)).$$

The difference between the two log Bayes factors is:

$$\begin{aligned} \log \text{BF}_{01}^{a=1} - \log \text{BF}_{01}^{a=50} &= \log(p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 1)) - \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 1)) \\ &\quad - \log(p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 50)) + \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 50)) \\ &= \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 50)) - \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 1)) \\ &\quad + \log(p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 1)) - \log(p(\boldsymbol{\theta}_0 \mid \text{data}, \mathbf{a} = 50)). \end{aligned}$$

As soon as the data have overwhelmed the prior, the posteriors under both  $\mathcal{H}_1^{a=1}$  and  $\mathcal{H}_1^{a=50}$  will be the same, hence the last two terms cancel and the difference of the two log Bayes factors – when  $N$  grows large – is given by the difference between the log prior densities:

$$\lim_{N \rightarrow \infty} [\log \text{BF}_{01}^{a=1} - \log \text{BF}_{01}^{a=50}] = \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 50)) - \log(p(\boldsymbol{\theta}_0 \mid \mathbf{a} = 1))$$

which is equal to 18.39.

---

# Informed Bayesian $T$ -Tests

---

## Abstract

Across the empirical sciences, few statistical procedures rival the popularity of the frequentist  $t$ -test. In contrast, the Bayesian versions of the  $t$ -test have languished in obscurity. In recent years, however, the theoretical and practical advantages of the Bayesian  $t$ -test have become increasingly apparent and various Bayesian  $t$ -tests have been proposed, both objective ones (based on general desiderata) and subjective ones (based on expert knowledge). Here we propose a flexible  $t$ -prior for standardized effect size that allows computation of the Bayes factor by evaluating a single numerical integral. This specification contains previous objective and subjective  $t$ -test Bayes factors as special cases. Furthermore, we propose two measures for informed prior distributions that quantify the departure from the objective Bayes factor desiderata of predictive matching and information consistency. We illustrate the use of informed prior distributions based on an expert prior elicitation effort.

## 11.1 Introduction

The  $t$ -test is designed to assess whether or not two means differ. The question is fundamental, and consequently the  $t$ -test has grown to be an inferential workhorse of the empirical sciences. The popularity of the  $t$ -test is underscored by considering the  $p$ -values published in eight major psychology journals from 1985 until 2013 (Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2016); out of a total of 258,105  $p$ -values, 26% tested the significance of a  $t$  statistic. For comparison, 4% of those  $p$ -values tested an  $r$  statistic, 4% a  $z$  statistic, 9% a  $\chi^2$  statistic, and 57% an  $F$  statistic. Similarly, Wetzels et al. (2011) found 855  $t$ -tests reported in 252 psychology articles, for an average of about 3.4  $t$ -tests per article.

---

This chapter is published as Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian  $t$ -tests. *The American Statistician*, 74, 137–143. doi: <https://doi.org/10.1080/00031305.2018.1562983>. Also available as *arXiv preprint*: <https://arxiv.org/abs/1704.02479>

The two-sample  $t$ -test typically assumes that the data are normally distributed with common standard deviation, that is,  $Y_{1i} \sim \mathcal{N}(\mu + \frac{\sigma\delta}{2}, \sigma^2)$  and  $Y_{2j} \sim \mathcal{N}(\mu - \frac{\sigma\delta}{2}, \sigma^2)$  for  $i = 1, \dots, n_1$  and  $j = 1, \dots, n_2$ . The parameter  $\mu$  is interpreted as a grand mean,  $\sigma$  as the common standard deviation, and  $\delta$  as the (standardized) effect size. A typical application involves a treatment group and a control group and the task is to infer whether or not the treatment has an effect. The null hypothesis of the treatment not being effective corresponds to  $\mathcal{H}_0 : \delta = 0$  and implies that the population means of the two groups are the same, while the two-sided alternative  $\mathcal{H}_1$  allows the effect size to vary freely, and implies that the population means of the two groups differ.

This chapter concerns the Bayesian  $t$ -test originally developed by Jeffreys (1948) in the one-sample setting, and recently extended to the two-sample set-up by Gönen, Johnson, Lu, and Westfall (2005) and, subsequently, Rouder et al. (2009). In his work on hypothesis testing, Jeffreys focused on the *Bayes factor* (Etz & Wagenmakers, 2017; Kass & Raftery, 1995; Ly et al., 2016a, 2016b; Robert, Chopin, & Rousseau, 2009), the predictive updating factor that quantifies the change in relative beliefs about the hypotheses  $\mathcal{H}_1$  and  $\mathcal{H}_0$  based on observed data  $d$  (Wrinch & Jeffreys, 1921, p. 387):

$$\underbrace{\frac{P(\mathcal{H}_1 | d)}{P(\mathcal{H}_0 | d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d | \mathcal{H}_1)}{p(d | \mathcal{H}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}}_{\text{Prior odds}}. \quad (11.1)$$

The Bayes factor is given by the ratio of the marginal likelihoods of  $\mathcal{H}_1$  and  $\mathcal{H}_0$  that are obtained by integrating out the model parameters with respect to the parameters' prior distribution. For the two-sample  $t$ -test, the null model  $\mathcal{H}_0$  specifies two free parameters  $\zeta = (\mu, \sigma)$ , while the alternative has three, namely,  $(\zeta, \delta) = (\mu, \sigma, \delta)$ . Once the priors  $\pi_0(\zeta)$  and  $\pi_1(\zeta, \delta)$  are specified, the parameters of each model can be integrated out as follows

$$\text{BF}_{10}(d) = \frac{\int_{\Delta} \int_Z f(d | \delta, \zeta, \mathcal{H}_1) \pi_1(\delta, \zeta) d\zeta d\delta}{\int_Z f(d | \zeta, \mathcal{H}_0) \pi_0(\zeta) d\zeta}. \quad (11.2)$$

Eq. 11.2 shows that the Bayes factor can be regarded as the ratio of two weighted averages where the weights correspond to the prior distribution for the parameters. Consequently, the choice of the prior distributions is crucial for the development of a Bayes factor hypothesis test. Jeffreys (1961) elaborated on various procedures to select priors for a Bayes factor and the construction of his one-sample  $t$ -test became the norm in objective Bayesian analysis (e.g., Bayarri et al., 2012; Berger & Pericchi, 2001; Liang, Paulo, Molina, Clyde, & Berger, 2008). Jeffreys's Bayes factor for the two-sample  $t$ -test, however, was needlessly complicated and it was Gönen et al. (2005) who provided the desired simplification.

The innovation of Gönen et al. (2005) was to reparameterize the means of the two groups,  $\mu_1$  and  $\mu_2$ , in terms of a grand mean and the effect size, as was introduced at the start of this section. Following Jeffreys, the second idea was to use a right Haar prior  $\pi_0(\mu, \sigma) \propto \sigma^{-1}$  on the nuisance parameters, the parameters common to both the null and the alternative model (Bayarri et al., 2012, Berger,

Pericchi, & Varshavsky, 1998, Severini, Mukerjee, & Ghosh, 2002). Using this prior choice, the marginal likelihood of the null model – the denominator of the Bayes factor  $\text{BF}_{10}(d)$  – is proportional to the density of a standard  $t$ -distribution evaluated at the observed  $t$ -value. The third idea was to decompose the prior under the alternative hypothesis into a product of the prior used under the null hypothesis, and a test-relevant prior on the (standardized) effect size, that is,  $\pi_1(\mu, \sigma, \delta) = \pi_0(\mu, \sigma)\pi(\delta)$ . Finally, Gönen et al. (2005) showed that a normal prior  $\delta \sim \mathcal{N}(\mu_\delta, g)$  on the effect size yields a Bayes factor for the two-sample  $t$ -test that is easily calculated:

$$\text{BF}_{10}(d; \mu_\delta, g) = \frac{\frac{1}{\sqrt{1+n_\delta g}} T_\nu\left(\frac{t}{\sqrt{1+n_\delta g}}; \sqrt{\frac{n_\delta}{1+n_\delta g}} \mu_\delta\right)}{T_\nu(t)}, \quad (11.3)$$

where  $\frac{1}{b} T_\nu(\frac{t}{b}; a)$  denotes the density of a  $t$ -distribution with  $\nu$  degrees of freedom, non-centrality parameter  $a$  and scale  $b$ ,  $T_\nu(t) = T_\nu(t; 0)$  denotes the density of a standard  $t$ -distribution, and  $d$  refers to the data consisting of degrees of freedom  $\nu = n_1 + n_2 - 2$ , the observed  $t$ -value  $t = \sqrt{n_\delta}(\bar{y}_1 - \bar{y}_2)/s_p$ , where  $n_\delta = (1/n_1 + 1/n_2)^{-1}$  is the effective sample size, and  $\nu s_p^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$  the pooled sums of squares.<sup>1</sup> This means that practitioners who can calculate a classical  $t$ -test can also easily conduct a Bayesian two-sample  $t$ -test: they only need to choose the hyperparameter  $\mu_\delta$  corresponding to the effect size prior mean and the hyperparameter  $g$  corresponding to the prior variance. For brevity, we refer to the latter choice  $\delta \sim \mathcal{N}(\mu_\delta, g)$  as a  $g$ -prior on  $\delta$ , since it resembles the priors Zellner (1986) proposed in the regression framework.<sup>2</sup>

Later Bayes factors for the two-sample  $t$ -test proposed by Rouder et al. (2009) and M. Wang and Liu (2016) retained the first three ideas: the parameterization in terms of the grand mean and effect size, the use of the right Haar prior on the nuisance parameters  $\pi_0(\mu, \sigma) \propto \sigma^{-1}$ , and the decomposition  $\pi_1(\mu, \sigma, \delta) = \pi_0(\mu, \sigma)\pi(\delta)$ , but they differ in the choice of the test relevant prior  $\pi(\delta)$ . M. Wang and Liu (2016) noted that the Bayes factors of Gönen et al. (2005) are *information inconsistent*, which implies that the Bayes factor in favor of the alternative does not go to infinity when the observed  $t$ -value increases indefinitely. To make the Bayes factor information consistent, M. Wang and Liu (2016) instead proposed to assign  $g$  a Pearson type VI/beta prime hyper-prior distribution (see also Maruyama & George, 2011, for this proposal in the regression context). Inspired by the developments of Liang et al. (2008) in the regression framework, Rouder et al. (2009) proposed to replace the normal prior on  $\delta$  by a Cauchy prior  $\pi(\delta) = \text{Cauchy}(\delta; 0, \gamma)$ , a choice that resembles that of Jeffreys (1948) proposition for the one-sample  $t$ -test with prior scale  $\gamma = 1$ . In their response to M. Wang and Liu (2016), Gönen, Johnson, Lu, and Westfall (2019) stressed the relevance of a subjective prior specification and noted that the Bayes factors proposed by Rouder

<sup>1</sup>In fact, the Bayes factors for the two-sample  $t$ -test discussed here also cover the one-sample case, by (1) replacing the effective sample size by the sample size  $n$ ; (2) replacing the degrees of freedom  $\nu$  by  $n - 1$ ; and (3) replacing the two-sample  $t$ -value by its one sample equivalent  $t = \sqrt{n}\bar{y}/s_y$ , where  $\nu s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ .

<sup>2</sup>When  $\mu_\delta = 0$ , the normal  $g$ -prior on  $\delta$  translates to Zellner's  $g$ -prior on the mean difference  $(\mu_1 - \mu_2) \sim \mathcal{N}(0, g\sigma^2)$ .

et al. (2009) and M. Wang and Liu (2016) are not flexible enough to incorporate available expert knowledge, since these objective Bayes factors are based on priors that are centered at zero. Here – without taking sides in the discussion between objective and subjective inference – we present a generalized form of the Bayes factor developed by Rouder et al. (2009) that allows the prior specification to be informed by substantive domain knowledge.

The remainder of this chapter is organized as follows: Section 2 presents the proposed Bayes factor and two measures for quantifying the departure from Jeffreys’s desiderata of predictive matching and information consistency. Section 3 demonstrates, using a concrete example, how the proposed Bayes factor can be used in practice to incorporate expert knowledge based on a prior elicitation effort. The chapter ends with concluding comments.

## 11.2 Theory

We use the framework of Gönen et al. (2005) and extend the priors proposed by Rouder et al. (2009) to allow for more informed Bayesian  $t$ -tests. We exploit the fact that, with  $\pi_0(\mu, \sigma) \propto \sigma^{-1}$ , the Bayes factor can be written as<sup>3</sup>

$$\text{BF}_{10}(d) = \frac{\int T_\nu(t | \sqrt{n_\delta} \delta) \pi(\delta) d\delta}{T_\nu(t)}, \quad (11.4)$$

where  $T_\nu(t | a)$  denotes the density of a  $t$ -distribution with  $\nu$  degrees of freedom and non-centrality parameter  $a$ . The numerator can be easily evaluated using numerical integration. Consequently, Eq. 11.4 shows that researchers can easily obtain a Bayes factor based on any proper prior for the standardized effect size  $\delta$  by inserting the prior density of interest for  $\pi(\delta)$ .

We propose the use of a flexible  $t$ -prior for  $\delta$ , that is,  $\pi(\delta) = \frac{1}{\gamma} T_\kappa(\frac{\delta - \mu_\delta}{\gamma})$ , allowing practitioners to incorporate expert knowledge about standardized effect size by specifying a location hyperparameter  $\mu_\delta$ , a scale hyperparameter  $\gamma$ , and a degrees of freedom hyperparameter  $\kappa$ . The resulting Bayes factor is given by:

$$\text{BF}_{10}(d; \mu_\delta, \gamma, \kappa) = \frac{\int T_\nu(t | \sqrt{n_\delta} \delta) \frac{1}{\gamma} T_\kappa(\frac{\delta - \mu_\delta}{\gamma}) d\delta}{T_\nu(t)}, \quad (11.5)$$

where the integral in the numerator can be easily calculated using free software packages such as R (R Core Team, 2019). We believe that the proposed Bayes factor based on a  $t$ -prior for effect size has a number of advantages. First, similar to the Bayes factor proposed by Gönen et al. (2005) – which is a special case obtained by taking  $\gamma = \sqrt{g}$  and  $\kappa \rightarrow \infty$  – it allows researchers, if desired, to incorporate existing expert knowledge about effect size into the prior specification furthering cumulative scientific learning. Second, this class of priors contains the Cauchy prior of Rouder et al. (2009) as a special case (obtained by setting  $\kappa = 1$ ,  $\mu_\delta = 0$ ). Therefore, using the same expression, researchers can incorporate expert prior knowledge or they can use an objective default prior. Third, this set-up

<sup>3</sup>A derivation is provided in the online appendix (Theorem A.1, Theorem A.2, and the associated corollaries).

allows researchers to quantify the departure from Jeffreys's predictive matching and information consistency desiderata based on departure measures proposed below. This enables a more formal assessment of differences between objective and subjective prior choices and may benefit the dialog between objective and subjective Bayesians (see, e.g., M. Wang & Liu, 2016, and Gönen et al., 2019).

## 11.2.1 Two Measures for the Departure from Jeffreys's Desiderata

### 11.2.1.1 Predictive Matching

Jeffreys considered two desiderata for prior choice. The first desideratum, *predictive matching*, states that the Bayes factor should be perfectly indifferent (i.e.,  $\text{BF}_{10}(d) = 1$ ) in case the data are completely uninformative. Recall that the alternative model has three free parameters; it is therefore natural to require at least three observations before conclusions can be drawn. Consequently, Jeffreys required a Bayes factor of 1 for any data set of size smaller or equal to 2, thus, for  $\nu = 0$ . As apparent from Eq. 11.1, this requirement guarantees the posterior model odds to be the same as the prior model odds for completely uninformative data sets. For instance, the data set  $d_{\nu < \min}$  consisting of only one observation in each group  $n_1 = n_2 = 1$  automatically has zero sums of squares, that is,  $\nu s_p^2 = 0$ . If  $\bar{y}_1 \neq \bar{y}_2$  the associated  $t$ -value would then be unbounded. Let  $f(d|\delta)$  denote the *reduced* likelihood (i.e., the likelihood with the nuisance parameters integrated out):  $f(d|\delta) = \int \int f(d|\mu, \sigma, \delta) \sigma^{-1} d\mu d\sigma$ . Using a lemma distilled from the Bateman project (Bateman et al., 1953, 1954; Ly, Marsman, & Wagenmakers, 2018), straightforward but tedious computations show that  $f(d|\delta)$  is proportional to the density of a  $t$ -distribution with  $\nu$  degrees of freedom and non-centrality parameter  $\sqrt{n_\delta} \delta$  (see Theorem A.2 in the online appendix for details). To convey that nothing is learned from the data set  $d_{\nu < \min}$ , Jeffreys chose  $\pi(\delta)$  such that

$$p(d_{\nu < \min} | \mathcal{H}_0) = p(d_{\nu < \min} | \mathcal{H}_1) = \int f(d_{\nu < \min} | \delta) \pi(\delta) d\delta. \quad (11.6)$$

As  $\nu s_p^2 = 0$ ,  $n_\delta = 1/2$ , and  $\bar{y}_1 \neq \bar{y}_2$ , we obtain

$$(2|\bar{y}_1 - \bar{y}_2|)^{-1} = \int (2|\bar{y}_1 - \bar{y}_2|)^{-1} [1 + \text{sign}(\bar{y}_1 - \bar{y}_2) \text{Erf}(\frac{\delta}{2})] \pi(\delta) d\delta, \quad (11.7)$$

where  $\text{sign}(z)$  is one when  $z$  is positive, minus one when  $z$  is negative, and zero otherwise (see Corollary A.1.3 and Corollary A.2.1 in the online appendix).  $\text{Erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du$  is the error function, an odd function of  $z$ . Note that the requirement Eq. 11.7 is fulfilled if a proper symmetric prior is used for  $\delta$ . Based on Eq. 11.7 we define the (two-sided) departure of any proper prior with respect to Jeffreys's predictive matching criterion as

$$D(\pi, \text{Pred} | d_{\nu < \min}) = \int \text{sign}(\bar{y}_1 - \bar{y}_2) \text{Erf}(\frac{\delta}{2}) \pi(\delta) d\delta, \quad (11.8)$$

and note that  $\text{BF}_{10}(d_{\nu < \min}) = 1 + D(\pi, \text{Pred} | d_{\nu < \min})$ . For instance, a  $t$ -prior located at  $\mu_\delta = 0.350$ , with scale  $\gamma = 0.103$  and  $\kappa = 3$  degrees of freedom, as used

later on in the example, has a departure of the predictive matching criterion of 0.0198 when  $\bar{y}_1 > \bar{y}_2$ . In other words, for completely uninformative data sets with  $\bar{y}_1 < \bar{y}_2$  the Bayes factor will be  $\text{BF}_{10}(d_{\nu < \min}) \approx 0.98$ , while if  $\bar{y}_1 > \bar{y}_2$  the Bayes factor would be  $\text{BF}_{10}(d_{\nu < \min}) \approx 1.02$ , instead.

### 11.2.1.2 Information Consistency

The second desideratum, *information consistency*, states that the Bayes factor should provide infinite support for the alternative in case the data are overwhelmingly informative (Bayarri et al., 2012; Jeffreys, 1942). An overwhelmingly informative data set for the two-sample  $t$ -test is denoted by  $d_{\text{info},\nu}$  with  $\nu \geq 1$ , effective sample size  $n_\delta > 1/2$ ,<sup>4</sup> a (pooled) sums of squares  $\nu s_p^2 = 0$ , and an observed mean difference  $\bar{y}_1 - \bar{y}_2 \neq 0$ , thus, an unbounded  $t$ -value. For such an overwhelmingly informative data set  $d_{\text{info},\nu}$  to provide infinite support for the alternative, Jeffreys required that  $p(d_{\text{info},\nu} | \mathcal{H}_0)$  is bounded and that  $\pi(\delta)$  is chosen such that  $\int f(d_{\text{info},\nu} | \delta) \pi(\delta) d\delta$  diverges. With  $\nu s_p^2 = 0$  and  $\bar{y}_1 \neq \bar{y}_2$  the marginal likelihood of the null model becomes

$$p(d_{\text{info},\nu} | \mathcal{H}_0) = \frac{\Gamma(\frac{\nu+1}{2})}{2\pi^{\frac{\nu+1}{2}} \sqrt{\nu+2}} (n_\delta(\bar{y}_1 - \bar{y}_2)^2)^{-\frac{\nu+1}{2}}, \quad (11.9)$$

which is indeed bounded (see Corollary A.1.3 in the online appendix). In Corollary A.2.2 of the online appendix it is shown that for  $\delta$  large, the reduced likelihood  $f(d_{\text{info},\nu} | \delta)$  with  $\nu s_p^2 = 0$  behaves like a polynomial with leading order  $\nu$ , that is,

$$f(d_{\text{info},\nu} | \delta) \sim \delta^\nu. \quad (11.10)$$

To guarantee for degrees of freedom  $\nu$  that  $\int f(d_{\text{info},\nu} | \delta) \pi(\delta) d\delta$  diverges, it suffices to take a prior that does not have the  $\nu$ th moment. As information consistency should hold for all  $\nu \geq 1$ , this implies that  $\pi(\delta)$  should be chosen such that it does not have a first moment. Based on the condition that the marginal likelihood should already diverge for  $\nu = 1$ , we define the departure of Jeffreys's information consistency criterion as

$$D(\pi, \text{InfoConsist}) = \arg \min \left\{ \nu \in \mathbb{N} : \int f(d_{\text{info},\nu} | \delta) \pi(\delta) d\delta \notin \mathbb{R} \right\} - 1. \quad (11.11)$$

If  $\pi(\delta)$  is taken to be a  $t$ -prior with  $\kappa$  degrees of freedom the departure from Jeffreys's information consistency criterion is  $\kappa - 1$ , since a  $t$ -distribution has  $\kappa - 1$  moments. For instance, a  $t$ -prior with  $\kappa = 3$  degrees of freedom has only two moments and, therefore, misses the information consistency by two samples. This means that the Bayes factor only goes to infinity for overwhelmingly informative data when  $\nu \geq 3$ . Therefore, an informed  $t$ -prior with degrees of freedom larger than one requires more observations to be “convinced” by the data than does an objective prior with degrees of freedom equal to 1.

---

<sup>4</sup>This condition implies that there is at least one observation per group.

### 11.2.1.3 Practical Value of the Proposed Departure Measures

The departure measures introduced above can be used to issue recommendations for researchers who would like to incorporate expert knowledge into the prior specification, but would also like to retain Jeffreys’s desiderata as much as possible. For the proposed  $t$ -prior, we recommend that researchers who would like to retain information consistency choose  $\kappa \in (0, 1]$ . For instance, setting  $\kappa = 1$  results in a Cauchy prior. Note that, crucially, information consistency still holds if this Cauchy prior is centered on a value other than zero which enables one to incorporate expert knowledge about effect size by shifting the prior away from zero. Researchers who want to retain predictive matching should specify the prior to be centered on zero (i.e.,  $\mu_\delta = 0$ ); however, the scale parameter  $\gamma$  and the degrees of freedom  $\kappa$  can be chosen freely. Next, we demonstrate with an example how the proposed Bayes factor can be used in practice. The example features a prior elicitation effort (e.g., Kadane & Wolfson, 1998) highlighting the practical feasibility of specifying an informed prior based on expert knowledge.

## 11.3 Practice

The *facial feedback hypothesis* states that affective responses can be influenced by one’s facial expression even when that facial expression is not the result of an emotional experience. In a seminal study, Strack, Martin, and Stepper (1988) found that participants who held a pen between their teeth (inducing a facial expression similar to a smile) rated cartoons as more funny on a 10-point Likert scale ranging from 0-9 than participants who held a pen with their lips (inducing a facial expression similar to a pout).

In a recently published Registered Replication Report (Wagenmakers, Beek, et al., 2016), 17 labs worldwide attempted to replicate this finding using a preregistered and independently vetted protocol. A classical random-effects meta-analysis yielded an estimate of the mean difference between the “smile” and “pout” condition equal to 0.03 [95% CI:  $-0.11, 0.16$ ]. Furthermore, one-sided default Bayesian unpaired  $t$ -tests (using a zero-centered Cauchy prior with scale  $1/\sqrt{2}$  for effect size, the current standard in the field of psychology; see Morey & Rouder, 2015) revealed that for all 17 studies, the Bayes factor indicated evidence in favor of the null hypothesis and for 13 out of the 17 studies, the Bayes factor in favor of the null was larger than 3. Overall, the authors concluded that “the results were inconsistent with the original result” (Wagenmakers, Beek, et al., 2016, p. 924).

Here we present an informed reanalysis of the data of one of the labs based on a prior elicitation effort with Dr. Suzanne Oosterwijk, a social psychologist at the University of Amsterdam with considerable expertise in this domain. The results for the other labs can be found in online appendix C.

### 11.3.1 Prior Elicitation

Before commencing the elicitation process, we asked our expert to ignore the knowledge about the failed replication of Strack et al. (1988). Next, we stressed that the goal of the elicitation effort was to obtain an informed prior distribu-

tion for  $\delta$  under the alternative hypothesis  $\mathcal{H}_1$ , that is, under the assumption that the effect is present. This was important in order to prevent unwittingly eliciting a prior that is a mixture between a point mass at zero and the distribution of interest. Then, we proceeded in steps of increasing sophistication. First, together with the expert we decided that the theory specified a direction, implying a one-sided hypothesis test. Next, we asked the expert to provide a value for the median of the effect size: this yielded a value of 0.35. Subsequently, we asked for values for the 33% and 66% percentile of the prior distribution for the effect size: this yielded values of 33%-tile = 0.25 and 66%-tile = 0.45. To finesse and validate the specified prior distribution we used the MATCH Uncertainty Elicitation Tool (<http://optics.eee.nottingham.ac.uk/match/uncertainty.php>; see also online appendix B), a web application that allows one to elicit probability distributions from experts (Morris, Oakley, & Crowe, 2014). Furthermore, we used R's (R Core Team, 2019) plotting capabilities for eliciting the prior number of degrees of freedom. The complete elicitation effort took approximately one hour and resulted in a  $t$ -distribution with location 0.350, scale 0.102, and 3 degrees of freedom. As shown in the theory part, this prior choice has a departure from the predictive matching criterion of  $\pm 0.0198$  and misses information consistency by two samples. It should be emphasized, however, that the goal of this prior elicitation was to construct a prior that truly reflects the expert's knowledge without being constrained by considerations about Bayes factor desiderata. Alternatively, in an elicitation effort that puts more emphasis on these desiderata, one could, for instance, fix the degrees of freedom to one and let the expert only choose the location and scale.

### 11.3.2 Reanalysis of the Oosterwijk Replication Study

Having elicited an informed prior distribution for  $\delta$  under the alternative hypothesis, we now turn to a detailed reanalysis of the facial feedback replication attempt from Dr. Oosterwijk's lab at the University of Amsterdam. This data set features 53 participants in the "smile" condition with an average funniness rating of 4.63 ( $SD = 1.48$ ), and 57 participants in the "pout" condition with an average funniness rating of 4.87 ( $SD = 1.32$ ); consequently, the observed  $t$  statistic is  $t(108) = -0.90$ .

The alternative hypothesis is directional, that is, the teeth condition is predicted to result in relatively high funniness ratings, not relatively low funniness ratings. In order to respect the directional nature of the alternative hypothesis the two-sided informed  $t$ -test outlined above requires an adjustment. Specifically, the Bayes factor that compares an alternative hypothesis that only allows for positive effect size values to the null hypothesis can be computed via a simply identity that exploits the transitive nature of the Bayes factor (Morey & Wagenmakers, 2014):

$$\text{BF}_{+0}(d) = \underbrace{\frac{p(d|\mathcal{H}_+)}{p(d|\mathcal{H}_1)}}_{\text{BF}_{+1}(d)} \underbrace{\frac{p(d|\mathcal{H}_1)}{p(d|\mathcal{H}_0)}}_{\text{BF}_{10}(d)} = \text{BF}_{+1}(d)\text{BF}_{10}(d). \quad (11.12)$$

We already showed how to obtain  $\text{BF}_{10}(d)$ , that is, the Bayes factor for the two-sided test of an informed alternative hypothesis; the correction term  $\text{BF}_{+1}(d)$  can

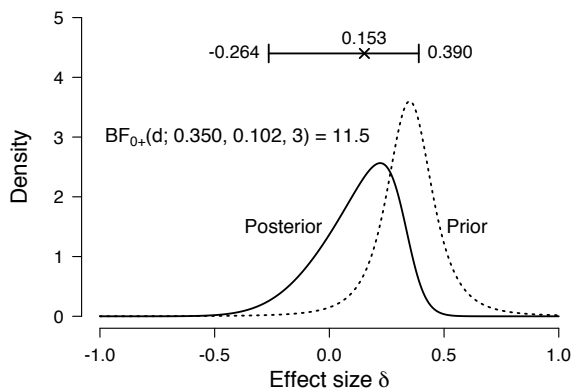


Figure 11.1: Results of an informed reanalysis of the facial feedback hypothesis replication data from the Oosterwijk lab. The dotted line corresponds to the elicited  $\frac{1}{0.102}T_3\left(\frac{\delta-0.350}{0.102}\right)$  prior distribution. The solid line corresponds to the associated posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favor of the null hypothesis over the one-sided informed alternative hypothesis equals  $BF_{0+}(d; 0.350, 0.102, 3) = 11.5$ . Figure available at <https://tinyurl.com/mk7uaxm> under CC license <https://creativecommons.org/licenses/by/2.0/>.

be obtained by simply dividing the posterior mass for  $\delta$  larger than zero by the prior mass for  $\delta$  larger than zero.<sup>5</sup> The Bayes factor hypothesis test that we report will respect the directional nature of the facial feedback hypothesis and include the correction term from Eq. 11.12.

Fig. 11.1 shows the results of the reanalysis of the data from the Oosterwijk lab. The displayed prior and posterior distribution do not impose the directional constraint. The one-sided Bayes factor based on the informed prior equals  $BF_{0+}(d; 0.350, 0.102, 3) = 11.5$ , indicating that the data are about twelve times more likely under the null hypothesis than under the one-sided alternative hypothesis.

For comparison, Fig. 11.2 displays the results based on the default one-sided zero-centered Cauchy distribution with scale  $1/\sqrt{2}$ . The one-sided default Bayes factor equals  $BF_{0+}(d; 0, 1/\sqrt{2}, 1) = 8.7$ , indicating that the data are about 9 times more likely under the null hypothesis than under the one-sided default alternative hypothesis. Hence, both the informed and the default Bayes factor yield the same qualitative conclusion, that is, evidence for the null hypothesis. However, the unrestricted posterior distributions differ noticeably between the informed and the

<sup>5</sup>The expression for the marginal posterior distribution for  $\delta$  is provided in Corollary A.2.3 in the online appendix. Using this expression, numerical integration can be used to obtain the desired posterior mass.

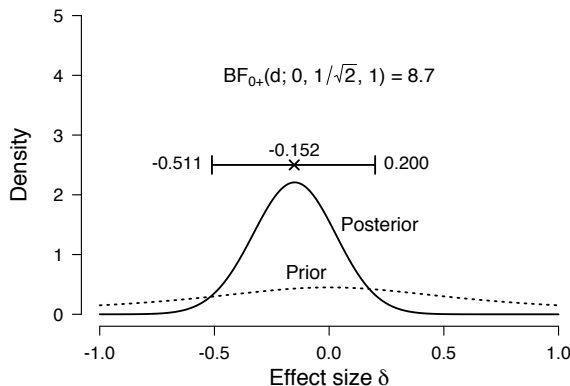


Figure 11.2: Results of the default analysis of the facial feedback hypothesis replication data from the Oosterwijk lab. The dotted line corresponds to the default Cauchy prior distribution with scale parameter  $1/\sqrt{2}$ . The solid line corresponds to the associated posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favor of the null hypothesis over the one-sided default alternative hypothesis equals  $\text{BF}_{0+}(d; 0, 1/\sqrt{2}, 1) = 8.7$ . Figure available at <https://tinyurl.com/mgs28ob> under CC license <https://creativecommons.org/licenses/by/2.0/>.

default analysis: the posterior median based on the informed prior specification is positive and equal to 0.153 (95% credible interval:  $[-0.264, 0.390]$ ) whereas the posterior median based on the default prior distribution is equal to  $-0.152$  (95% credible interval:  $[-0.511, 0.200]$ ).

## 11.4 Concluding Comments

The comparison between two means is a quintessential inference problem. Originally developed by Jeffreys (1948) in the one-sample setting, the Bayesian  $t$ -test has recently been extended to the two-sample set-up by Gönen et al. (2005) and, subsequently, by Rouder et al. (2009) and M. Wang and Liu (2016). Here we showed that practitioners can easily and intuitively use a generalized version of the Bayes factor by Rouder et al. (2009) to inform their two-sample Bayesian  $t$ -tests. We used the framework of Gönen et al. (2005) and extended the priors by Rouder et al. (2009) to allow for more informed Bayesian  $t$ -tests that can incorporate expert knowledge by using a flexible  $t$ -prior. An advantage of the flexible  $t$ -prior is that it contains the objective default prior by Rouder et al. (2009) as a special case and the subjective prior proposed by Gönen et al. (2005) as a limiting case. Therefore, practitioners can use the same formula to compute subjective and objective Bayesian  $t$ -tests. To encourage its adoption in applied work, we have

implemented the proposed Bayesian  $t$ -test set-up in the open-source statistical program JASP (JASP Team, 2020, [jasp-stats.org](https://jasp-stats.org)). In the theoretical part of this chapter, we investigated theoretical properties of the informed  $t$ -prior. Specifically, we discussed popular Bayes factor desiderata and proposed measures to quantify the deviation of an informed  $t$ -test from its objective counterpart. In the practical part of the chapter, we illustrated the use of the informed Bayes factor with an example. Similar to the prior proposed by Gönen et al. (2005), the flexible  $t$ -prior may encourage the use of prior distributions that better represent the predictions from the hypothesis under test, allowing more meaningful conclusions to be drawn from the same data (Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016; Rouder, Morey, & Wagenmakers, 2016).

Other choices than a  $t$ -prior for effect size are conceivable. Eq. 11.3 shows that one can obtain a Bayes factor for any scale-mixture of normals by integrating Eq. 11.3 with respect to a prior on  $g$  (see Theorem A.3 in the online appendix; for possible choices see, e.g., Bayarri et al., 2012). This also includes the prior proposed by M. Wang and Liu (2016) and highlights that it is straightforward to extend this prior to include a location parameter that can be specified based on expert knowledge. In fact, the expressions for the Bayes factor that we presented make it relatively straightforward to use *any* proper prior on standardized effect size (see Eq. 11.4). The proposed departure measures can then be used to investigate information consistency and predictive matching for different choices.

In this chapter, we focused on the Bayes factor as the inferential tool for quantifying the relative evidence for competing hypotheses based on observed data. However, it could be argued that a complete Bayesian analysis requires one to also specify the prior plausibilities of the competing hypotheses. This is of particular importance in situations where unlikely hypotheses are tested or when multiple comparisons are considered (Scott & Berger, 2010). Although specifying the prior plausibilities of the competing hypotheses may not be trivial, once this has been achieved, the Bayes factor can be simply multiplied by the prior odds to obtain the posterior odds of interest.

R code and the online appendix are available at: <https://osf.io/37vch/>.



---

# Informed Bayesian Inference for the A/B Test

---

## Abstract

Booming in business and a staple analysis in medical trials, the A/B test assesses the effect of an intervention or treatment by comparing its success rate with that of a control condition. Across many practical applications, it is desirable that (1) evidence can be obtained in favor of the null hypothesis that the treatment is ineffective; (2) evidence can be monitored as the data accumulate; (3) expert prior knowledge can be taken into account. Most existing approaches do not fulfill these desiderata. In this chapter we describe a Bayesian A/B procedure based on Kass and Vaidyanathan (1992) that allows one to monitor the evidence for the hypotheses that the treatment has either a positive effect, a negative effect, or, crucially, no effect. Furthermore, this approach enables one to incorporate expert knowledge about the relative prior plausibility of the rival hypotheses and about the expected size of the effect, given that it is non-zero. To facilitate the wider adoption of this Bayesian procedure we developed the **abtest** package in R. We illustrate the package options and the associated statistical results with a synthetic example.

## 12.1 Introduction

Does the modification of a company website increase the number of online purchases? Does a new drug result in a lower mortality rate? These are just two examples of the kinds of questions that can be addressed with A/B testing, a procedure popular not only in business and medical clinical trials, but also in

---

This chapter has been submitted for publication as Gronau, Q. F., Raj K. N., A., & Wagenmakers, E.-J. (2019). Informed Bayesian inference for the A/B test. Available as *arXiv preprint*: <https://arxiv.org/abs/1905.02068>

fields such as psychology, neuroscience, and biology. An A/B test compares the success rate of two options or treatment arms, A and B, and therefore can be conceptualized as a test for a difference between two proportions (Little, 1989).<sup>1</sup> Typically, options A and B correspond to a control condition and an intervention or treatment of interest.

Regardless of the specific field of application, we believe three general desiderata for A/B tests can be identified. First, it is desirable that evidence can be obtained in favor of the null hypothesis that there is no difference between options A and B. For instance, suppose a programmer alters code that should leave the appearance of a website unaffected. An A/B test may be conducted to confirm that the code changes did not lead to unintended consequences. Alternatively, suppose that a cheaper drug is introduced as a replacement of the standard drug; here, an A/B test may confirm that the cheaper drug is as effective as the drug that is currently standard.

Second, it is desirable that evidence can be monitored as the data accumulate. Data collection can be time-consuming and expensive, and interim tests allow one to assess whether the results in hand are already sufficiently compelling or whether additional data ought to be obtained. There is also an ethical aspect to this desideratum, one that is particularly pronounced in case of new clinical treatments that are potentially beneficial or harmful; it is unethical to withhold treatment that interim analysis shows to be beneficial, just as it is unethical to continue to administer a treatment that interim analysis shows to be harmful (e.g., Armitage, 1960; see also Ware, 1989 and the accompanying discussion).

Third, it is desirable that expert knowledge can be taken into account (e.g., O'Hagan, 2019). In many A/B testing applications, there exists considerable expert knowledge about what size of effect to expect. For instance, the effect of website changes on conversion rates is often less than 0.5% (Berman, Pekelis, Scott, & Van den Bulte, 2018). Incorporating such expert knowledge into the statistical analysis will yield a more targeted test.

Unfortunately, the majority of A/B testing procedures that are currently in vogue do not fulfill the above desiderata. Specifically, many companies apply standard  $p$ -value-based null hypothesis significance testing to assess whether or not options A and B differ. This procedure has the advantage that it is readily available in software such as R (R Core Team, 2019, e.g., via the functions `prop.test`, `fisher.test`, and `chisq.test`). However, this approach cannot distinguish between *absence of evidence* (i.e., the data are inconclusive) and *evidence of absence* (i.e., the data provide support for the null hypothesis that options A and B do not differ; e.g., Dienes, 2014). Furthermore, although common practice, sequentially monitoring the uncorrected  $p$ -value (and stopping data collection as soon as the  $p$ -value is smaller than some fixed  $\alpha$ -level) invalidates the analysis (e.g., Feller, 1940). However, there exist valid classical sequential procedures that enable one to monitor a corrected  $p$ -value as data accumulate (e.g., Malek, Katariya, Chow, & Ghavamzadeh, 2017). For instance, *Optimizely*, one of the leading commercial A/B testing platforms, has recently implemented an alter-

---

<sup>1</sup>The A/B test set-up discussed in this chapter assumes that the dependent variable is binary. Nevertheless, the dependent variable could in principle also be continuous.

native  $p$ -value-based approach that allows users to continuously monitor the test outcome (Johari, Koomen, Pekelis, & Walsh, 2017). Nevertheless, these sequential  $p$ -value-based procedures retain the inability to quantify evidence for the absence of an effect. Furthermore, (sequential)  $p$ -value-based A/B testing does not allow one to incorporate expert knowledge into the statistical analysis in a straightforward manner.

An alternative A/B testing approach that has become more popular of late is Bayesian estimation. For instance, *VWO*, another leading A/B testing platform, has recently implemented a Bayesian estimation approach (Stucchio, 2015). A Bayesian estimation approach is also available via the **bayesAB** package (Portman, 2019).<sup>2</sup> Since Bayesian inference is immune to optional stopping (Berger & Wolpert, 1988), this approach allows one to monitor the analysis output as data accumulate. A Bayesian estimation approach also enables the incorporation of expert knowledge via the specification of a prior distribution that captures the expert’s knowledge about a parameter of interest. However, this approach operates under the assumption that an effect exists – since a continuous prior assigns zero probability to a single null value – and consequently does not allow one to obtain evidence in favor of the null hypothesis of no effect. For instance, **bayesAB** provides the user with the posterior probability that one option yields more successes than the other, but this ignores the fact that both options could be equally effective. Furthermore, the currently used Bayesian estimation approaches – such as the one implemented in **bayesAB** – typically assign independent priors to the success probabilities of the control and treatment condition, a practice that was critiqued by Howard (1998).<sup>3</sup>

To overcome the limitations of the current A/B tests we developed the **abtest** package in R (R Core Team, 2019). The **abtest** package implements Bayesian inference for the A/B test, using informed prior distributions that induce a dependency between the two success probabilities. The analysis approach is based on a model by Kass and Vaidyanathan (1992); for alternative approaches see Deng, Lu, and Chen (2016), Jamil, Marsman, Ly, Morey, and Wagenmakers (2017), Pham-Gia, Van Thin, and Doan (2017), and Skorski (2019). The implemented Bayesian procedure allows users (1) to obtain evidence in favor of the null hypothesis (e.g., Berger & Delampady, 1987; Wagenmakers, Marsman, et al., 2018); (2) monitor the evidence as the data accumulate (e.g., Rouder, 2014); and (3) elicit and incorporate expert prior knowledge (e.g., O’Hagan, 2019). The **abtest** package thus fulfills all three desiderata mentioned above.

The **abtest** package provides functionality for both hypothesis testing and parameter estimation. In line with Jeffreys (1939) and Fisher (1928), we believe that testing and estimation are complementary activities (Haaf et al., 2019): before a

<sup>2</sup>The **bayesAB** package provides a range of functions for Bayesian A/B testing. One advantage is that users can choose from a range of different data distributions (e.g., Bernoulli, normal, Poisson, etc.).

<sup>3</sup>“do English or Scots cattle have a higher proportion of cows infected with a certain virus? Suppose we were informed (before collecting any data) that the proportion of English cows infected was 0.8. With independent uniform priors we would now give  $H_1$  ( $p_1 > p_2$ ) a probability of 0.8 (because the chance that  $p_2 > 0.8$  is still 0.2). In very many cases this would not be appropriate. Often we will believe (for example) that if  $p_1$  is 80%,  $p_2$  will be near 80% as well and will be almost equally likely to be larger or smaller.” (p. 363)

parameter is estimated, it should be tested whether there is anything to justify estimation at all. Jeffreys (1939, p. 345) related this principle to Occam’s razor: “variation must be taken as random until there is positive evidence to the contrary” (see also Kass & Raftery, 1995, Section 8.1). However, some researchers and practitioners oppose this idea, for instance because they believe that one should replace hypothesis testing with parameter estimation (e.g., Gelman & Rubin, 1995; Cumming, 2014). Nevertheless, the **abtest** package may also be useful for researchers without an interest in hypothesis testing, since the package can also be used exclusively for Bayesian parameter estimation (and prior elicitation).

This chapter is organized as follows: The next section discusses the implementation details of the Bayesian A/B test procedure used in **abtest**. Subsequently, the functionality of the **abtest** package and the practical benefits of the implemented approach are demonstrated using a synthetic example. The chapter ends with concluding comments.

## 12.2 Implementation Details

The Bayesian A/B test implemented in the **abtest** package is based on Kass and Vaidyanathan (1992, Section 3, “Testing Equality of Two Binomial Proportions”). Appendix A-C provide detailed derivations.

### 12.2.1 Model

Let  $y_1$  denote the number of successes for option A with  $n_1$  denoting the corresponding total number of observations for option A. Similarly,  $y_2$  denotes the number of successes for option B with  $n_2$  denoting the corresponding total number of observations for option B. The Bayesian A/B test model based on Kass and Vaidyanathan (1992) is specified as follows:<sup>4</sup>

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= \beta - \frac{\psi}{2} \\ \log\left(\frac{p_2}{1-p_2}\right) &= \beta + \frac{\psi}{2} \\ y_1 &\sim \text{Binomial}(n_1, p_1) \\ y_2 &\sim \text{Binomial}(n_2, p_2).\end{aligned}\tag{12.1}$$

Therefore, the model assumes that  $y_1$  and  $y_2$  follow binomial distributions with success probabilities  $p_1$  and  $p_2$ . These probabilities are functions of the two model parameters,  $\beta$  and  $\psi$ . Specifically, the log odds corresponding to  $p_1$  are given by  $\beta - \psi/2$  and the log odds corresponding to  $p_2$  are given by  $\beta + \psi/2$ . The nuisance parameter  $\beta$  corresponds to the grand mean of the log odds and the test-relevant parameter  $\psi$  corresponds to the log odds ratio. When  $\psi$  is positive, this implies that  $p_2 > p_1$  (i.e., option B has a higher success probability than option A); when  $\psi$  is negative this implies that  $p_2 < p_1$  (i.e., option B has a lower success probability than option A).

---

<sup>4</sup>Note that this is equivalent to a logistic regression model with a binary covariate (i.e., group membership) that is coded using  $\pm 0.5$ .

### 12.2.2 Hypotheses

The **abtest** package enables both estimation of the model parameters and testing of hypotheses about the test-relevant log odds ratio parameter  $\psi$ . There are four hypotheses that are of potential interest:

1. The null hypothesis  $\mathcal{H}_0$  which states that the success probabilities  $p_1$  and  $p_2$  are identical, that is,  $p_1 = p_2$ . This is equivalent to  $\mathcal{H}_0 : \psi = 0$ . This hypothesis corresponds to the claim that there is no difference between options A and B (i.e., the “A/A test”).
2. The two-sided alternative hypothesis  $\mathcal{H}_1$  which states that the two success probabilities  $p_1$  and  $p_2$  are not equal (i.e.,  $p_1 \neq p_2$ ), but does not specify which of the two is larger. This is equivalent to  $\mathcal{H}_1 : \psi \neq 0$ . This hypothesis corresponds to the claim that options A and B differ but it is not specified which one yields more successes.
3. The one-sided hypothesis  $\mathcal{H}_+$  which states that the second success probability  $p_2$  is larger than the first success probability  $p_1$ . This is equivalent to  $\mathcal{H}_+ : \psi > 0$ . This hypothesis corresponds to the claim that option B yields more successes than option A.
4. The one-sided hypothesis  $\mathcal{H}_-$  which states that the first success probability  $p_1$  is larger than the second success probability  $p_2$ . This is equivalent to  $\mathcal{H}_- : \psi < 0$ . This hypothesis corresponds to the claim that option A yields more successes than option B.

Researchers who conduct an A/B test are usually interested in answering the question: Does option B yield more successes than option A (i.e.,  $\mathcal{H}_+$ ), fewer successes than option A (i.e.,  $\mathcal{H}_-$ ), or is there no difference between options A and B (i.e.,  $\mathcal{H}_0$ )? Therefore, it may be argued that the hypotheses of interest are typically  $\mathcal{H}_+$ ,  $\mathcal{H}_-$ , and  $\mathcal{H}_0$ . Consequently, by default, only these three hypotheses are assigned non-zero prior probability in the **abtest** package. Specifically, a default prior probability of .50 is assigned to the hypothesis that there is no effect (i.e.,  $\mathcal{H}_0$ ), and the remaining prior probability is split evenly across the hypothesis that there is a positive effect (i.e.,  $\mathcal{H}_+$  receives .25) and a negative effect (i.e.,  $\mathcal{H}_-$  also receives .25). The user may change these default prior probabilities to custom values. Table 12.1 provides an overview of five qualitatively different tests that can be conducted by assigning prior probabilities to hypotheses in certain ways.<sup>5</sup> The first column displays the default setting that assigns probability .50 to the null hypothesis and splits the remaining probability evenly across  $\mathcal{H}_+$  and  $\mathcal{H}_-$ . The second column displays a prior probability assignment that implements an undirected test (i.e.,  $\mathcal{H}_0$  is compared to the undirected  $\mathcal{H}_1$ ). The third column displays a prior probability assignment for testing whether the effect is non-existent

<sup>5</sup>Note that, except for the first column of Table 12.1 which displays the default setting, the remaining examples use equal prior probabilities for all hypotheses that are assigned non-zero prior probability. However, the user can of course also assign prior probability unevenly to the hypotheses of interest (e.g., if prior knowledge exists about the relative plausibility of the rival hypotheses).

Table 12.1: Changing the prior probability assignments across rival hypotheses produces different tests.

Hypothesis	Test				
	Default	Undirected	Positive	Negative	Direction
$\mathcal{H}_0$	<b>.50</b>	<b>.50</b>	<b>.50</b>	<b>.50</b>	0
$\mathcal{H}_1$	0	<b>.50</b>	0	0	0
$\mathcal{H}_+$	<b>.25</b>	0	<b>.50</b>	0	<b>.50</b>
$\mathcal{H}_-$	<b>.25</b>	0	0	<b>.50</b>	<b>.50</b>

or positive. The fourth column displays a prior probability assignment for testing whether the effect is non-existent or negative. Finally, the fifth column displays a prior probability assignment for a test of direction, that is, for testing whether the effect is positive or negative. This last setting may be of interest whenever the null hypothesis is a priori deemed implausible, uninteresting, or irrelevant.

### 12.2.3 Parameter Priors

The **abtest** package assigns normal priors to the model parameters:  $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$  and  $\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)$ . As illustrated in the example below, these priors result in a dependency in the implied prior for the success probabilities  $p_1$  and  $p_2$ , which is generally desirable (Howard, 1998).

For the one-sided hypotheses  $\mathcal{H}_+$  and  $\mathcal{H}_-$ , the prior on  $\psi$  is truncated at zero. Specifically, for  $\mathcal{H}_+$ , the prior on  $\psi$  is a truncated normal distribution with parameters  $\mu_\psi$  and  $\sigma_\psi$  and lower bound at zero. For  $\mathcal{H}_-$ , the prior on  $\psi$  is a truncated normal distribution with parameters  $\mu_\psi$  and  $\sigma_\psi$  and upper bound at zero. These normal priors are computationally convenient and sufficiently flexible to encode a wide range of prior information.

By default, the **abtest** package assigns standard normal priors to both  $\beta$  and  $\psi$ . For the nuisance parameter  $\beta$ , a standard normal prior results in a relatively flat implied prior on  $p_1$  and  $p_2$  when  $\psi = 0$ . Generally, the choice of a prior for the nuisance parameter  $\beta$  is relatively inconsequential (Kass & Vaidyanathan, 1992). In contrast, the prior on the test-relevant parameter  $\psi$  is consequential, as it defines the extent to which the hypotheses of interest differ from  $\mathcal{H}_0$ . Our choice for a default standard normal prior on the test-relevant parameter  $\psi$  is motivated by the fact that a zero-centered prior does not favor any of the two options A or B a priori. Furthermore, the standard deviation of 1 results in a prior distribution that assigns mass to a wide range of reasonable log odds ratios (H. Chen, Cohen, & Chen, 2010) without being so uninformative that the results unduly favor  $\mathcal{H}_0$  (Bartlett, 1957; Lindley, 1957).<sup>6</sup> However, large changes in the prior standard deviation of the test-relevant parameter may result in large changes in the results, as the prior standard deviation governs the degree to which the hypothesis of

<sup>6</sup>Note that the default implied prior on the absolute risk  $p_2 - p_1$  is considerably more narrow than the prior induced by the popular default choice that assigns  $p_1$  and  $p_2$  independent uniform distributions (Jeffreys, 1935).

interest makes predictions that differ from  $\mathcal{H}_0$ . To include prior knowledge about the expected results, the **abtest** package allows the user to change the default values of the prior distributions for the nuisance parameter  $\beta$  and the test-relevant parameter  $\psi$ , either by changing the location of the normal prior distribution, the scale, or both.

### 12.2.4 Encoding Prior Information

A straightforward way to encode prior information about the model parameters is to set  $\mu_\beta$ ,  $\sigma_\beta$ ,  $\mu_\psi$ , and  $\sigma_\psi$  directly. However, it may sometimes be easier to specify prior distributions based on quantities such as the (log) odds ratio, relative risk (i.e.,  $p_2/p_1$ , the ratio of the success probability in condition B and condition A), and absolute risk (i.e.,  $p_2 - p_1$ , the difference of the success probability in condition B and condition A). The `elicit_prior` function allows users to encode prior information about a quantity of interest (either log odds ratio, odds ratio, relative risk, or absolute risk). The function assumes that the prior on  $\beta$  is not the primary target of prior elicitation and is fixed by the user a priori (using the arguments `mu_beta` and `sigma_beta`) – for instance, to a standard normal prior which corresponds to a relatively flat implied prior on  $p_1$  and  $p_2$  when  $\psi = 0$ .

To encode prior information, the user needs to provide quantiles for a quantity of interest. Let  $q_i, i = 1, \dots, I$  denote the values of  $I$  quantiles provided by the user and let  $\text{prob}_i, i = 1, \dots, I$  denote the corresponding probabilities (e.g., for the median,  $\text{prob}_i = 0.5$ ). Least-squares minimization is used to obtain  $\mu_\psi$  and  $\sigma_\psi$  as follows:

$$(\mu_\psi, \sigma_\psi) = \arg \min_{\mu_\psi, \sigma_\psi} \sum_{i=1}^I (F(q_i; \mu_\psi, \sigma_\psi) - \text{prob}_i)^2, \quad (12.2)$$

where  $F(\cdot; \mu_\psi, \sigma_\psi)$  corresponds to the cumulative distribution function (cdf) for the quantity of interest implied by the normal prior on  $\psi$ . For some quantities, this cdf also depends on the prior for  $\beta$ ; however, as described above, it is assumed that  $\mu_\beta$  and  $\sigma_\beta$  are fixed a priori.

### 12.2.5 Hypothesis Testing

To quantify the evidence that the data provide for  $\mathcal{H}_0$ ,  $\mathcal{H}_1$ ,  $\mathcal{H}_+$ , and  $\mathcal{H}_-$ , one can compute Bayes factors (Jeffreys, 1939; Kass & Raftery, 1995) and posterior probabilities of the rival hypotheses. The posterior probability of hypothesis  $\mathcal{H}_j$ ,  $j \in \{0, 1, +, -\}$  is given by:

$$\underbrace{p(\mathcal{H}_j | \text{data})}_{\text{posterior probability}} = \underbrace{\frac{p(\text{data} | \mathcal{H}_j)}{\sum_k p(\text{data} | \mathcal{H}_k) p(\mathcal{H}_k)}}_{\text{updating factor}} \times \underbrace{p(\mathcal{H}_j)}_{\text{prior probability}}. \quad (12.3)$$

The Bayes factor for comparing hypotheses  $\mathcal{H}_j$  and  $\mathcal{H}_k$  equals the change from prior to posterior odds:

$$\underbrace{\frac{p(\mathcal{H}_j | \text{data})}{p(\mathcal{H}_k | \text{data})}}_{\text{posterior odds}} = \underbrace{\frac{p(\text{data} | \mathcal{H}_j)}{p(\text{data} | \mathcal{H}_k)}}_{\text{Bayes factor BF}_{jk}} \times \underbrace{\frac{p(\mathcal{H}_j)}{p(\mathcal{H}_k)}}_{\text{prior odds}}. \quad (12.4)$$

In order to obtain posterior probabilities of the hypotheses and Bayes factors one needs to evaluate the marginal likelihood  $p(\text{data} \mid \mathcal{H}_j)$  for each hypothesis  $j \in \{0, 1, +, -\}$ . For  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , we evaluate the marginal likelihood using Laplace approximations as suggested by Kass and Vaidyanathan (1992). Specifically, the marginal likelihood for  $\mathcal{H}_0$  is approximated by:

$$\begin{aligned} p(\text{data} \mid \mathcal{H}_0) &= \int \underbrace{p(\text{data} \mid \beta)}_{\text{likelihood}} \underbrace{\pi_0(\beta)}_{\text{prior}} d\beta \\ &\approx (2\pi\sigma_0^2)^{\frac{1}{2}} \exp \{l_0^*(\beta_0^*)\}, \end{aligned} \quad (12.5)$$

where  $l_0^*(\beta) = \log \{p(\text{data} \mid \beta) \pi_0(\beta)\}$ ,  $\beta_0^*$  corresponds to the mode of  $l_0^*(\beta)$ , and  $\sigma_0^2 = \left( -\frac{d^2}{d\beta^2} l_0^*(\beta) \right)^{-1} \Big|_{\beta=\beta_0^*}$  denotes the inverse of the negative second derivative of  $l_0^*(\beta)$  evaluated at the mode  $\beta_0^*$ .

The marginal likelihood for  $\mathcal{H}_1$  is approximated by:

$$\begin{aligned} p(\text{data} \mid \mathcal{H}_1) &= \int \int \underbrace{p(\text{data} \mid \beta, \psi)}_{\text{likelihood}} \underbrace{\pi(\beta, \psi)}_{\text{prior}} d\beta d\psi \\ &\approx 2\pi \det(\mathbf{\Sigma}_1)^{\frac{1}{2}} \exp \{l^*(\beta^*, \psi^*)\}, \end{aligned} \quad (12.6)$$

where  $l^*(\beta, \psi) = \log \{p(\text{data} \mid \beta, \psi) \pi(\beta, \psi)\}$ ,  $(\beta^*, \psi^*)$  denotes the mode of  $l^*(\beta, \psi)$ , and  $\mathbf{\Sigma}_1 = (-\mathbf{H}_1)^{-1} \Big|_{(\beta, \psi)=(\beta^*, \psi^*)}$  denotes the inverse of the negative Hessian  $\mathbf{H}_1$  (i.e., the matrix with second-order partial derivatives) of  $l^*(\beta, \psi)$  evaluated at the mode  $(\beta^*, \psi^*)$ .

These Laplace approximations work well in practice, even for sample sizes that are extremely small. As a demonstration, for a range of synthetic data sets we computed the (log of the) Bayes factor  $\text{BF}_{10}$  which compares  $\mathcal{H}_1$  to  $\mathcal{H}_0$  using the above Laplace approximations and, as a comparison, also using bridge sampling (Gronau, Singmann, & Wagenmakers, 2020; Meng & Wong, 1996). The priors on  $\beta$  and  $\psi$  were standard normal distributions. Figure 12.1 displays the results and confirms that the Laplace approximation yields accurate results, even for sample sizes as small as  $n_1 = n_2 = 5$ .

For the one-sided hypotheses  $\mathcal{H}_+$  and  $\mathcal{H}_-$ , Laplace approximations did not appear to yield accurate results for small sample sizes, even after removing the constraint on  $\psi$  through the parameterization  $(\beta, \xi) = (\beta, \log(\psi))$  for  $\mathcal{H}_+$  and  $(\beta, \xi) = (\beta, \log(-\psi))$  for  $\mathcal{H}_-$ . The **abtest** package therefore uses importance sampling to increase the accuracy of the Laplace approximations when computing the marginal likelihoods for  $\mathcal{H}_+$  and  $\mathcal{H}_-$ . Specifically, a Laplace approximation is used to approximate the mode and covariance matrix of the posterior. The importance density is then given by a multivariate  $t$  distribution with location set to the approximated posterior mode, scale matrix set to the approximated posterior covariance matrix, and five degrees of freedom (note that the user can change the degrees of freedom). The marginal likelihood for  $\mathcal{H}_+$  is then estimated

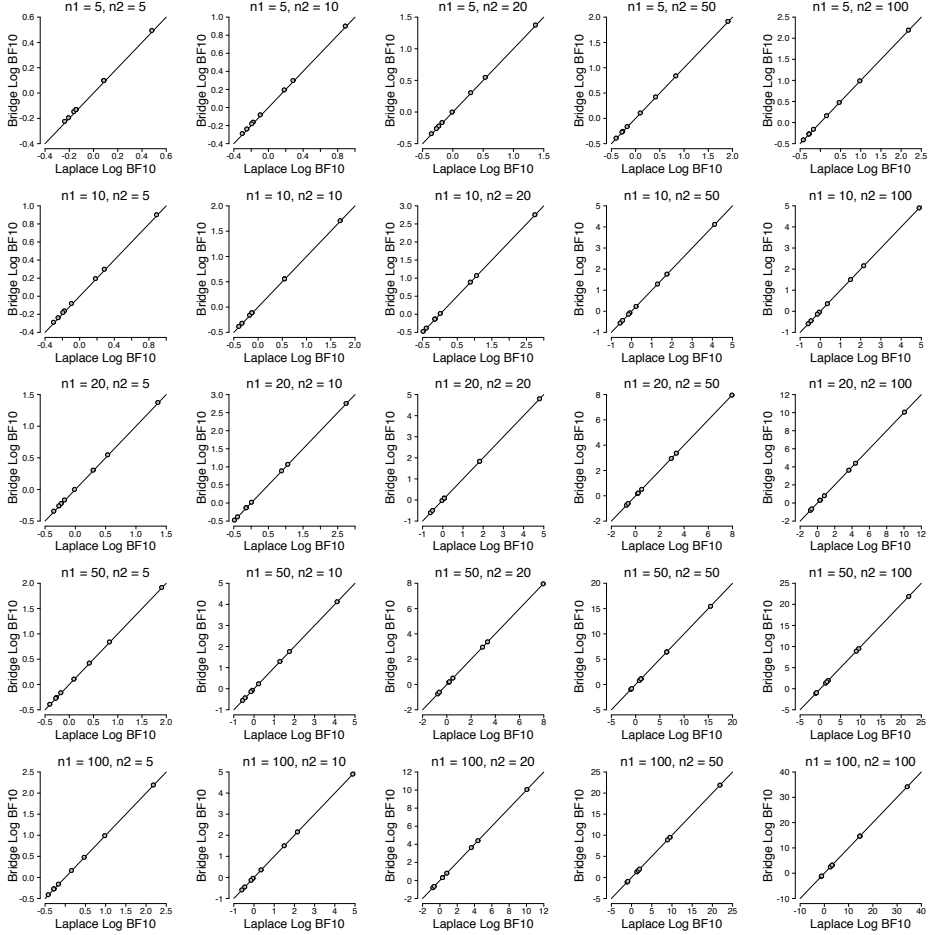


Figure 12.1: Comparison of the Laplace approximation and bridge sampling for computing the (log of the) Bayes factor  $\text{BF}_{10}$ . We considered all possible combinations of  $n_1 \in \{5, 10, 20, 50, 100\}$  and  $n_2 \in \{5, 10, 20, 50, 100\}$ . For each of the  $n_1$ - $n_2$  combinations, we considered all possible combinations of  $y_1 \in \{\frac{1}{5}n_1, \frac{2}{5}n_1, \frac{3}{5}n_1, \frac{4}{5}n_1\}$  and  $y_2 \in \{\frac{1}{5}n_2, \frac{2}{5}n_2, \frac{3}{5}n_2, \frac{4}{5}n_2\}$ . The results reveal that the two methods yield highly similar results, even when sample size is very small.

as follows:

$$\begin{aligned}
 p(\text{data} \mid \mathcal{H}_+) &= \int \int \underbrace{p(\text{data} \mid \beta, \xi)}_{\text{likelihood}} \underbrace{\pi_+(\beta, \xi)}_{\text{prior}} d\beta d\xi \\
 &\approx \frac{1}{S} \sum_{s=1}^S \frac{p(\text{data} \mid \tilde{\beta}_s, \tilde{\xi}_s) \pi_+(\tilde{\beta}_s, \tilde{\xi}_s)}{g_{\text{is}}(\tilde{\beta}_s, \tilde{\xi}_s)},
 \end{aligned} \tag{12.7}$$

where  $\left\{ \tilde{\beta}_s, \tilde{\xi}_s \right\}_{s=1}^S$  denotes  $S$  samples from the multivariate  $t$  importance density  $g_{\text{is}}$ , and

$$\pi_+(\beta, \xi) = \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2) \mathcal{N}_+(\exp(\xi); \mu_\psi, \sigma_\psi^2) \exp(\xi), \tag{12.8}$$

where  $\mathcal{N}(x; y, z)$  denotes the probability density function of a normal distribution with mean  $y$  and variance  $z$  that is evaluated at  $x$ . Furthermore,  $\mathcal{N}_+(x; y, z)$  denotes the density of a normal distribution that is truncated to allow only positive values for  $x$ . The marginal likelihood for  $\mathcal{H}_-$  is computed analogously.

### 12.2.6 Obtaining Posterior Samples

In a Bayesian A/B test application, one may not only be interested in testing hypotheses, but also in obtaining posterior samples for the model parameters under  $\mathcal{H}_1$ ,  $\mathcal{H}_+$ , and  $\mathcal{H}_-$ . The **abtest** package allows the user to obtain posterior samples using sampling importance resampling (e.g., Robert & Casella, 2010). Specifically, posterior samples for  $\mathcal{H}_+$  are obtained as follows (samples for the other hypotheses are obtained in an analogous manner):

1. Generate  $S$  samples from the multivariate  $t$  proposal distribution mentioned before, denoted by  $\left\{ \tilde{\beta}_s, \tilde{\xi}_s \right\}_{s=1}^S$ .
2. Compute the importance weights:

$$w_s = \frac{p(\text{data} \mid \tilde{\beta}_s, \tilde{\xi}_s) \pi_+(\tilde{\beta}_s, \tilde{\xi}_s)}{g_{\text{is}}(\tilde{\beta}_s, \tilde{\xi}_s)}, \quad s = 1, 2, \dots, S. \tag{12.9}$$

3. Renormalize the importance weights:  $v_s = w_s / \sum_{t=1}^S w_t$ ,  $s = 1, 2, \dots, S$ .
4. Resample (with replacement) from the samples obtained from the importance density according to the normalized importance weights  $v_s$  which yields (approximate) samples from the posterior distribution.

## 12.3 Example: Effectiveness of Resilience Training

Suppose the managers of a large consultancy firm are interested in reducing the number of employees who quit within the first six months, possibly due to the high stress involved in the job. A coaching company offers a resilience training and claims that this training greatly reduces the number of employees who quit.

Implementing the training for all newly hired employees would be expensive and some of the managers are not completely convinced that the training is at all effective. Therefore, the managers decide to run an A/B test where half of a sample of newly hired employees will receive the training, the other half will not be trained. The dependent variable is whether or not an employee quit within the first six months (1 = still on the job, 0 = quit).

### 12.3.1 Prior Specification

Before commencing the A/B test, the managers ask the coaching company to specify how effective they believe the training will be. The coaching company claims that, based on past experience with the training, they expect the proportion of employees who do not quit within the first six months to be 15% larger for the group who received the training, with a 95% uncertainty interval ranging from a 2.5% benefit to a 27.5% benefit. Assuming that the claimed 15% corresponds to the prior median, this expectation corresponds to a median absolute risk (i.e.,  $p_2 - p_1$ ) of 0.15 with a 95% uncertainty interval ranging from 0.025 to 0.275. The `elicit_prior` function can be used to encode this prior information:<sup>7</sup>

```
R> library("abtest")
R> prior_par <- elicit_prior(q = c(0.025, 0.15, 0.275),
+                             prob = c(.025, .5, .975),
+                             what = "arisk")
```

The obtained prior on the absolute risk can be visualized as follows:

```
R> plot_prior(prior_par, what = "arisk")
```

The resulting graph is shown in the top panel of Figure 12.2. The user can also visualize the (implied) prior for other quantities. For instance, the prior on the log odds ratio (middle panel of Figure 12.2) is obtained as follows:

```
R> plot_prior(prior_par, what = "logor")
```

The implied prior on the success probabilities  $p_1$  and  $p_2$  (bottom panel of Figure 12.2) is obtained as follows:

```
R> plot_prior(prior_par, what = "p1p2")
```

The bottom panel of Figure 12.2 illustrates that there is a dependency between  $p_1$  and  $p_2$  which is arguably desirable (Howard, 1998): When one of the success probabilities is very (small) large, it is likely that the other one will also be (small) large.

---

<sup>7</sup>All code and plots are also available at <https://osf.io/t3ajr/>.

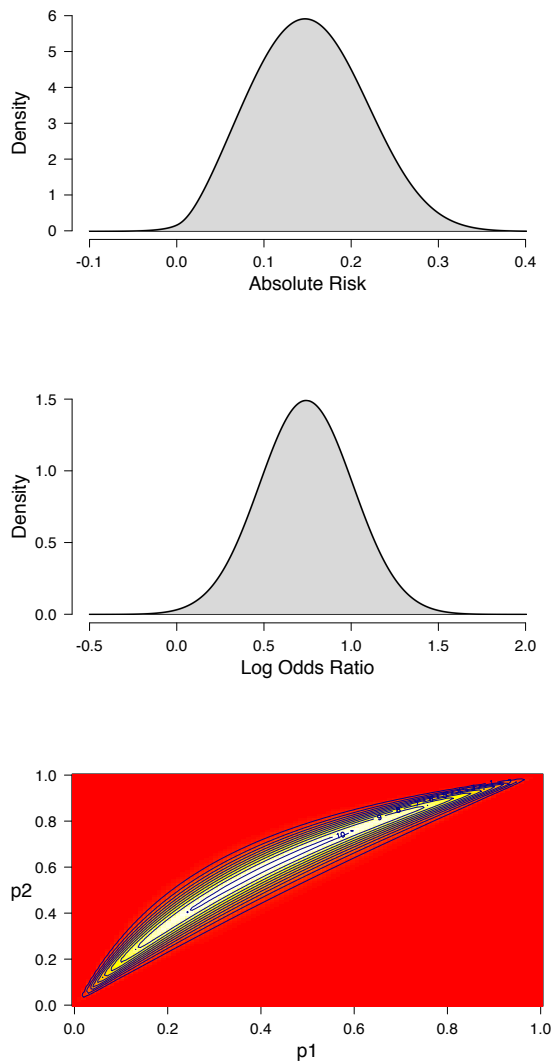


Figure 12.2: Elicited (implied) prior distributions for the effectiveness of the resilience training. The top panel displays the prior distribution for the absolute risk which corresponds to the difference between the probability of still being on the job for the trained and the non-trained employees (i.e.,  $p_2 - p_1$ ). The middle panel shows the prior distribution for the log odds ratio parameter  $\psi$ . The bottom panel displays the implied joint prior distribution for the success probabilities  $p_1$  and  $p_2$ . The bottom panel illustrates that the two success probabilities are assigned dependent priors. Furthermore, most prior mass is above the main diagonal which represents the coaching company's prior expectation that the training is successful.

### 12.3.2 Hypothesis Testing

After having specified the prior distribution for the test-relevant parameter, the consultancy firm starts to collect data. These (synthetic) data<sup>8</sup> are included in the **abtest** package (i.e., **seqdata**) and consist of a total of 1,000 observations (500 in each group). The number of employees still on the job after six months is 249 in the group without training and 269 in the trained group. Therefore, the observed success probabilities are  $\hat{p}_1 = .498$  in the control group and  $\hat{p}_2 = .538$  in the group that received training. Consequently, the observed success probabilities suggest that there is a positive effect of the training of 4%; however, a statistical analysis is required to assess whether this observed difference is statistically compelling. The **ab\_test** function can be used to conduct a Bayesian A/B test as follows:

```
R> data("seqdata")
R> set.seed(1)
R> ab <- ab_test(data = seqdata, prior_par = prior_par)
```

This yields the following output:

```
R> print(ab)
```

Bayesian A/B Test Results:

Bayes Factors:

```
BF10: 0.1406443
BF+0: 0.13823
BF-0: 0.4920187
```

Prior Probabilities Hypotheses:

```
H+: 0.25
H-: 0.25
H0: 0.5
```

Posterior Probabilities Hypotheses:

```
H+: 0.0526
H-: 0.1871
H0: 0.7604
```

The first part of the output presents Bayes factors in favor of the hypotheses  $\mathcal{H}_1$ ,  $\mathcal{H}_+$ , and  $\mathcal{H}_-$ , where the reference hypothesis (i.e., denominator of the Bayes factor) is  $\mathcal{H}_0$ . Since all three Bayes factors are smaller than 1, they all indicate evidence in favor of the null hypothesis of no effect. The next part of the output displays the

---

<sup>8</sup>The data set is structured such that the sequential nature of the data is retained: the data set contains the number of observations and the number of successes in each of the two groups after each observation.

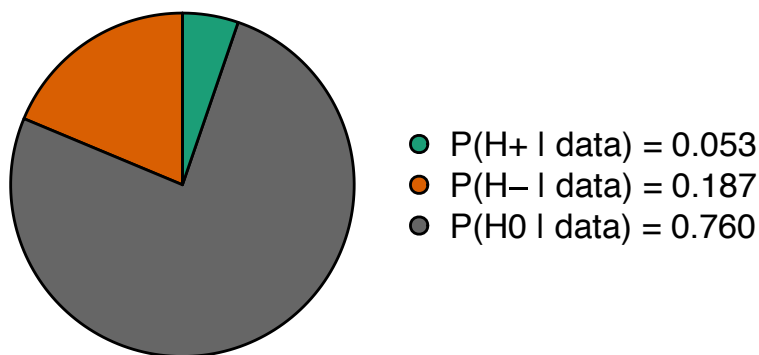


Figure 12.3: Posterior probabilities of the hypotheses visualized as a probability wheel.

prior probabilities of the hypotheses with non-zero prior probability. As explained before, the default setting assigns probability .50 to the null hypothesis and splits the remaining probability evenly across  $\mathcal{H}_+$  and  $\mathcal{H}_-$ . The user can change this default setting via the `prior_prob` argument (e.g., to assign non-zero probability to  $\mathcal{H}_1$ ). The final part of the output displays the posterior probabilities of the hypotheses with non-zero prior probability. The posterior probability of the null hypothesis  $\mathcal{H}_0$  indicates that the data have increased the plausibility of the null hypothesis from .50 to .76. Furthermore, the data have decreased the plausibility of both  $\mathcal{H}_+$  and  $\mathcal{H}_-$ .

As an aside, it may appear paradoxical that the data indicate a 4% positive effect of the training and yet the posterior probability of  $\mathcal{H}_-$  is larger than that of  $\mathcal{H}_+$ . The reason for this result is that the company's prior was overly ambitious, and  $\mathcal{H}_+$  is penalized for having predicted effects that are much too large. Furthermore, note that the test-relevant prior distribution under  $\mathcal{H}_-$  is obtained by truncating the prior on  $\psi$  at zero and renormalizing. Since the company's prior assigns almost all mass to positive log odds ratio values, renormalizing the negative part of the distribution results in a prior that is highly similar to  $\mathcal{H}_0$ ; this explains why  $\mathcal{H}_-$  receives non-trivial posterior probability. These considerations underscore the fact that the outcome of a Bayesian analysis is always relative to the specific set of models (and associated prior distributions) under consideration. Because highly informed priors can exert a large influence on the results, it is generally wise to examine the robustness of the conclusions by executing the default analysis as well. This analysis is reported in Appendix D.

The **abtest** package allows users to visualize the posterior probabilities of the hypotheses by means of a probability wheel (Figure 12.3):

```
R> prob_wheel(ab)
```

Overall, the data support the hypothesis that the training is ineffective over the company's hypothesis that the training is highly effective. The Bayes factor for

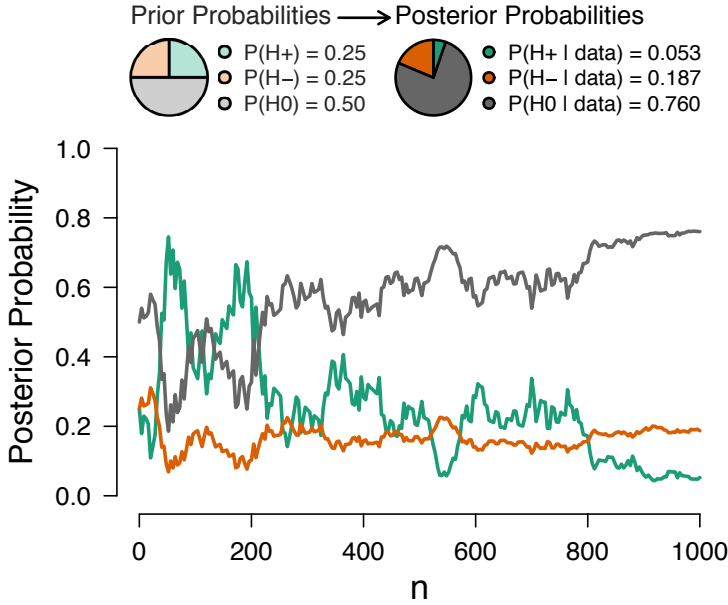


Figure 12.4: Sequential analysis results. The posterior probability of each hypothesis is plotted as a function of the number of observations across groups. On top, two probability wheels visualize the prior probabilities of the hypotheses and the posterior probabilities after taking into account all observations.

$\mathcal{H}_0$  over  $\mathcal{H}_+$  equals  $1/0.138 \approx 7.2$ , which indicates moderate evidence (Jeffreys, 1939, Appendix I).

Since the data set is of a sequential nature, it may be of interest to consider not only the result based on all observations, but to conduct also a sequential analysis that tracks the evidential flow as a function of the total number of observations (i.e., the number of observations across both groups). This sequential analysis can be conducted as follows:

```
R> plot_sequential(ab, thin = 4)
```

Setting the `thin` argument to 4 indicates that the evidence is computed after every 4th observation. Thinning can be useful to speed up the analysis in case the data set is very large or in case observations arrive in batches. Figure 12.4 displays the result of the sequential analysis. The posterior probability of each hypothesis with non-zero prior probability is plotted as a function of the total number of observations. At the top, two probability wheels visualize the prior probabilities of the hypotheses and the posterior probabilities of the hypotheses based on all available data. Figure 12.4 shows that after some initial fluctuation, adding more

observations increased the probability of the null hypothesis that there is no effect of the training.

### 12.3.3 Parameter Estimation

The data indicate evidence in favor of the null hypothesis versus the hypothesis that the training is highly effective, leaving open the possibility that the training does have an effect, but of a more modest size than the company anticipated. To assess this possibility one may investigate the potential size of the effect under the assumption that the effect is non-zero.<sup>9</sup> For parameter estimation, we generally prefer to investigate the posterior distribution for the unconstrained alternative hypothesis  $\mathcal{H}_1$ ; however, the **abtest** package also provides posterior samples and plotting functionality for the constrained hypotheses  $\mathcal{H}_+$  and  $\mathcal{H}_-$ .

The top panel of Figure 12.5 displays the posterior distribution for the absolute risk (i.e.,  $p_2 - p_1$ ) that can be obtained as follows:

```
R> plot_posterior(ab, what = "arisk")
```

The top panel of Figure 12.5 shows the prior distribution as a dotted line and the posterior distribution (with 95% central credible interval) as a solid line. The plot indicates that, under the assumption that the difference between the two success probabilities is not exactly zero, it is likely to be smaller than expected: the posterior median is 0.067 and the 95% central credible interval ranges from 0.011 to 0.122.

The middle panel of Figure 12.5 displays the posterior distribution for the log odds ratio  $\psi$  that can be obtained as follows:

```
R> plot_posterior(ab, what = "logor")
```

The middle panel of Figure 12.5 indicates that, given the log odds ratio is not exactly zero, it is likely to be between 0.043 and 0.492, where the posterior median is 0.267.

It may also be of interest to consider the marginal posterior distributions of the success probabilities  $p_1$  and  $p_2$ . This plot can be produced as follows:

```
R> plot_posterior(ab, what = "p1p2")
```

The bottom panel of Figure 12.5 displays the resulting plot. In this example,  $p_1$  and  $p_2$  correspond to the probability of still being on the job after six months for the non-trained employees and the employees that received the training, respectively. The bottom panel of Figure 12.5 indicates that the posterior median for  $p_1$  is 0.485, with 95% credible ranging from 0.443 to 0.527, and the posterior median for  $p_2$  is 0.551, with 95% credible interval ranging from 0.509 to 0.592.

In sum, this synthetic data set offers modest evidence in favor of the null hypothesis which states that the training is not effective over the hypothesis that the training is highly effective; nevertheless, the consultancy firm should probably

---

<sup>9</sup>For consistency, we continue this analysis with the company's prior; an analysis with the less enthusiastic default prior is provided in Appendix D.

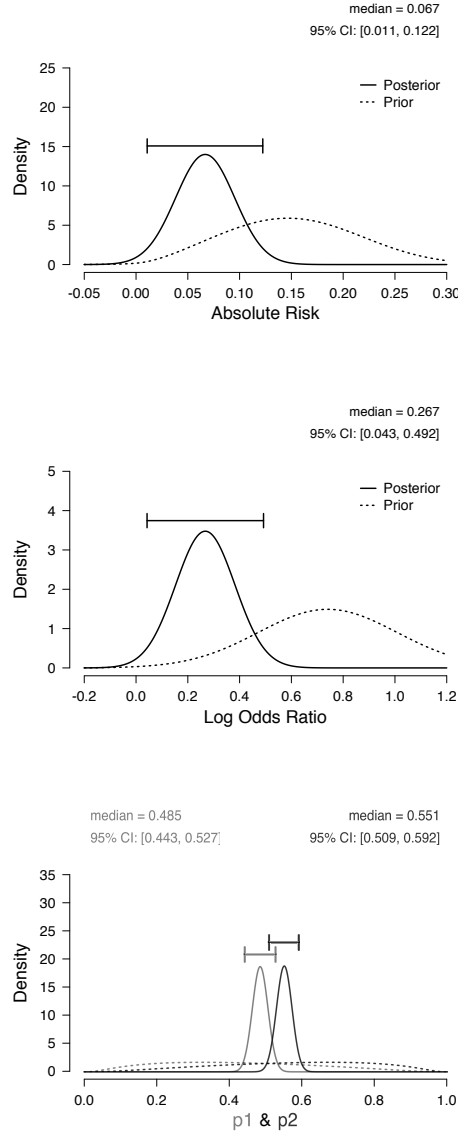


Figure 12.5: (Implied) prior and posterior distributions under  $\mathcal{H}_1$ . The dotted lines display the prior distributions, the solid lines display the posterior distributions (with 95% central credible intervals). The medians and the bounds of the 95% central credible intervals are displayed on top of each panel. The top panel displays the posterior distribution for the absolute risk (i.e.,  $p_2 - p_1$ ); the middle panel shows the posterior distribution for the log odds ratio parameter  $\psi$ ; the bottom panel displays the marginal posterior distributions for the success probabilities  $p_1$  and  $p_2$ .

continue to collect data in order to obtain more compelling evidence before deciding whether or not the training should be implemented. If the true effect is as small as 4%, continued testing will ultimately show compelling evidence for  $\mathcal{H}_+$  over  $\mathcal{H}_0$ . Note that continued testing is trivial in the Bayesian framework: the results can simply be updated as new observations arrive.

## 12.4 Concluding Comments

In this chapter, we have introduced the **abtest** package that implements both Bayesian hypothesis testing and Bayesian estimation for the A/B test using informed priors. The procedure allows users to (1) obtain evidence in favor of the null hypothesis; (2) monitor the evidence as data accumulate; and (3) elicit and incorporate expert prior distributions. We hope that the provided analysis approach is useful across different fields that apply A/B testing on a routine basis, particularly business and medicine.

Despite the practical benefits that the package offers right now, there are areas for future improvement. For instance, **abtest** currently allows users to compare two groups; however, there are applications in which one may be interested in simultaneously comparing more than two groups. Furthermore, at the moment, **abtest** expects the dependent variable to be binary. Nevertheless, in certain scenarios, it may be more natural to compare the two groups based on a continuous outcome variable. This scenario resembles an independent samples *t*-test for which well-established Bayesian procedures exist (e.g., Ly et al., 2016b; Rouder et al., 2009) which are available, for instance, in the **BayesFactor** package (Morey & Rouder, 2015) and JASP (JASP Team, 2020).<sup>10</sup> Moreover, currently, the **abtest** package does not provide functions for generating predictions. Note, however, that users can generate predictions in a straightforward manner themselves based on the posterior samples that are provided by **abtest**. The implementation also does not allow users to incorporate utilities explicitly (e.g., Lindley, 1985). However, again, based on the provided posterior probabilities and posterior samples, users who wish to take into account utilities may do so in a relatively straightforward way. Furthermore, users interested in adjusting the model used in **abtest** (e.g., to account for hierarchically-structured data or covariates) are referred to general-purpose Bayesian software such as **Stan** (Carpenter et al., 2017; Stan Development Team, 2016) and the related R package **brms** (Bürkner, 2017). In combination with the **bridgesampling** package (Gronau, Singmann, & Wagenmakers, 2020), this enables the user to compare custom models using Bayes factors and posterior model probabilities. A more structural limitation of **abtest** is that it has been developed to analyze A/B test data, but not to run the A/B test experiment itself.

In sum, A/B testing is ubiquitous in business and medicine. Here we have demonstrated how the **abtest** package enables relatively complete Bayesian inference including the capability to obtain support for the null, continuously monitor the results, and elicit and incorporate expert prior knowledge. Hopefully, this

---

<sup>10</sup>For a list of Bayesian R packages, see <https://cran.r-project.org/web/views/Bayesian.html>.

approach forms a basis for evidence-based conclusions that will benefit both businesses and patients.

R code for reproducing the analyses presented in this chapter is available at <https://osf.io/t3ajr/>.

## 12.A Interpretation of the Parameters

Here we show that  $\beta$  corresponds to the grand mean of the log odds and that  $\psi$  corresponds to the log odds ratio (for the model definition, see Equation 12.1). The nuisance parameter  $\beta$  corresponds to the grand mean of the log odds since

$$\frac{1}{2} \log \left( \frac{p_1}{1-p_1} \right) + \frac{1}{2} \log \left( \frac{p_2}{1-p_2} \right) = \frac{1}{2} \beta - \frac{1}{4} \psi + \frac{1}{2} \beta + \frac{1}{4} \psi = \beta.$$

The test-relevant parameter  $\psi$  corresponds to the log odds ratio since

$$\log \left( \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} \right) = \log \left( \frac{p_2}{1-p_2} \right) - \log \left( \frac{p_1}{1-p_1} \right) = \beta + \frac{\psi}{2} - \left( \beta - \frac{\psi}{2} \right) = \psi.$$

## 12.B Prior Elicitation: Implied Distributions

The prior elicitation approach described in Equation 12.2 requires the cdf's for the quantities of interest. Here, we derive the implied cdf's for these quantities; we also derive the corresponding probability density functions (pdf's). Additionally, we derive four further implied distributions of interest: the joint pdf of  $p_1$  and  $p_2$ , the conditional pdf of  $p_2$  given  $p_1$  is fixed to a particular value, the marginal distribution for  $p_1$ , and the marginal distribution for  $p_2$ . A few of these expressions will contain a one-dimensional integral which can easily be evaluated using numerical integration.

### 12.B.1 Log Odds Ratio

Since  $\psi$  itself corresponds to the log odds ratio,  $F(\cdot; \mu_\psi, \sigma_\psi)$  corresponds in this case to the cdf of a normal distribution with mean  $\mu_\psi$  and standard deviation  $\sigma_\psi$ . The corresponding pdf is the normal probability density function.

### 12.B.2 Odds Ratio

The implied prior on the odds ratio  $\omega = \exp(\psi)$  is a log-normal distribution. Hence,  $F(\cdot; \mu_\psi, \sigma_\psi)$  corresponds in this case to the cdf of a log-normal distribution with parameters  $\mu_\psi$  and  $\sigma_\psi$ . The corresponding pdf is the log-normal probability density function.

### 12.B.3 Relative Risk

The relative risk is given by  $\Lambda = \frac{p_2}{p_1}$ . We use a capital letter (i.e.,  $\Lambda$ ) to refer to the random variable and use a lower-case letter (i.e.,  $\lambda$ ) to refer to a concrete realization. Note that so far, we have abused notation by only using lower-case letters, but it should be clear from the context when we referred to a random variable or a concrete realization. However, for deriving the following cdf, we need the distinction to keep the notation clear. To derive the implied cdf for the relative

risk, we proceed as follows:

$$\begin{aligned} P(\Lambda \leq \lambda) &= P\left(\frac{p_2}{p_1} \leq \lambda\right) \\ &= P(p_2 \leq \lambda p_1) \\ &= P\left(\frac{1}{1 + \exp\left(-\beta - \frac{\psi}{2}\right)} \leq \frac{\lambda}{1 + \exp\left(-\beta + \frac{\psi}{2}\right)}\right). \end{aligned}$$

Taking reciprocals and some algebra yields

$$P\left(\left(\exp\left(\frac{\psi}{2}\right)\right)^2 + (1 - \lambda) \exp(\beta) \exp\left(\frac{\psi}{2}\right) - \lambda \leq 0\right).$$

When we set

$$\left(\exp\left(\frac{\psi}{2}\right)\right)^2 + (1 - \lambda) \exp(\beta) \exp\left(\frac{\psi}{2}\right) - \lambda = 0,$$

we can solve for  $\psi$  using the fact that this is a quadratic equation in  $\exp\left(\frac{\psi}{2}\right)$  and we obtain:

$$\exp\left(\frac{\psi}{2}\right) = \frac{-(1 - \lambda) \exp(\beta) + \sqrt{(1 - \lambda)^2 \exp(2\beta) + 4\lambda}}{2},$$

where we took into account that  $\exp\left(\frac{\psi}{2}\right)$  needs to be positive (i.e., we omitted the solution corresponding to minus the square root). Hence,

$$\psi = 2 \log \left( \frac{-(1 - \lambda) \exp(\beta) + \sqrt{(1 - \lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right).$$

Therefore,  $\left(\exp\left(\frac{\psi}{2}\right)\right)^2 + (1 - \lambda) \exp(\beta) \exp\left(\frac{\psi}{2}\right) - \lambda \leq 0$  whenever

$$\psi \leq 2 \log \left( \frac{-(1 - \lambda) \exp(\beta) + \sqrt{(1 - \lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right).$$

Hence, the desired cdf can be written as

$$\begin{aligned} &P\left(\psi \leq 2 \log \left( \frac{-(1 - \lambda) \exp(\beta) + \sqrt{(1 - \lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right)\right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{2 \log \left( \frac{-(1 - \lambda) \exp(\beta) + \sqrt{(1 - \lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right)} \mathcal{N}(\psi; \mu_\psi, \sigma_\psi^2) \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2) d\psi d\beta \\ &= \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2) \Phi\left(2 \log \left( \frac{-(1 - \lambda) \exp(\beta) + \sqrt{(1 - \lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right); \mu_\psi, \sigma_\psi^2\right) d\beta, \end{aligned} \tag{12.10}$$

where  $\Phi(\cdot; \mu_\psi, \sigma_\psi^2)$  denotes the cdf of a normal distribution with mean  $\mu_\psi$  and variance  $\sigma_\psi^2$ , and  $\mathcal{N}(\cdot; \mu_\beta, \sigma_\beta^2)$  denotes the corresponding pdf.

The pdf of the relative risk is obtained by taking the derivative with respect to  $\lambda$ :

$$\begin{aligned} \frac{d}{d\lambda} & \left[ \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2) \Phi \left( 2 \log \left( \frac{-(1-\lambda) \exp(\beta) + \sqrt{(1-\lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right); \mu_\psi, \sigma_\psi^2 \right) d\beta \right] \\ &= \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_\beta, \sigma_\beta^2) \mathcal{N} \left( 2 \log \left( \frac{-(1-\lambda) \exp(\beta) + \sqrt{(1-\lambda)^2 \exp(2\beta) + 4\lambda}}{2} \right); \mu_\psi, \sigma_\psi^2 \right) \\ & \quad \times 2 \left[ \frac{\exp(\beta) + \frac{2-(1-\lambda) \exp(2\beta)}{\sqrt{(1-\lambda)^2 \exp(2\beta) + 4\lambda}}}{-(1-\lambda) \exp(\beta) + \sqrt{(1-\lambda)^2 \exp(2\beta) + 4\lambda}} \right] d\beta. \end{aligned} \quad (12.11)$$

### 12.B.4 Absolute Risk

The absolute risk is given by  $\Upsilon = p_2 - p_1$ . We use the upper-case letter  $\Upsilon$  to refer to the random variable and the lower-case letter  $v$  to refer to a concrete realization. To derive the implied cdf for the absolute risk, we proceed as follows:

$$\begin{aligned} P(\Upsilon \leq v) &= P(p_2 - p_1 \leq v) \\ &= P(p_2 \leq v + p_1) \\ &= P \left( \frac{1}{1 + \exp \left( -\beta - \frac{\psi}{2} \right)} \leq v + \frac{1}{1 + \exp \left( -\beta + \frac{\psi}{2} \right)} \right). \end{aligned}$$

After some algebra, we obtain

$$P \left( \exp(\beta) (1 - v) \left( \exp \left( \frac{\psi}{2} \right) \right)^2 - v (\exp(2\beta) + 1) \exp \left( \frac{\psi}{2} \right) - \exp(\beta) (v + 1) \leq 0 \right).$$

When we set

$$\exp(\beta) (1 - v) \left( \exp \left( \frac{\psi}{2} \right) \right)^2 - v (\exp(2\beta) + 1) \exp \left( \frac{\psi}{2} \right) - \exp(\beta) (v + 1) = 0,$$

we can solve for  $\psi$  using the fact that this is a quadratic equation in  $\exp \left( \frac{\psi}{2} \right)$  and we obtain:

$$\exp \left( \frac{\psi}{2} \right) = \frac{v (\exp(2\beta) + 1) + \sqrt{v^2 (\exp(2\beta) - 1)^2 + 4 \exp(2\beta)}}{2 \exp(\beta) (1 - v)},$$

where we took into account that  $\exp \left( \frac{\psi}{2} \right)$  needs to be positive (i.e., we omitted the solution corresponding to minus the square root). Hence,

$$\psi = 2 \log \left( \frac{v (\exp(2\beta) + 1) + \sqrt{v^2 (\exp(2\beta) - 1)^2 + 4 \exp(2\beta)}}{2 \exp(\beta) (1 - v)} \right).$$

Therefore,

$$\exp(\beta)(1-v)\left(\exp\left(\frac{\psi}{2}\right)\right)^2 - v(\exp(2\beta)+1)\exp\left(\frac{\psi}{2}\right) - \exp(\beta)(v+1) \leq 0$$

whenever

$$\psi \leq 2 \log \left( \frac{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-v)} \right).$$

Hence, the desired cdf can be written as

$$\begin{aligned} & P \left( \psi \leq 2 \log \left( \frac{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-v)} \right) \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{2 \log \left( \frac{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-v)} \right)} \mathcal{N}(\psi; \mu_{\psi}, \sigma_{\psi}^2) \mathcal{N}(\beta; \mu_{\beta}, \sigma_{\beta}^2) d\psi d\beta \\ &= \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_{\beta}, \sigma_{\beta}^2) \Phi \left( 2 \log \left( \frac{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-v)} \right); \mu_{\psi}, \sigma_{\psi}^2 \right) d\beta. \end{aligned} \quad (12.12)$$

The pdf of the absolute risk is obtained by taking the derivative with respect to  $v$ :

$$\begin{aligned} & \frac{d}{dv} \left[ \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_{\beta}, \sigma_{\beta}^2) \Phi \left( 2 \log \left( \frac{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-v)} \right); \mu_{\psi}, \sigma_{\psi}^2 \right) d\beta \right] \\ &= \int_{-\infty}^{\infty} \mathcal{N}(\beta; \mu_{\beta}, \sigma_{\beta}^2) \mathcal{N} \left( 2 \log \left( \frac{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}}{2\exp(\beta)(1-v)} \right); \mu_{\psi}, \sigma_{\psi}^2 \right) \\ & \quad \times 2 \left[ \frac{\exp(2\beta) + \frac{v(\exp(2\beta)-1)^2}{\sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}} + 1}{v(\exp(2\beta)+1) + \sqrt{v^2(\exp(2\beta)-1)^2 + 4\exp(2\beta)}} + \frac{1}{1-v} \right] d\beta. \end{aligned} \quad (12.13)$$

### 12.B.5 Joint Distribution of $p_1$ and $p_2$

Another distribution of interest is the implied joint distribution of the two success probabilities  $p_1$  and  $p_2$ . This distribution will not be used to elicit the prior on  $\psi$  which is the reason why we only derive the pdf and not the cdf. The model parameters  $\beta$  and  $\psi$  are related to  $p_1$  and  $p_2$  as follows:

$$\begin{aligned} \log \left( \frac{p_1}{1-p_1} \right) &= \beta - \frac{\psi}{2} \\ \log \left( \frac{p_2}{1-p_2} \right) &= \beta + \frac{\psi}{2}. \end{aligned}$$

Hence, the inverse transformation is given by:

$$\begin{aligned}\beta &= \frac{1}{2} \log \left( \frac{p_1}{1-p_1} \right) + \frac{1}{2} \log \left( \frac{p_2}{1-p_2} \right) \\ \psi &= \log \left( \frac{p_2}{1-p_2} \right) - \log \left( \frac{p_1}{1-p_1} \right).\end{aligned}$$

The corresponding Jacobian is:

$$\begin{aligned}|J| &= \left| \begin{pmatrix} \frac{\partial \beta}{\partial p_1} & \frac{\partial \beta}{\partial p_2} \\ \frac{\partial \psi}{\partial p_1} & \frac{\partial \psi}{\partial p_2} \end{pmatrix} \right| \\ &= \left| \begin{pmatrix} \frac{1}{2} \frac{1}{p_1(1-p_1)} & \frac{1}{2} \frac{1}{p_2(1-p_2)} \\ -\frac{1}{p_1(1-p_1)} & \frac{1}{p_2(1-p_2)} \end{pmatrix} \right| \\ &= \frac{1}{p_1 p_2 (1-p_1)(1-p_2)}.\end{aligned}$$

Therefore, the joint pdf of  $p_1$  and  $p_2$  is given by:

$$\begin{aligned}p(p_1, p_2) &= \frac{1}{p_1 p_2 (1-p_1)(1-p_2)} \mathcal{N} \left( \frac{1}{2} \left[ \log \left( \frac{p_1}{1-p_1} \right) + \log \left( \frac{p_2}{1-p_2} \right) \right]; \mu_\beta, \sigma_\beta^2 \right) \\ &\quad \times \mathcal{N} \left( \log \left( \frac{p_2}{1-p_2} \right) - \log \left( \frac{p_1}{1-p_1} \right); \mu_\psi, \sigma_\psi^2 \right).\end{aligned}\tag{12.14}$$

### 12.B.6 Marginal Distribution of $p_1$

The marginal distribution of  $p_1$  is given by:

$$\begin{aligned}p(p_1) &= \int_0^1 p(p_1, p'_2) dp'_2 \\ &= \int_0^1 \frac{1}{p_1 p'_2 (1-p_1)(1-p'_2)} \mathcal{N} \left( \frac{1}{2} \left[ \log \left( \frac{p_1}{1-p_1} \right) + \log \left( \frac{p'_2}{1-p'_2} \right) \right]; \mu_\beta, \sigma_\beta^2 \right) \\ &\quad \times \mathcal{N} \left( \log \left( \frac{p'_2}{1-p'_2} \right) - \log \left( \frac{p_1}{1-p_1} \right); \mu_\psi, \sigma_\psi^2 \right) dp'_2.\end{aligned}\tag{12.15}$$

### 12.B.7 Marginal Distribution of $p_2$

The marginal distribution of  $p_2$  is given by:

$$\begin{aligned}p(p_2) &= \int_0^1 p(p'_1, p_2) dp'_1 \\ &= \int_0^1 \frac{1}{p'_1 p_2 (1-p'_1)(1-p_2)} \mathcal{N} \left( \frac{1}{2} \left[ \log \left( \frac{p'_1}{1-p'_1} \right) + \log \left( \frac{p_2}{1-p_2} \right) \right]; \mu_\beta, \sigma_\beta^2 \right) \\ &\quad \times \mathcal{N} \left( \log \left( \frac{p_2}{1-p_2} \right) - \log \left( \frac{p'_1}{1-p'_1} \right); \mu_\psi, \sigma_\psi^2 \right) dp'_1.\end{aligned}\tag{12.16}$$

### 12.B.8 Conditional Distribution of $p_2$ given $p_1$

Another distribution of interest is the conditional distribution of the second success probability  $p_2$  given a particular value of  $p_1$ . This distribution will not be used for prior elicitation which is the reason why we only present the expression for the pdf which is given by:

$$\begin{aligned} p(p_2 | p_1) &= \frac{p(p_1, p_2)}{\int_0^1 p(p_1, p'_2) dp'_2} \\ &= \frac{\frac{1}{p_2(1-p_2)} \mathcal{N}\left(\frac{1}{2} \left[ \log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2}{1-p_2}\right) \right]; \mu_\beta, \sigma_\beta^2\right) \mathcal{N}\left(\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right); \mu_\psi, \sigma_\psi^2\right)}{\int_0^1 \frac{1}{p'_2(1-p'_2)} \mathcal{N}\left(\frac{1}{2} \left[ \log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p'_2}{1-p'_2}\right) \right]; \mu_\beta, \sigma_\beta^2\right) \mathcal{N}\left(\log\left(\frac{p'_2}{1-p'_2}\right) - \log\left(\frac{p_1}{1-p_1}\right); \mu_\psi, \sigma_\psi^2\right) dp'_2}. \end{aligned} \quad (12.17)$$

### 12.B.9 Implied Distributions for Truncated Priors on the Log Odds Ratio

Note that the above expressions can be all easily modified in case the prior on the log odds ratio  $\psi$  is a truncated normal distribution (e.g., restricting  $\psi$  to be larger/smaller than zero) which is the case for the hypotheses  $\mathcal{H}_+$  and  $\mathcal{H}_-$ . In this case, the normal prior density function and cumulative distribution function for  $\psi$  simply need to be replaced by the truncated versions. For the implied log-normal prior on the odds ratio, the truncation bounds simply need to be exponentiated to obtain the truncation bounds with respect to the log-normal prior.

## 12.C Laplace Approximation Details

The Laplace approximations require first-order and second-order derivatives. Let us first state explicitly the functions for which we need to find the derivatives. For  $\mathcal{H}_0$  we have:

$$\begin{aligned} l_0^*(\beta) &= \log \{p(y | \beta) \pi_0(\beta)\} \\ &= (y_1 + y_2) \log \left( \frac{\exp(\beta)}{1 + \exp(\beta)} \right) + (n_1 + n_2 - y_1 - y_2) \log \left( 1 - \frac{\exp(\beta)}{1 + \exp(\beta)} \right) \\ &\quad - \frac{1}{2} \log (2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2} (\beta - \mu_\beta)^2. \end{aligned} \quad (12.18)$$

For  $\mathcal{H}_1$  we have:

$$\begin{aligned}
 l^*(\beta, \psi) &= \log \{p(y \mid \beta, \psi) \pi(\beta, \psi)\} \\
 &= y_1 \log \left( \frac{\exp(\beta - \frac{\psi}{2})}{1 + \exp(\beta - \frac{\psi}{2})} \right) + (n_1 - y_1) \log \left( 1 - \frac{\exp(\beta - \frac{\psi}{2})}{1 + \exp(\beta - \frac{\psi}{2})} \right) \\
 &\quad + y_2 \log \left( \frac{\exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})} \right) + (n_2 - y_2) \log \left( 1 - \frac{\exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})} \right) \\
 &\quad - \frac{1}{2} \log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2 - \frac{1}{2} \log(2\pi\sigma_\psi^2) - \frac{1}{2\sigma_\psi^2}(\psi - \mu_\psi)^2.
 \end{aligned} \tag{12.19}$$

For  $\mathcal{H}_+$  we have:

$$\begin{aligned}
 l_+^*(\beta, \xi) &= \log \{p(y \mid \beta, \xi) \pi_+(\beta, \xi)\} \\
 &= y_1 \log \left( \frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} \right) + (n_1 - y_1) \log \left( 1 - \frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} \right) \\
 &\quad + y_2 \log \left( \frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} \right) + (n_2 - y_2) \log \left( 1 - \frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} \right) \\
 &\quad - \frac{1}{2} \log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2 \\
 &\quad - \frac{1}{2} \log(2\pi\sigma_\psi^2) - \frac{1}{2\sigma_\psi^2}(\exp(\xi) - \mu_\psi)^2 - \log(1 - \Phi(0; \mu_\psi, \sigma_\psi^2)) + \xi.
 \end{aligned} \tag{12.20}$$

Finally, for  $\mathcal{H}_-$  we have

$$\begin{aligned}
 l_-^*(\beta, \xi) &= \log \{p(y \mid \beta, \xi) \pi_-(\beta, \xi)\} \\
 &= y_1 \log \left( \frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} \right) + (n_1 - y_1) \log \left( 1 - \frac{\exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} \right) \\
 &\quad + y_2 \log \left( \frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} \right) + (n_2 - y_2) \log \left( 1 - \frac{\exp(\beta - \frac{\exp(\xi)}{2})}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} \right) \\
 &\quad - \frac{1}{2} \log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2 \\
 &\quad - \frac{1}{2} \log(2\pi\sigma_\psi^2) - \frac{1}{2\sigma_\psi^2}(-\exp(\xi) - \mu_\psi)^2 - \log(\Phi(0; \mu_\psi, \sigma_\psi^2)) + \xi.
 \end{aligned} \tag{12.21}$$

### 12.C.1 First-order Derivatives

The first-order derivatives are used to find the modes for the Laplace approximations. As shown below, we can find these derivatives analytically; however, setting

the derivatives equal to zero and solving for the parameters is not straightforward. Nevertheless, having these derivatives is useful not only as an intermediate step to finding the second-order derivatives but also for finding the modes: This allows us to provide numerical optimizers with the analytic expressions for the derivatives which can increase speed and accuracy for numerically finding the modes of the relevant functions.

The first-order derivative for  $l_0(\beta)$  is given by:

$$\frac{d}{d\beta} l_0^*(\beta) = \frac{y_1 + y_2 - (n_1 + n_2 - y_1 - y_2) \exp(\beta)}{1 + \exp(\beta)} - \frac{\beta - \mu_\beta}{\sigma_\beta^2}. \quad (12.22)$$

The first-order partial derivatives for  $l^*(\beta, \psi)$  are given by

$$\frac{\partial}{\partial \beta} l^*(\beta, \psi) = \frac{y_1 - (n_1 - y_1) \exp(\beta - \frac{\psi}{2})}{1 + \exp(\beta - \frac{\psi}{2})} + \frac{y_2 - (n_2 - y_2) \exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})} - \frac{\beta - \mu_\beta}{\sigma_\beta^2}, \quad (12.23)$$

and

$$\begin{aligned} \frac{\partial}{\partial \psi} l^*(\beta, \psi) = & \frac{1}{2} \left( \frac{(n_1 - y_1) \exp(\beta - \frac{\psi}{2}) - y_1}{1 + \exp(\beta - \frac{\psi}{2})} + \frac{y_2 - (n_2 - y_2) \exp(\beta + \frac{\psi}{2})}{1 + \exp(\beta + \frac{\psi}{2})} \right) \\ & - \frac{\psi - \mu_\psi}{\sigma_\psi^2}. \end{aligned} \quad (12.24)$$

The first-order partial derivatives for  $l_+^*(\beta, \xi)$  are given by:

$$\begin{aligned} \frac{\partial}{\partial \beta} l_+^*(\beta, \xi) = & \frac{y_1 - (n_1 - y_1) \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)} + \frac{y_2 - (n_2 - y_2) \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)} \\ & - \frac{\beta - \mu_\beta}{\sigma_\beta^2}, \end{aligned} \quad (12.25)$$

and

$$\begin{aligned} \frac{\partial}{\partial \xi} l_+^*(\beta, \xi) = & \frac{\exp(\xi)}{2} \left( \frac{(n_1 - y_1) \exp(\beta - \frac{\exp(\xi)}{2}) - y_1}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} + \frac{y_2 - (n_2 - y_2) \exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} \right) \\ & - \exp(\xi) \frac{\exp(\xi) - \mu_\psi}{\sigma_\psi^2} + 1. \end{aligned} \quad (12.26)$$

The first-order partial derivatives for  $l_-^*(\beta, \xi)$  are given by:

$$\begin{aligned} \frac{\partial}{\partial \beta} l_-^*(\beta, \xi) = & \frac{y_1 - (n_1 - y_1) \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)} + \frac{y_2 - (n_2 - y_2) \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)} \\ & - \frac{\beta - \mu_\beta}{\sigma_\beta^2}, \end{aligned} \quad (12.27)$$

and

$$\begin{aligned} \frac{\partial}{\partial \xi} l_{-}^{*}(\beta, \xi) = & \frac{\exp(\xi)}{2} \left( \frac{y_1 - (n_1 - y_1) \exp(\beta + \frac{\exp(\xi)}{2})}{1 + \exp(\beta + \frac{\exp(\xi)}{2})} + \frac{(n_2 - y_2) \exp(\beta - \frac{\exp(\xi)}{2}) - y_2}{1 + \exp(\beta - \frac{\exp(\xi)}{2})} \right) \\ & + \exp(\xi) \frac{-\exp(\xi) - \mu_{\psi}}{\sigma_{\psi}^2} + 1. \end{aligned} \quad (12.28)$$

### 12.C.2 Second-order Derivatives

For the Laplace approximations, we also need the inverse of the negative Hessians. The Hessian is the matrix with the second-order partial derivatives which is the reason why we now present expressions for the second-order partial derivatives. Note that under all hypotheses there are either one or two parameters. Hence, the Hessians will be at most 2 by 2 matrices. For matrices up to 2 by 2, it is straightforward to find the inverse and the determinant which makes it easy to obtain the quantities needed for the Laplace approximations once we have the required derivatives.

For  $l_0^{*}(\beta)$ , there is only one parameter and the second-order derivative is given by:

$$\frac{d^2}{d\beta^2} l_0^{*}(\beta) = -\frac{(n_1 + n_2) \exp(\beta)}{(1 + \exp(\beta))^2} - \frac{1}{\sigma_{\beta}^2}. \quad (12.29)$$

For  $l^{*}(\beta, \psi)$  the second-order partial derivatives are given by

$$\frac{\partial^2}{\partial \beta^2} l^{*}(\beta, \psi) = -\frac{n_1 \exp(\beta - \frac{\psi}{2})}{\left(1 + \exp(\beta - \frac{\psi}{2})\right)^2} - \frac{n_2 \exp(\beta + \frac{\psi}{2})}{\left(1 + \exp(\beta + \frac{\psi}{2})\right)^2} - \frac{1}{\sigma_{\beta}^2}, \quad (12.30)$$

and

$$\frac{\partial^2}{\partial \beta \partial \psi} l^{*}(\beta, \psi) = \frac{1}{2} \left( \frac{n_1 \exp(\beta - \frac{\psi}{2})}{\left(1 + \exp(\beta - \frac{\psi}{2})\right)^2} - \frac{n_2 \exp(\beta + \frac{\psi}{2})}{\left(1 + \exp(\beta + \frac{\psi}{2})\right)^2} \right), \quad (12.31)$$

and

$$\frac{\partial^2}{\partial \psi^2} l^{*}(\beta, \psi) = -\frac{1}{4} \left( \frac{n_1 \exp(\beta - \frac{\psi}{2})}{\left(1 + \exp(\beta - \frac{\psi}{2})\right)^2} + \frac{n_2 \exp(\beta + \frac{\psi}{2})}{\left(1 + \exp(\beta + \frac{\psi}{2})\right)^2} \right) - \frac{1}{\sigma_{\psi}^2}. \quad (12.32)$$

For  $l_{+}^{*}(\beta, \xi)$  the second-order partial derivatives are given by

$$\frac{\partial^2}{\partial \beta^2} l_{+}^{*}(\beta, \xi) = -\frac{n_1 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{\sigma_{\beta}^2}, \quad (12.33)$$

and

$$\frac{\partial^2}{\partial\beta\partial\xi} l_+^*(\beta, \xi) = \frac{\exp(\xi)}{2} \left( \frac{n_1 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} \right), \quad (12.34)$$

and

$$\begin{aligned} \frac{\partial^2}{\partial\xi^2} l_+^*(\beta, \xi) &= \frac{\exp(\xi)}{2} \left( \frac{(n_1 - y_1) \exp\left(\beta - \frac{\exp(\xi)}{2}\right) - y_1}{1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)} + \frac{y_2 - (n_2 - y_2) \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)} \right. \\ &\quad \left. - \frac{1}{2} \exp(\xi) \frac{n_1 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{2} \exp(\xi) \frac{n_2 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} \right) \\ &\quad - \exp(\xi) \frac{2 \exp(\xi) - \mu_\psi}{\sigma_\psi^2}. \end{aligned} \quad (12.35)$$

For  $l_-^*(\beta, \xi)$  the second-order partial derivatives are given by

$$\frac{\partial^2}{\partial\beta^2} l_-^*(\beta, \xi) = - \frac{n_1 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{\sigma_\beta^2}, \quad (12.36)$$

and

$$\frac{\partial^2}{\partial\beta\partial\xi} l_-^*(\beta, \xi) = - \frac{\exp(\xi)}{2} \left( \frac{n_1 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{n_2 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} \right), \quad (12.37)$$

and

$$\begin{aligned} \frac{\partial^2}{\partial\xi^2} l_-^*(\beta, \xi) &= - \frac{\exp(\xi)}{2} \left( \frac{(n_1 - y_1) \exp\left(\beta + \frac{\exp(\xi)}{2}\right) - y_1}{1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)} + \frac{y_2 - (n_2 - y_2) \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)} \right. \\ &\quad \left. - \frac{1}{2} \exp(\xi) \frac{n_1 \exp\left(\beta + \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta + \frac{\exp(\xi)}{2}\right)\right)^2} - \frac{1}{2} \exp(\xi) \frac{n_2 \exp\left(\beta - \frac{\exp(\xi)}{2}\right)}{\left(1 + \exp\left(\beta - \frac{\exp(\xi)}{2}\right)\right)^2} \right) \\ &\quad + \exp(\xi) \frac{2 \exp(\xi) - \mu_\psi}{\sigma_\psi^2}. \end{aligned} \quad (12.38)$$

### 12.C.3 Hessians

Having derived the relevant second-order partial derivatives, we can simply build the Hessian matrices of interest by inserting the relevant expressions. Next, we present symbolically the Hessians of interest, that is, we show which of the second-order partial derivatives need to be inserted where. Note that we omit the one

for  $\mathcal{H}_0$  since this is a single number which is simply the second-order derivative of  $l_0^*(\beta)$ .

The Hessian for  $\mathcal{H}_1$  is given by:

$$\mathbf{H}_1 = \begin{pmatrix} \frac{\partial^2}{\partial \beta^2} l^*(\beta, \psi) & \frac{\partial^2}{\partial \beta \partial \psi} l^*(\beta, \psi) \\ \frac{\partial^2}{\partial \beta \partial \psi} l^*(\beta, \psi) & \frac{\partial^2}{\partial \psi^2} l^*(\beta, \psi) \end{pmatrix}. \quad (12.39)$$

The Hessian for  $\mathcal{H}_+$  is given by:

$$\mathbf{H}_+ = \begin{pmatrix} \frac{\partial^2}{\partial \beta^2} l_+^*(\beta, \xi) & \frac{\partial^2}{\partial \beta \partial \xi} l_+^*(\beta, \xi) \\ \frac{\partial^2}{\partial \beta \partial \xi} l_+^*(\beta, \xi) & \frac{\partial^2}{\partial \xi^2} l_+^*(\beta, \xi) \end{pmatrix}. \quad (12.40)$$

The Hessian for  $\mathcal{H}_-$  is given by:

$$\mathbf{H}_- = \begin{pmatrix} \frac{\partial^2}{\partial \beta^2} l_-^*(\beta, \xi) & \frac{\partial^2}{\partial \beta \partial \xi} l_-^*(\beta, \xi) \\ \frac{\partial^2}{\partial \beta \partial \xi} l_-^*(\beta, \xi) & \frac{\partial^2}{\partial \xi^2} l_-^*(\beta, \xi) \end{pmatrix}. \quad (12.41)$$

### 12.C.3.1 Computing the Inverse of the Negative Hessians

Note that computing the inverses of the 2 by 2 negative Hessians is straightforward: We simply need to attach minus signs to each element of the Hessians and then make use of the fact that the inverse of a 2 by 2 matrix  $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is given by

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \text{ where } \det(\mathbf{A}) = ad - bc.$$

## 12.D Example: Effectiveness of Resilience Training (Default Analysis)

Here we present the results for the resilience training example obtained using the default prior setting.

### 12.D.1 Prior Specification

We use the default prior setting in the **abtest** package that assigns both  $\beta$  and  $\psi$  standard normal prior distributions. The implied prior on the absolute risk can be visualized as follows:

```
R> library("abtest")
R> plot_prior(what = "arisk")
```

The resulting graph is shown in the top panel of Figure 12.6. The user can also visualize the (implied) prior for other quantities. For instance, the prior on the log odds ratio (middle panel of Figure 12.6) is obtained as follows:

```
R> plot_prior(what = "logor")
```

The implied prior on the success probabilities  $p_1$  and  $p_2$  (bottom panel of Figure 12.6) is obtained as follows:

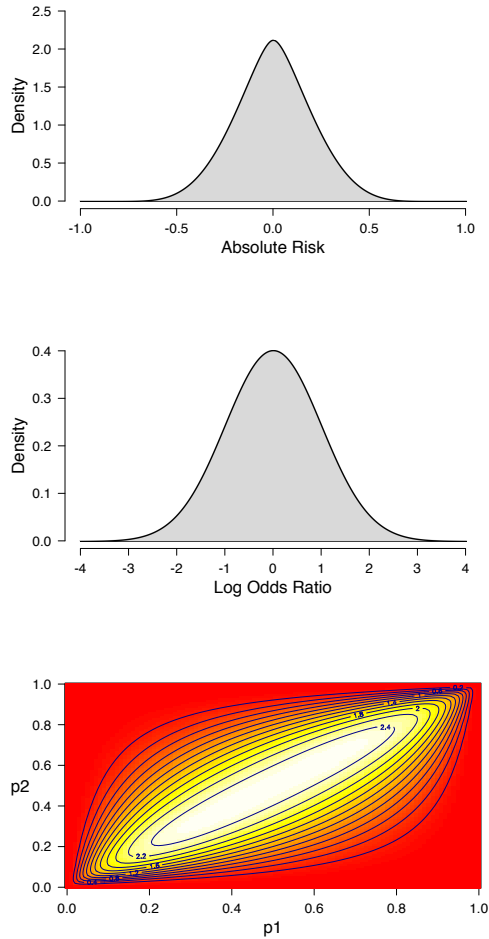


Figure 12.6: Default (implied) prior distributions. The top panel displays the prior distribution for the absolute risk which corresponds to the difference between the probability of still being on the job for the trained and the non-trained employees (i.e.,  $p_2 - p_1$ ). The middle panel shows the prior distribution for the log odds ratio parameter  $\psi$ . The bottom panel displays the implied joint prior distribution for the success probabilities  $p_1$  and  $p_2$ . The bottom panel illustrates that the two success probabilities are assigned dependent priors.

```
R> plot_prior(what = "p1p2")
```

The bottom panel of Figure 12.6 illustrates that there is a dependency between  $p_1$  and  $p_2$  which is arguably desirable (Howard, 1998): When one of the success probabilities is very (small) large, it is likely that the other one will also be (small) large.

### 12.D.2 Hypothesis Testing

The `ab.test` function can be used to conduct a Bayesian A/B test using the default prior setting as follows:

```
R> data("seqdata")
R> set.seed(1)
R> ab_default <- ab.test(data = seqdata)
```

This yields the following output:

```
R> print(ab_default)
```

Bayesian A/B Test Results:

Bayes Factors:

```
BF10: 0.2767214
BF+0: 0.4890489
BF-0: 0.05778357
```

Prior Probabilities Hypotheses:

```
H+: 0.25
H-: 0.25
H0: 0.5
```

Posterior Probabilities Hypotheses:

```
H+: 0.192
H-: 0.0227
H0: 0.7853
```

The first part of the output presents Bayes factors in favor of the hypotheses  $\mathcal{H}_1$ ,  $\mathcal{H}_+$ , and  $\mathcal{H}_-$ , where the reference hypothesis (i.e., denominator of the Bayes factor) is  $\mathcal{H}_0$ . Since all three Bayes factors are smaller than 1, they all indicate evidence in favor of the null hypothesis of no effect. The next part of the output displays the prior probabilities of the hypotheses with non-zero prior probability. The final part of the output displays the posterior probabilities of the hypotheses with non-zero prior probability. The posterior probability of the null hypothesis  $\mathcal{H}_0$  indicates that the data have increased the plausibility of the null hypothesis

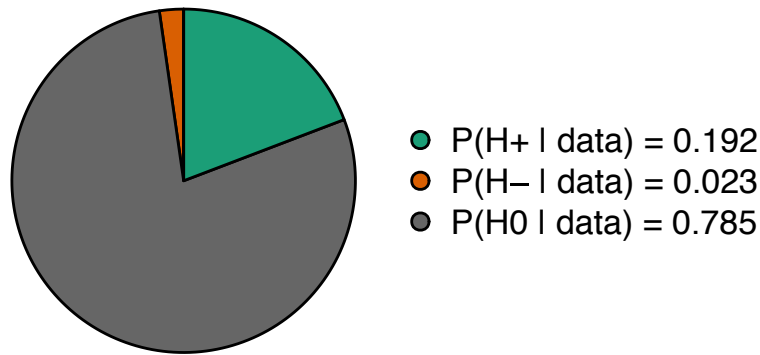


Figure 12.7: Posterior probabilities of the hypotheses visualized as a probability wheel.

from .50 to .79. Furthermore, the data have decreased the plausibility of both  $\mathcal{H}_+$  and  $\mathcal{H}_-$ .

The **abtest** package allows users to visualize the posterior probabilities of the hypotheses by means of a probability wheel (Figure 12.7):

```
R> prob_wheel(ab_default)
```

Overall, the data support the hypothesis that the training is ineffective over the hypothesis that the training has a positive effect. The Bayes factor for  $\mathcal{H}_0$  over  $\mathcal{H}_+$  equals  $1/0.489 \approx 2.04$ ; however, this indicates only anecdotal evidence (Jeffreys, 1939, Appendix I).

Since the data set is of a sequential nature, it may be of interest to consider not only the result based on all observations, but to conduct also a sequential analysis that tracks the evidential flow as a function of the total number of observations (i.e., the number of observations across both groups). This sequential analysis can be conducted as follows:

```
R> plot_sequential(ab_default, thin = 4)
```

Figure 12.8 displays the result of the sequential analysis. The sequential analysis indicates that after some initial fluctuation, adding more observations increased the probability of the null hypothesis that there is no effect of the training.

### 12.D.3 Parameter Estimation

The data indicate only anecdotal evidence in favor of the null hypothesis versus the hypothesis that the training is effective, leaving open the possibility that the training does have an effect. To assess this possibility one may investigate the potential size of the effect under the assumption that the effect is non-zero. For parameter estimation, we generally prefer to investigate the posterior distribution for the unconstrained alternative hypothesis  $\mathcal{H}_1$ .

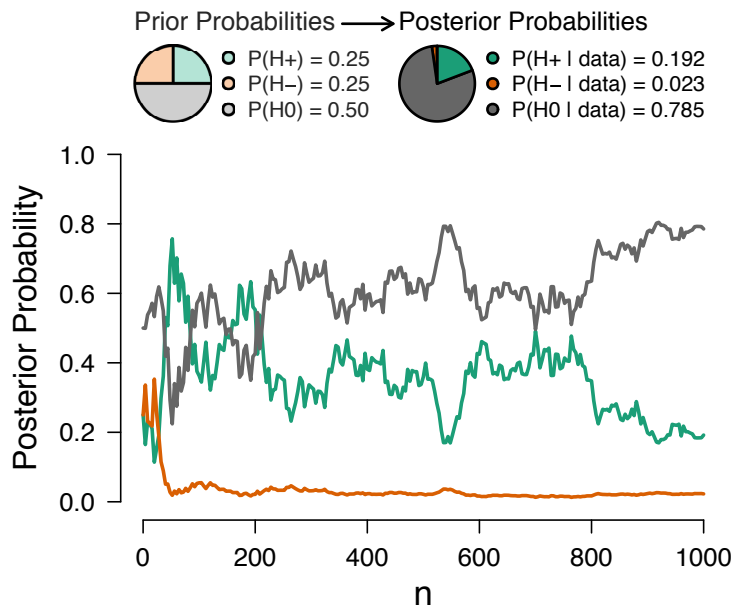


Figure 12.8: Sequential analysis results. The posterior probability of each hypothesis is plotted as a function of the number of observations across groups. On top, two probability wheels visualize the prior probabilities of the hypotheses and the posterior probabilities after taking into account all observations.

The top panel of Figure 12.9 displays the posterior distribution for the absolute risk (i.e.,  $p_2 - p_1$ ) that can be obtained as follows:

```
R> plot_posterior(ab_default, what = "arisk")
```

The top panel of Figure 12.9 shows the prior distribution as a dotted line and the posterior distribution (with 95% central credible interval) as a solid line. The plot indicates that, under the assumption that the difference between the two success probabilities is not exactly zero, the posterior median is 0.039 and the 95% central credible interval ranges from  $-0.022$  to  $0.101$ .

The middle panel of Figure 12.9 displays the posterior distribution for the log odds ratio  $\psi$  that can be obtained as follows:

```
R> plot_posterior(ab_default, what = "logor")
```

The middle panel of Figure 12.9 indicates that, given the log odds ratio is not exactly zero, it is likely to be between  $-0.089$  and  $0.406$ , where the posterior median is 0.159.

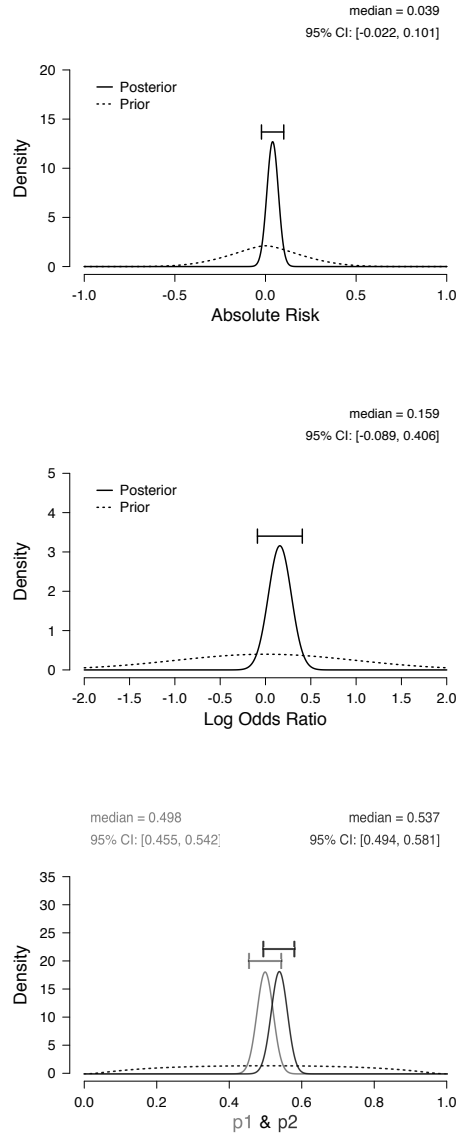


Figure 12.9: (Implied) prior and posterior distributions under  $\mathcal{H}_1$ . The dotted lines display the prior distributions, the solid lines display the posterior distributions (with 95% central credible intervals). The medians and the bounds of the 95% central credible intervals are displayed on top of each panel. The top panel displays the posterior distribution for the absolute risk (i.e.,  $p_2 - p_1$ ); the middle panel shows the posterior distribution for the log odds ratio parameter  $\psi$ ; the bottom panel displays the marginal posterior distributions for the success probabilities  $p_1$  and  $p_2$ .

It may also be of interest to consider the marginal posterior distributions of the success probabilities  $p_1$  and  $p_2$ . This plot can be produced as follows:

```
R> plot_posterior(ab_default, what = "p1p2")
```

The bottom panel of Figure 12.9 displays the resulting plot. In this example,  $p_1$  and  $p_2$  correspond to the probability of still being on the job after six month for the non-trained employees and the employees that received the training, respectively. The bottom panel of Figure 12.9 indicates that the posterior median for  $p_1$  is 0.498, with 95% credible ranging from 0.455 to 0.542, and the posterior median for  $p_2$  is 0.537, with 95% credible interval ranging from 0.494 to 0.581.

In sum, based on a default prior analysis, this synthetic data set offers anecdotal evidence in favor of the null hypothesis which states that the training is not effective over the hypothesis that the training is effective; the consultancy firm should probably continue to collect data in order to obtain more compelling evidence before deciding whether or not the training should be implemented. If the true effect is as small as 4%, continued testing will ultimately show compelling evidence for  $\mathcal{H}_+$  over  $\mathcal{H}_0$ . Note that continued testing is trivial in the Bayesian framework: the results can simply be updated as new observations arrive.

## 12.E Progesterone in Women with Bleeding in Early Pregnancy: Absence of Evidence, Not Evidence of Absence

As an example application of the **abtest** package, here we present the results of a reanalysis of a recent medical trial.<sup>11</sup>

A recent trial assessed the effectiveness of progesterone in preventing miscarriages (Coomarasamy et al., 2019). The number of live births was 74.7% (1513/2025) in the progesterone group and 72.5% (1459/2013) in the placebo group ( $p = .08$ ). The authors concluded: “The incidence of adverse events did not differ significantly between the groups.”

This conclusion leaves unaddressed the degree to which the data undercut or support the progesterone hypothesis. To quantify such evidence we conducted Bayesian logistic regression (Gronau, Raj K. N., & Wagenmakers, 2019; Kass & Vaidyanathan, 1992). Under the no-effect model  $\mathcal{H}_0$ , the log odds ratio equals  $\psi = 0$ , whereas under the positive-effect model  $\mathcal{H}_+$ ,  $\psi$  is assigned a positive-only normal prior  $\mathcal{N}_+(\mu, \sigma^2)$ . A default analysis (i.e.,  $\mu = 0$ ,  $\sigma = 1$ ) reveals only weak evidence for  $\mathcal{H}_0$  (Jeffreys, 1939). Figure 12.10 shows the evidence is weak for all combinations of  $\mu$  in  $[0, 0.30]$  and  $\sigma$  in  $[0.25, 1]$ .

In sum, these data neither undercut nor support the progesterone hypothesis in compelling fashion.

---

<sup>11</sup>This reanalysis is available on *PsyArXiv*: Gronau, Q. F., & Wagenmakers, E.-J. (2019). Progesterone in women with bleeding in early pregnancy: Absence of evidence, not evidence of absence. <https://psyarxiv.com/etk7g/>

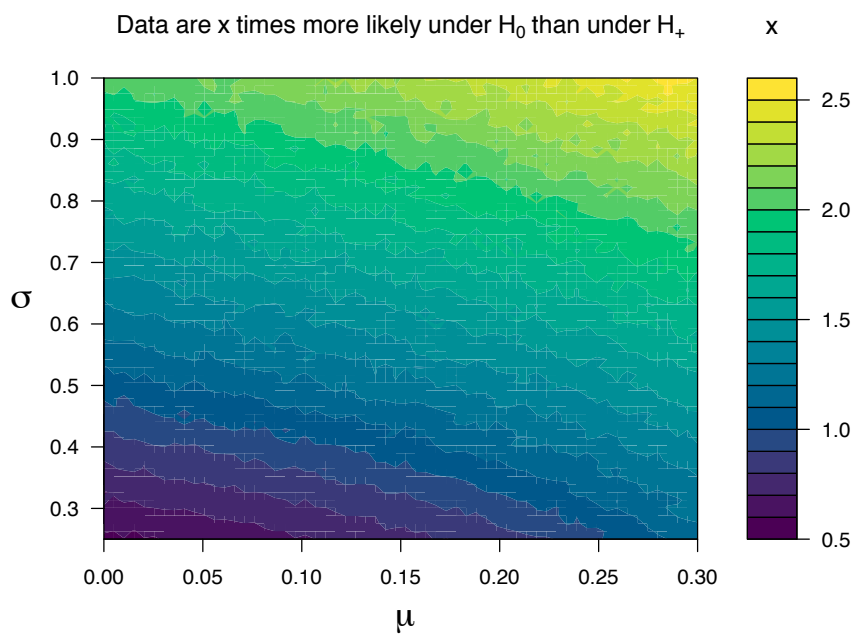


Figure 12.10: Across different priors, the evidence for no-effect  $\mathcal{H}_0$  over positive-effect  $\mathcal{H}_+$  is weak.



---

# Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection

---

## Abstract

Cross-validation (CV) is increasingly popular as a generic method to adjudicate between mathematical models of cognition and behavior. In order to measure model generalizability, CV quantifies out-of-sample predictive performance, and the CV preference goes to the model that predicted the out-of-sample data best. The advantages of CV include theoretic simplicity and practical feasibility. Despite its prominence, however, the limitations of CV are often underappreciated. Here we demonstrate the limitations of a particular form of CV – Bayesian leave-one-out cross-validation or LOO – with three concrete examples. In each example, a data set of infinite size is perfectly in line with the predictions of a simple model (i.e., a general law or invariance). Nevertheless, LOO shows bounded and relatively modest support for the simple model. We conclude that CV is not a panacea for model selection.

[...] if you can't do simple problems, how can you do complicated ones?

---

Dennis Lindley (1985, p. 65)

---

This chapter is published as Gronau, Q. F., & Wagenmakers, E.-J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior*, 2, 1–11. doi: <https://doi.org/10.1007/s42113-018-0011-7>. Also available as *PsyArXiv preprint*: <https://psyarxiv.com/at7cx/>

### 13.1 Introduction

Model selection is a perennial problem, both in mathematical psychology (e.g., the three special issues for the *Journal of Mathematical Psychology*: J. Mulder & Wagenmakers, 2016; Myung, Forster, & Browne, 2000b; Wagenmakers & Waldorp, 2006b) and in statistics (e.g., Ando, 2010; Burnham & Anderson, 2002; Claeskens & Hjort, 2008; Grünwald, Myung, & Pitt, 2005; Wrinch & Jeffreys, 1921). The main challenge for model selection is known both as the bias-variance tradeoff and as the parsimony-fit tradeoff (e.g., Myung, 2000; Myung & Pitt, 1997). These tradeoffs form the basis of what may be called the *fundamental law of model selection*: when the goal is to assess a model's predictive performance, goodness-of-fit ought to be discounted by model complexity. For instance, consider the comparison between two regression models,  $\mathcal{M}_S$  and  $\mathcal{M}_C$ ; the 'simple' model  $\mathcal{M}_S$  has  $k$  predictors, whereas the 'complex' model  $\mathcal{M}_C$  has  $l$  predictors more, for a total of  $k + l$ . Hence,  $\mathcal{M}_S$  is said to be nested under  $\mathcal{M}_C$ . In such cases,  $\mathcal{M}_C$  always outperforms  $\mathcal{M}_S$  in terms of goodness-of-fit (e.g., variance explained), even when the  $l$  extra predictors are useless in the sense that they capture only the idiosyncratic, nonreplicable noise in the sample at hand. Consequently, model selection methods that violate the fundamental law trivially fail, because they prefer the most complex model regardless of the data.

All popular methods of model selection adhere to the fundamental law in that they seek to chart a route that avoids the Scylla of 'overfitting' (i.e., overweighting goodness-of-fit such that complex models receive an undue preference) and the Charybdis of 'underfitting' (i.e., overweighting parsimony such that simple models receive an undue preference). Both Scylla and Charybdis result in the selection of models with poor predictive performance; models that fall prey to Scylla mistake what is idiosyncratic noise in the sample for replicable signal, leading to excess variability in the parameter estimates; in contrast, models that fall prey to Charybdis mistake what is replicable signal for idiosyncratic noise, leading to bias in the parameter estimates. Both excess variability and bias result in suboptimal predictions, that is, poor generalizability.

The cornucopia of model selection methods includes (1) approximate methods such as AIC (Akaike, 1973) and BIC (Nathoo & Masson, 2016; Schwarz, 1978), which punish complexity by an additive term that includes the number of free parameters; (2) methods that quantify predictive performance by averaging goodness-of-fit across the model's entire parameter space (i.e., the Bayes factor, e.g., Jeffreys, 1961; Kass & Raftery, 1995; Ly et al., 2016b; Rouder et al., 2012); note that the averaging process indirectly penalizes complexity, as a vast parameter space will generally contain large swathes that produce a poor fit (Vandekerckhove et al., 2015); (3) methods based on minimum description length (Grünwald, 2007; Myung, Navarro, & Pitt, 2006; Rissanen, 2007), where the goal is the efficient transmission of information, that is, a model and the data it encodes; complex models take more bits to describe and transmit; (4) methods such as cross-validation (CV; Browne, 2000; M. Stone, 1974) that assess predictive performance directly, namely by separating the data in a part that is used for fitting (i.e., the calibration set or training set) and a part that is used to assess predictive adequacy (i.e., the validation set or test set).

Each model selection method comes with its own set of assumptions and operating characteristics which may or may not be appropriate for the application at hand. For instance, AIC and BIC assume that model complexity can be approximated by counting the number of free parameters, and the Bayes factor presupposes the availability of a reasonable joint prior distribution across the parameter space (Lee & Vanpaemel, 2018). The focus of the current chapter is on CV, an increasingly popular and generic model selection procedure (e.g., Doxas, Dennis, & Oliver, 2010; Hastie, Tibshirani, Friedman, & Vetterling, 2008; Yarkoni & Westfall, 2017). Specifically, our investigation concerns leave-one-out CV, where the model is trained on all observations except one, which then forms the test set. The procedure is repeated for all  $n$  observations, and the overall predictive CV performance is the sum of the predictive scores for each of the  $n$  test sets.

Originally developed within a frequentist framework, leave-one-out CV can also be executed within a Bayesian framework; in the Bayesian framework, the predictions for the test sets are based not on a point estimate but on the entire posterior distribution (Geisser & Eddy, 1979; Gelfand, Dey, & Chang, 1992; see also Geisser, 1975). Henceforth we will refer to this Bayesian version of leave-one-out CV as LOO (e.g., Gelman, Hwang, & Vehtari, 2014; Vehtari et al., 2017; Vehtari & Ojanen, 2012).<sup>1</sup>

To foreshadow our conclusion, we demonstrate below with three concrete examples how LOO can yield conclusions that appear undesirable; specifically, in the idealized case where there exist a data set of infinite size that is perfectly consistent with the simple model  $\mathcal{M}_S$ , LOO will nevertheless fail to strongly endorse  $\mathcal{M}_S$ . It has long been known that CV has this property, termed “inconsistency” (e.g., Shao, 1993).<sup>2</sup> Our examples demonstrate not just that CV is inconsistent, but also serve to explicate the reason for the inconsistency. Moreover, the examples show not only that CV is inconsistent, that is, the support for the true  $\mathcal{M}_S$  does not increase without bound,<sup>3</sup> but they also show that the degree of the support for the true  $\mathcal{M}_S$  is surprisingly modest. One of our examples also reveals that, in contrast to what is commonly assumed, the results for LOO can depend strongly on the prior distribution, even asymptotically; finally, in all three examples the observation of data perfectly consistent with  $\mathcal{M}_S$  may nevertheless cause LOO to decrease its preference for  $\mathcal{M}_S$ . Before we turn to the three examples we first introduce LOO in more detail.

<sup>1</sup>The LOO functionality is available through the R package “loo” (Vehtari, Gabry, Yao, & Gelman, 2018), see also <http://mc-stan.org/loo/>.

<sup>2</sup>“[...] it is known to many statisticians (although a rigorous statement has probably not been given in the literature) that the cross-validation with  $n_v \equiv 1$  is asymptotically incorrect (inconsistent) and is too conservative in the sense that it tends to select an unnecessarily large model” (Shao, 1993, p. 486).

<sup>3</sup>The authors agree with Bayarri et al. (2012, p. 1553) who argued that “[...] it would be philosophically troubling to be in a situation with infinite data generated from one of the models being considered, and not choosing the correct model.”

### 13.2 Bayesian Leave-One-Out Cross-Validation

The general principle of cross-validation is to partition a data set consisting of  $n$  observations  $y_1, y_2, \dots, y_n$  into a training set and a test set. The training set is used to fit the model and the test set is used to evaluate the fitted model's predictive adequacy. LOO repeatedly partitions the data set into a training set which consists of all data points except the  $i$ th one, denoted as  $y_{-i}$ , and then evaluates the predictive density for the held-out data point  $y_i$ . The log of these predictive densities for all data points is summed to obtain the LOO estimate of the expected log pointwise predictive density (elpd; Gelman, Hwang, & Vehtari, 2014; Vehtari et al., 2017):<sup>4</sup>

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^n \log p(y_i | y_{-i}), \quad (13.1)$$

where

$$p(y_i | y_{-i}) = \int p(y_i | \theta) p(\theta | y_{-i}) d\theta \quad (13.2)$$

is the leave-one-out predictive density for data point  $y_i$  given the remaining data points  $y_{-i}$  and  $\theta$  denotes the model parameters.

It is insightful to note the close connection of LOO to what Gelfand and Dey (1994) called the *pseudo-Bayes factor* (PSBF) which they attribute to Geisser and Eddy (1979). Recall that the Bayes factor that compares models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  (Kass & Raftery, 1995) is defined as:

$$\text{BF}_{12} = \frac{p(y | \mathcal{M}_1)}{p(y | \mathcal{M}_2)}, \quad (13.3)$$

where  $y = (y_1, y_2, \dots, y_n)$  and  $p(y | \mathcal{M}_m) = \int_{\Theta_m} p(y | \theta_m, \mathcal{M}_m) p(\theta_m | \mathcal{M}_m) d\theta_m$  denotes the marginal likelihood of model  $\mathcal{M}_m$ ,  $m \in \{1, 2\}$ . The pseudo-Bayes factor PSBF replaces the marginal likelihood of each model by the product of the leave-one-out predictive densities so that:

$$\begin{aligned} \text{PSBF}_{12} &= \frac{\prod_{i=1}^n p(y_i | y_{-i}, \mathcal{M}_1)}{\prod_{i=1}^n p(y_i | y_{-i}, \mathcal{M}_2)} \\ &= \exp \left\{ \Delta \text{elpd}_{\text{loo}}^{\mathcal{M}_1, \mathcal{M}_2} \right\}, \end{aligned} \quad (13.4)$$

where  $\Delta \text{elpd}_{\text{loo}}^{\mathcal{M}_1, \mathcal{M}_2} = \text{elpd}_{\text{loo}}^{\mathcal{M}_1} - \text{elpd}_{\text{loo}}^{\mathcal{M}_2}$  and  $\text{elpd}_{\text{loo}}^{\mathcal{M}_m}$  denotes the LOO estimate for model  $\mathcal{M}_m$ ,  $m \in \{1, 2\}$ . It is also worth mentioning that LOO can be used to compute model weights (e.g., Yao, Vehtari, Simpson, & Gelman, 2018; see also Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004) as follows:

$$w_m = \frac{\exp \left\{ \text{elpd}_{\text{loo}}^{\mathcal{M}_m} \right\}}{\sum_{j=1}^M \exp \left\{ \text{elpd}_{\text{loo}}^{\mathcal{M}_j} \right\}}, \quad (13.5)$$

---

<sup>4</sup>Note that the following expressions are conditional on a specific model. However, we have omitted conditioning on the model for enhanced legibility.

where  $w_m$  denotes the model weight for model  $\mathcal{M}_m$  and  $M$  is the number of models under consideration. The LOO results from the three examples below will be primarily presented as weights.

### 13.3 Example 1: Induction

As a first example, we consider what is perhaps the world’s oldest inference problem, one that has occupied philosophers for over two millennia: given a general law such as “all X’s have property Y”, how does the accumulation of confirmatory instances (i.e., X’s that indeed have property Y) increase our confidence in the general law? Examples of such general laws include “all ravens are black”, “all apples grow on apple trees”, “all neutral atoms have the same number of protons and electrons”, and “all children with Down syndrome have all or part of a third copy of chromosome 21”.

To address this question statistically we can compare two models (e.g., Etz & Wagenmakers, 2017; Wrinch & Jeffreys, 1921). The first model corresponds to the general law and can be conceptualized as  $\mathcal{H}_0 : \theta = 1$ , where  $\theta$  is a Bernoulli probability parameter. This model predicts that only confirmatory instances are encountered. The second model relaxes the general law and is therefore more complex; it assigns  $\theta$  a prior distribution, which, for mathematical convenience, we take to be from the beta family – consequently, we have  $\mathcal{H}_1 : \theta \sim \text{Beta}(a, b)$ .

In the following, we assume that, in line with the prediction from  $\mathcal{H}_0$ , only confirmatory instances are observed. In such a scenario, we submit that there are at least three desiderata for model selection. First, for any sample size  $n > 0$  of confirmatory instances, the data ought to support the general law  $\mathcal{H}_0$ ; second, as  $n$  increases, so should the level of support in favor of  $\mathcal{H}_0$ ; third, as  $n$  increases without bound, the support in favor of  $\mathcal{H}_0$  should grow infinitely large.

How does LOO perform in this scenario? Before proceeding, note that when LOO makes predictions based on the maximum likelihood estimate (MLE), none of the above desiderata are fulfilled. Any training set of size  $n - 1$  will contain  $k = n - 1$  confirmatory instances, such that the MLE under  $\mathcal{H}_1$  is  $\hat{\theta} = k/(n - 1) = 1$ ; of course, the general law  $\mathcal{H}_0$  does not contain any adjustable parameters and simply stipulates that  $\theta = 1$ . When the models’ predictive performance is evaluated for the test set observation, it then transpires that both  $\mathcal{H}_0$  and  $\mathcal{H}_1$  have  $\theta$  set to 1 ( $\mathcal{H}_0$  on principle,  $\mathcal{H}_1$  by virtue of having seen the  $n - 1$  confirmatory instances from the training set), so that they make identical predictions. Consequently, according to the maximum likelihood version of LOO, the data are completely uninformative, no matter how many confirmatory instances are observed.<sup>5</sup>

The Bayesian LOO makes predictions using the leave-one-out posterior distribution for  $\theta$  under  $\mathcal{H}_1$ , and this means that it at least fulfills the first desideratum: the prediction under  $\mathcal{H}_0 : \theta = 1$  is perfect, whereas the prediction under  $\mathcal{H}_1 : \theta \sim \text{Beta}(a + n - 1, b)$  involves values of  $\theta$  that do not make such perfect predictions. As a result, the Bayesian LOO will show that the general law  $\mathcal{H}_0$  outpredicts  $\mathcal{H}_1$  for the test set.

<sup>5</sup>This holds for  $k$ -fold CV in general.

What happens when sample size  $n$  grows large? Intuitively, two forces are in opposition: on the one hand, as  $n$  grows large, the leave-one-out posterior distribution of  $\theta$  under the complex model  $\mathcal{H}_1$  will be increasingly concentrated near 1, generating predictions for the test set data that are increasingly similar to those made by  $\mathcal{H}_0$ . On the other hand, even with  $n$  large, the predictions from  $\mathcal{H}_1$  will still be inferior to those from  $\mathcal{H}_0$ , and these inferior predictions are multiplied by  $n$ , the number of test sets.

As it turns out, these two forces are asymptotically in balance, so that the level of support in favor of  $\mathcal{H}_0$  approaches a bound as  $n$  grows large. We first provide the mathematical result and then show the outcome for a few select scenarios.

### 13.3.1 Mathematical Result

In Example 1 the data consist of  $n$  realizations drawn from a Bernoulli distribution, denoted by  $y_i$ ,  $i = 1, 2, \dots, n$ . Under  $\mathcal{H}_0$ , the success probability  $\theta$  is fixed to 1 and under  $\mathcal{H}_1$ ,  $\theta$  is assigned a  $\text{Beta}(a, b)$  prior. We consider the case where only successes are observed, that is,  $y_i = 1, \forall i \in \{1, 2, \dots, n\}$ . The model corresponding to  $\mathcal{H}_0 : \theta = 1$  has no free parameters and predicts  $y_i = 1$  with probability one. Therefore, the Bayesian LOO estimate  $\text{elpd}_{\text{loo}}^{\mathcal{H}_0}$  is equal to 0. To calculate the LOO estimate under  $\mathcal{H}_1$ , one needs to be able to evaluate the predictive density for a single data point given the remaining data points. Recall that the posterior based on  $n - 1$  observations is a  $\text{Beta}(a + n - 1, b)$  distribution. Consequently, the leave-one-out predictive density is obtained as a generalization (with  $a$  and  $b$  potentially different from 1) of Laplace's rule of succession applied to  $n - 1$  observations,

$$\begin{aligned} p(y_i | y_{-i}) &= \int_0^1 \underbrace{\theta}_{p(y_i|\theta)} \underbrace{\frac{\Gamma(a+n-1+b)}{\Gamma(a+n-1)\Gamma(b)} \theta^{a+n-2} (1-\theta)^{b-1}}_{p(\theta|y_{-i})} d\theta \\ &= \frac{a + n - 1}{a + n - 1 + b}, \end{aligned} \quad (13.6)$$

and the Bayesian LOO estimate under  $\mathcal{H}_1$  is given by

$$\text{elpd}_{\text{loo}}^{\mathcal{H}_1} = n \log \left( \frac{a + n - 1}{a + n - 1 + b} \right). \quad (13.7)$$

The difference in the LOO estimates is

$$\begin{aligned} \Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} &= \text{elpd}_{\text{loo}}^{\mathcal{H}_0} - \text{elpd}_{\text{loo}}^{\mathcal{H}_1} \\ &= -n \log \left( \frac{a + n - 1}{a + n - 1 + b} \right). \end{aligned} \quad (13.8)$$

As the number of confirmatory instances  $n$  grows large, the difference in the LOO estimates approaches a bound (see Appendix A for a derivation):

$$\lim_{n \rightarrow \infty} \Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = b. \quad (13.9)$$

Hence, the asymptotic difference in the Bayesian LOO estimates under  $\mathcal{H}_0$  and under  $\mathcal{H}_1$  equals the Beta prior parameter  $b$ . Consequently, the limit of the pseudo-Bayes factor is

$$\lim_{n \rightarrow \infty} \text{PSBF}_{01} = \exp \{b\}, \quad (13.10)$$

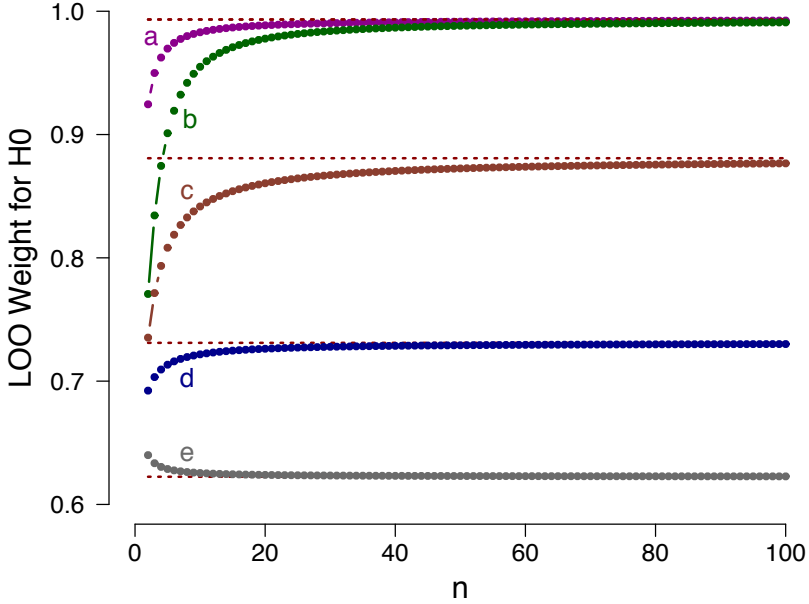


Figure 13.1: Example 1: LOO weights for  $\mathcal{H}_0 : \theta = 1$  as a function of the number of confirmatory instances  $n$ , evaluated in relation to five different prior specifications for  $\mathcal{H}_1$ : (a)  $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 5)$ ; (b)  $\mathcal{H}_1 : \theta \sim \text{Beta}(5, 5)$ ; (c)  $\mathcal{H}_1 : \theta \sim \text{Beta}(2, 2)$ ; (d)  $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 1)$ ; (e)  $\mathcal{H}_1 : \theta \sim \text{Beta}(0.5, 0.5)$ . The dotted horizontal lines indicate the corresponding analytical asymptotic bounds. See text for details. Available at <https://tinyurl.com/ya2r4gx8> under CC license <https://creativecommons.org/licenses/by/2.0/>.

and the limit of the model weight for  $\mathcal{H}_0$  is

$$\lim_{n \rightarrow \infty} w_0 = \frac{\exp\{b\}}{1 + \exp\{b\}}. \quad (13.11)$$

### 13.3.2 Select Scenarios

The mathematical result can be applied to a series of select scenarios. Figure 13.1 shows the LOO weight in favor of the general law  $\mathcal{H}_0$  as a function of the number of confirmatory instances  $n$ , separately for five different prior specifications under  $\mathcal{H}_1$ . The figure confirms that for each prior specification, the LOO weight for  $\mathcal{H}_0$  approaches its asymptotic bound as  $n$  grows large.

We conclude the following: (1) as  $n$  grows large, the support for the general law  $\mathcal{H}_0$  approaches a bound; (2) for many common prior distributions, this bound is surprisingly low. For instance, the Laplace prior  $\theta \sim \text{Beta}(1, 1)$  (case d) yields

a weight of  $e/(1+e) \approx 0.731$ ; (3) contrary to popular belief, our results provide an example of a situation in which the results from LOO are highly dependent on the prior distribution, even asymptotically. This is clear from Equation 13.11 and evidenced in Figure 13.1; (4) as shown by case (e) in Figure 13.1, the choice of Jeffreys's prior (i.e.,  $\theta \sim \text{Beta}(0.5, 0.5)$ ) results in a function that approaches the asymptote from above. This means that, according to LOO, the observation of additional confirmatory instances actually decreases the support for the general law, violating the second desideratum outlined above. This violation can be explained by the fact that the confirmatory instances help the complex model  $\mathcal{H}_1$  concentrate more mass near 1, thereby better mimicking the predictions from the simple model  $\mathcal{H}_0$ . For some prior choices, this increased ability to mimic outweighs the fact that the additional confirmatory instances are better predicted by  $\mathcal{H}_0$  than by  $\mathcal{H}_1$ .

One counterargument to this demonstration could be that, despite its venerable history, the case of induction is somewhat idiosyncratic, having to do more with logic than with statistics. To rebut this argument we present two additional examples.

## 13.4 Example 2: Chance

As a second example, we consider the case where the general law states that the Bernoulli probability parameter  $\theta$  equals  $1/2$  rather than 1. Processes that may be guided by such a law include “the probability that a randomly chosen digit from the decimal expansion of  $\pi$  is odd rather than even” (Gronau & Wagenmakers, 2018), “the probability that a particular Uranium-238 atom will decay in the next 4.5 billion years”, or “the probability that an extrovert participant in an experiment on extra-sensory perception correctly predicts whether an erotic picture will appear on the right or on the left side of a computer screen” (Bem, 2011).

Hence, the general law holds that  $\mathcal{H}_0 : \theta = 1/2$ , and the model that relaxes that law is given by  $\mathcal{H}_1 : \theta \sim \text{Beta}(a, b)$ , as in Example 1. Also, similar to Example 1, we consider the situation where the observed data are perfectly consistent with the predictions from  $\mathcal{H}_0$ . To accomplish this, we consider only even sample sizes  $n$  and set the number of successes  $k$  equal to  $n/2$ . In other words, the binary data come as pairs, where one member is a success and the other is a failure. The general desiderata are similar to those from Example 1: First, for any sample size with  $k = n/2$  successes, the data ought to support the general law  $\mathcal{H}_0$ ; second, as  $n$  increases (for  $n$  even and with  $k = n/2$  successes), so should the level of support in favor of  $\mathcal{H}_0$ ; third, as  $n$  increases without bound, the support in favor of  $\mathcal{H}_0$  should grow infinity large.

### 13.4.1 Mathematical Result

In Example 2 the data consist again of  $n$  realizations drawn from a Bernoulli distribution, denoted by  $y_i$ ,  $i = 1, 2, \dots, n$ . Under  $\mathcal{H}_0$ , the success probability  $\theta$  is now fixed to  $1/2$ ; under  $\mathcal{H}_1$ ,  $\theta$  is again assigned a  $\text{Beta}(a, b)$  prior. The model corresponding to  $\mathcal{H}_0 : \theta = 1/2$  has no free parameters and predicts  $y_i = 0$  with

probability  $1/2$  and  $y_i = 1$  with probability  $1/2$ . Therefore, the LOO estimate is given by  $\text{elpd}_{\text{loo}}^{\mathcal{H}_0} = -n \log(2)$ . To calculate the LOO estimate under  $\mathcal{H}_1$ , one needs to be able to evaluate the predictive density for a single data point given the remaining data points. Recall that the posterior based on  $n - 1$  observations is a  $\text{Beta}(a + k_{-i}, b + n - 1 - k_{-i})$  distribution, where  $k_{-i} = \sum_{j \neq i} y_j$  denotes the number of successes based on all data points except the  $i$ th one. Consequently, the leave-one-out predictive density is given by:

$$\begin{aligned} p(y_i | y_{-i}) &= \int_0^1 \underbrace{\theta^{y_i} (1 - \theta)^{1-y_i}}_{p(y_i | \theta)} \times \underbrace{\frac{\Gamma(a+b+n-1)}{\Gamma(a+k_{-i})\Gamma(b+n-k_{-i}-1)} \theta^{a+k_{-i}-1} (1 - \theta)^{b+n-k_{-i}-2}}_{p(\theta | y_{-i})} d\theta \\ &= \begin{cases} \frac{a+k-1}{a+b+n-1} & \text{if } y_i = 1 \\ \frac{b+n-k-1}{a+b+n-1} & \text{if } y_i = 0, \end{cases} \end{aligned} \quad (13.12)$$

where  $k = \sum_{i=1}^n y_i$  denotes the total number of successes. Example 2 considers the case where  $n$  is even and the number of successes  $k$  equals  $\frac{n}{2}$ . The Bayesian LOO estimate under  $\mathcal{H}_1$  is then given by:

$$\text{elpd}_{\text{loo}}^{\mathcal{H}_1} = \frac{n}{2} \log \left( \frac{a + \frac{n}{2} - 1}{a + b + n - 1} \right) + \frac{n}{2} \log \left( \frac{b + \frac{n}{2} - 1}{a + b + n - 1} \right). \quad (13.13)$$

The difference in the LOO estimates can be written as

$$\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = \frac{n}{2} \log \left( \frac{a + b + n - 1}{2a + n - 2} \right) + \frac{n}{2} \log \left( \frac{a + b + n - 1}{2b + n - 2} \right). \quad (13.14)$$

As the even sample size  $n$  grows large, the difference in the LOO estimates approaches a bound (see Appendix B for a derivation):

$$\lim_{n \rightarrow \infty} \Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = 1. \quad (13.15)$$

Consequently, the limit of the pseudo-Bayes factor is

$$\lim_{n \rightarrow \infty} \text{PSBF}_{01} = e \approx 2.718, \quad (13.16)$$

and the limit of the model weight for  $\mathcal{H}_0$  is

$$\lim_{n \rightarrow \infty} w_0 = \frac{e}{1 + e} \approx 0.731. \quad (13.17)$$

### 13.4.2 Select Scenarios

The mathematical result can be applied to a series of select scenarios, as before. Figure 13.2 shows the LOO weight in favor of the general law  $\mathcal{H}_0$  as a function of the even number of observations  $n$ , separately for five different prior specifications

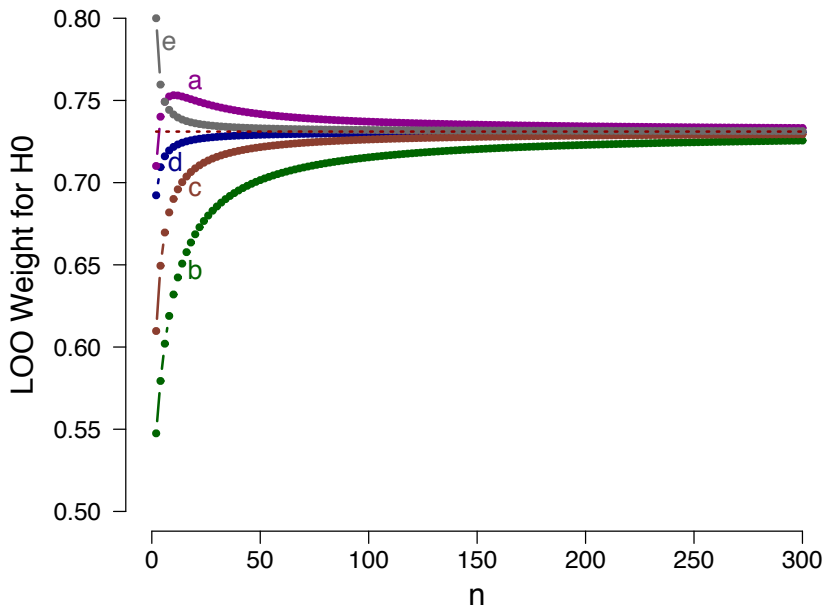


Figure 13.2: Example 2: LOO weights for  $\mathcal{H}_0 : \theta = 1/2$  as a function of the number of observations  $n$ , where the number of successes  $k = n/2$ , evaluated in relation to five different prior specifications for  $\mathcal{H}_1$ : (a)  $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 5)$ ; (b)  $\mathcal{H}_1 : \theta \sim \text{Beta}(5, 5)$ ; (c)  $\mathcal{H}_1 : \theta \sim \text{Beta}(2, 2)$ ; (d)  $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 1)$ ; (e)  $\mathcal{H}_1 : \theta \sim \text{Beta}(0.5, 0.5)$ . The dotted horizontal line indicates the corresponding analytical asymptotic bound. Note that only even sample sizes are displayed. See text for details. Available at <https://tinyurl.com/y8azu4hc> under CC license <https://creativecommons.org/licenses/by/2.0/>.

under  $\mathcal{H}_1$ . The figure confirms that for each prior specification, the LOO weight for  $\mathcal{H}_0$  approaches its asymptotic bound as  $n$  grows large.

We conclude the following: (1) as  $n$  grows large, the support for the general law  $\mathcal{H}_0$  approaches a bound; (2) in contrast to Example 1, this bound is independent of the particular choice of Beta prior distribution for  $\theta$  under  $\mathcal{H}_1$ ; however, consistent with Example 1, this bound is surprisingly low. Even with an infinite number of observations, exactly half of which are successes and half of which are failures, the model weight for the general law  $\mathcal{H}_0$  does not exceed a modest 0.731; (3) as shown by case (e) in Figure 13.2, the choice of Jeffreys's prior (i.e.,  $\theta \sim \text{Beta}(0.5, 0.5)$ ) results in a function that approaches the asymptote from above. This means that, according to LOO, the observation of additional success-failure pairs actually decreases the support for the general law, violating the second desideratum outlined above; (4) as shown by case (a) in Figure 13.2, the choice of a Beta(1, 5)

prior results in a nonmonotonic relation, where the addition of  $\mathcal{H}_0$ -consistent pairs initially increases the support for  $\mathcal{H}_0$ , and later decreases it.

In sum, the result of the LOO procedure for a test against a chance process,  $\mathcal{H}_0 : \theta = 1/2$ , reveals behavior that is broadly similar to that for the test of induction ( $\mathcal{H}_0 : \theta = 0$  or  $\mathcal{H}_0 : \theta = 1$ ), and that violates two seemingly uncontroversial desiderata, namely that the additional observation of data that are perfectly consistent with the general law  $\mathcal{H}_0$  ought to result in more support for  $\mathcal{H}_0$ , and do so without bound as  $n$  grows indefinitely. The final example concerns continuous data.

### 13.5 Example 3: Nullity of a Normal Mean

As a final example, we consider the case of the  $z$ -test: data are normally distributed with unknown mean  $\mu$  and known variance  $\sigma^2 = 1$ . For concreteness we consider a general law which states that the mean  $\mu$  equals 0, that is,  $\mathcal{H}_0 : \mu = 0$ . The model that relaxes the general law assigns a prior distribution to  $\mu$ ; specifically, we consider  $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, \sigma_0^2)$ . Similar to Examples 1 and 2, we consider the situation where the observed data are perfectly consistent with the predictions from  $\mathcal{H}_0$ . Consequently, we consider data for which the sample mean  $\bar{y}$  is exactly 0 and the sample variance  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$  is exactly 1.

The general desiderata are similar to those from Example 1 and 2: First, for any sample size  $n$  with sample mean equal to zero and sample variance equal to one, the data ought to support the general law  $\mathcal{H}_0$ ; second, as  $n$  increases, so should the level of support in favor of  $\mathcal{H}_0$ ; third, as  $n$  increases without bound, the support in favor of  $\mathcal{H}_0$  should grow infinitely large.

#### 13.5.1 Mathematical Result

In Example 3 the data consist of  $n$  realizations drawn from a normal distribution with mean  $\mu$  and known variance  $\sigma^2 = 1$ :  $y_i \sim \mathcal{N}(\mu, 1)$ ,  $i = 1, 2, \dots, n$ . Under  $\mathcal{H}_0$  the mean  $\mu$  is fixed to 0; under  $\mathcal{H}_1$ ,  $\mu$  is assigned a  $\mathcal{N}(0, \sigma_0^2)$  prior. The model corresponding to  $\mathcal{H}_0 : \mu = 0$  has no free parameters so that the Bayesian LOO estimate is obtained by summing the log likelihood values:

$$\text{elpd}_{\text{loo}}^{\mathcal{H}_0} = -\frac{n}{2} \log(2\pi) - \frac{n-1}{2}. \quad (13.18)$$

To calculate the LOO estimate under  $\mathcal{H}_1$ , one needs to be able to evaluate the predictive density for a single data point given the remaining data points. Recall that the posterior for  $\mu$  based on  $n-1$  observations is a  $\mathcal{N}(\mu_{-i}, \sigma_{-i}^2)$  normal distribution, with

$$\mu_{-i} = \frac{(n-1)\bar{y}_{-i}}{n-1 + \frac{1}{\sigma_0^2}}, \quad (13.19)$$

and

$$\sigma_{-i}^2 = \frac{1}{n-1 + \frac{1}{\sigma_0^2}}, \quad (13.20)$$

### 13. LIMITATIONS OF BAYESIAN LEAVE-ONE-OUT CROSS-VALIDATION FOR MODEL SELECTION

---

where  $\bar{y}_{-i} = \frac{1}{n-1} \sum_{j \neq i} y_j$  denotes the mean of the observations without the  $i$ th data point. Consequently, the leave-one-out predictive density is given by a  $\mathcal{N}(\mu_{-i}, 1 + \sigma_{-i}^2)$  distribution which follows from well-known properties of a product of normal distributions. Example 3 considers data sets that convey the maximal possible evidence for  $\mathcal{H}_0$  by having a sample mean of  $\bar{y} = 0$  and a sample variance of  $s^2 = 1$ . The Bayesian LOO estimate under  $\mathcal{H}_1$  is then given by:

$$\text{elpd}_{\text{loo}}^{\mathcal{H}_1} = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left( \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right) - \frac{(n-1) \left( n + \frac{1}{\sigma_0^2} \right)}{2 \left( n-1 + \frac{1}{\sigma_0^2} \right)}. \quad (13.21)$$

The difference in the LOO estimates can be written as

$$\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = \frac{n}{2} \log \left( \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right) + \frac{n-1}{2 \left( n-1 + \frac{1}{\sigma_0^2} \right)}. \quad (13.22)$$

As the sample size  $n$  grows without bound, the difference in the LOO estimates approaches a bound (see Appendix C for a derivation):

$$\lim_{n \rightarrow \infty} \Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = 1. \quad (13.23)$$

Consequently, the limit of the pseudo-Bayes factor is

$$\lim_{n \rightarrow \infty} \text{PSBF}_{01} = e \approx 2.718, \quad (13.24)$$

and the limit of the model weight for  $\mathcal{H}_0$  is

$$\lim_{n \rightarrow \infty} w_0 = \frac{e}{1+e} \approx 0.731, \quad (13.25)$$

which is identical to the limit obtained in Example 2.

#### 13.5.2 Select Scenarios

As in the previous two examples, the mathematical result can be applied to a series of select scenarios. Figure 13.3 shows the LOO weight in favor of the general law  $\mathcal{H}_0$  as a function of the sample size  $n$  with sample mean exactly zero and sample variance exactly one, separately for four different prior specifications of  $\mathcal{H}_1$ . The figure confirms that for each prior specification, the LOO weight for  $\mathcal{H}_0$  approaches the asymptotic bound as  $n$  grows large.

We conclude the following: (1) as  $n$  grows large, the support for the general law  $\mathcal{H}_0$  approaches a bound; (2) in contrast to Example 1, but consistent with Example 2, this bound is independent of the particular choice of normal prior distribution for  $\mu$  under  $\mathcal{H}_1$ ; however, consistent with both earlier examples, this bound is surprisingly low. Even with an infinite number of observations and a sample mean of exactly zero, the model weight on the general law  $\mathcal{H}_0$  does not exceed a modest 0.731; (3) as shown by case (a) in Figure 13.3, the choice of a  $\mathcal{N}(0, 3^2)$  prior distributions results in a function that approaches the asymptote

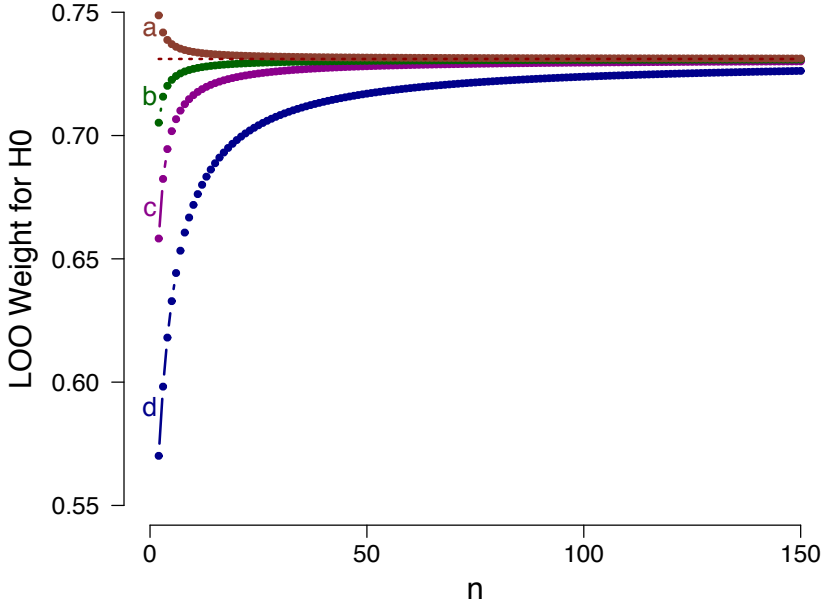


Figure 13.3: Example 3: LOO weights for  $\mathcal{H}_0 : \mu = 0$  as a function of sample size  $n$ , for data sets with sample mean equal to zero and sample variance equal to one, evaluated in relation to four different prior specifications for  $\mathcal{H}_1$ : (a)  $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, 3^2)$ ; (b)  $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, 1.5^2)$ ; (c)  $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, 1)$ ; (d)  $\mathcal{H}_1 : \mu \sim \mathcal{N}(0, 0.5^2)$ . The dotted horizontal line indicates the corresponding analytical asymptotic bound. See text for details. Available at <https://tinyurl.com/y7qhtp3o> under CC license <https://creativecommons.org/licenses/by/2.0/>.

from above. This means that, according to LOO, increasing the sample size of observations that are perfectly consistent with  $\mathcal{H}_0$  actually decreases the support for  $\mathcal{H}_0$ , violating the second desideratum outlined earlier; (4) some prior distributions (e.g.,  $\mu \sim \mathcal{N}(0, 2.035^2)$ ) result in a nonmonotonic relation, where the addition of  $\mathcal{H}_0$ -consistent observations initially increases the support for  $\mathcal{H}_0$ , and later decreases it toward asymptote.<sup>6</sup>

In sum, the result of the LOO procedure for a  $z$ -test involving  $\mathcal{H}_0 : \mu = 0$  shows behavior similar to that for the test of induction ( $\mathcal{H}_0 : \theta = 0$  or  $\mathcal{H}_0 : \theta = 1$ ) and the test against chance ( $\mathcal{H}_0 : \theta = 1/2$ ); this behavior violates two seemingly uncontroversial desiderata of inference, namely that the additional observation of data that are perfectly consistent with the general law  $\mathcal{H}_0$  ought to result in more

<sup>6</sup>Because the size of this nonmonotonicity is relatively small, we have omitted it from the figure. The OSF project page <https://osf.io/6s5zp/> contains a figure that zooms in on the nonmonotonicity.

support for  $\mathcal{H}_0$ , and do so without bound.

## 13.6 Closing Comments

Three simple examples revealed some expected as well as some unexpected limitations of Bayesian leave-one-out cross-validation or LOO. In the statistical literature it is already well-known that LOO is inconsistent (Shao, 1993), meaning that the true data-generating model will not be chosen with certainty as the sample size approaches infinity. Our examples provide a concrete demonstration of this phenomenon; moreover, our examples highlighted that, as the number of  $\mathcal{H}_0$ -consistent observations  $n$  increases indefinitely, the bound on support in favor of  $\mathcal{H}_0$  may remain modest. Inconsistency is arguably not a practical problem when the support is bounded at a level of evidence that is astronomically large, say a weight of 0.99999999; however, for both the test against chance and the  $z$ -test, the level of asymptotic LOO support for  $\mathcal{H}_0$  was categorized by Jeffreys (1939) as “not worth more than a bare comment” (p. 357).

It thus appears that, when the data are generated from a simple model, LOO falls prey to the Scylla of overfitting, giving undue preference to the complex model. The reason for this cuts to the heart of cross-validation: when two candidate models are given access to the same training set, this benefits the complex more than it benefits the simple model. In our examples, the simple model did not have any free parameters at all, and consequently these models gained no benefit whatsoever from having been given access to the training data; in contrast, the more complex models did have free parameters, and these parameters greatly profited from having been given access to the data set. Perhaps this bias may be overcome by introducing a cost function, such that the price for advance information (i.e., the training set) depends on the complexity of the model – models that stand to benefit more from the training set should pay a higher price for being granted access to it. Another approach is to abandon the leave-*one*-out idea and instead decrease the size of the training set as the number of observations  $n$  increases;<sup>7</sup> Shao (1993) demonstrated that this approach can yield consistency.

In order to better understand the behavior of leave-one-out cross-validation it is also useful to consider AIC, a method to which it is asymptotically equivalent (M. Stone, 1977). Indeed, for Example 2 and Example 3, the asymptotic LOO model weight equals that obtained when using AIC (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004). In addition, as pointed out by O’Hagan and Forster (2004, p. 187), “AIC corresponds to a partial Bayes factor in which one-fifth of the data are applied as a training sample and four-fifths are used for model comparison”. O’Hagan and Forster (2004) further note that this method is not consistent. It is also not immediately clear, in general, why setting aside one-fifth of the data for training is a recommendable course of action.

---

<sup>7</sup>Critics of cross-validation might argue that one weakness of the approach is that it is not a unique method for assessing predictive performance. That is, users of cross-validation need to decide which form to use exactly (e.g., leave-one-out, leave-two-out,  $k$ -fold, etc.) and different choices generally yield different results.

Another unexpected result was that, depending on the prior distribution, adding  $\mathcal{H}_0$ -consistent information may decrease the LOO preference for  $\mathcal{H}_0$ ; sometimes, as the  $\mathcal{H}_0$ -consistent observations accumulate, the LOO preference for  $\mathcal{H}_0$  may even be nonmonotonic, first increasing (or decreasing) and later decreasing (or increasing).

The examples outlined here are simple, and a LOO proponent may argue that, in real-world applications of substantive interest, simple models are never true, that is, the asymptotic data are never fully consistent with a simple model. Nevertheless, when researchers use LOO to compare two different models, it is important to keep in mind that the comparison is not between the predictive adequacy of the two models as originally entertained; the comparison is between predictive adequacy of two models where both have had advance access to all of the observations except one.

In sum, cross-validation is an appealing method for model selection. It directly assesses predictive ability, it is intuitive, and oftentimes it can be implemented with little effort. In the literature, it is occasionally mentioned that a drawback of cross-validation (and specifically LOO) is the computational burden involved. We believe that there is another, more fundamental drawback that deserves attention, namely the fact that LOO violates several common-sense desiderata of statistical support. Researchers who use LOO to adjudicate between competing mathematical models for cognition and behavior should be aware of this limitation and perhaps assess the robustness of their LOO conclusions by employing alternative procedures for model selection as well.

R code and more detailed derivations can be found at: <https://osf.io/6s5zp/>.

### 13.A Derivation Example 1 – Induction

To investigate how the difference in the LOO estimates

$$\begin{aligned}\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} &= \text{elpd}_{\text{loo}}^{\mathcal{H}_0} - \text{elpd}_{\text{loo}}^{\mathcal{H}_1} \\ &= -\log \left[ \left( \frac{a+n-1}{a+n-1+b} \right)^n \right]\end{aligned}$$

behaves as the number of observations goes to infinity, one can consider the limit of  $\left( \frac{a+n-1}{a+n-1+b} \right)^n$  as  $n \rightarrow \infty$ :

$$\lim_{n \rightarrow \infty} \left( \frac{a+n-1}{a+n-1+b} \right)^n = \exp \left\{ \lim_{n \rightarrow \infty} \frac{\log \left[ \frac{a+n-1}{a+n-1+b} \right]}{\frac{1}{n}} \right\}.$$

The limit of the denominator is  $\lim_{n \rightarrow \infty} \frac{1}{n} = 0$  and it is also straightforward to show that  $\lim_{n \rightarrow \infty} \log \left[ \frac{a+n-1}{a+n-1+b} \right] = 0$ . Therefore, both the limit of the numerator and of the denominator are 0 and L'Hôpital's rule can be applied which yields

$$\lim_{n \rightarrow \infty} \left( \frac{a+n-1}{a+n-1+b} \right)^n = \exp \left\{ - \lim_{n \rightarrow \infty} \frac{b}{1 + (2a-2+b)\frac{1}{n} + \frac{a^2-2a+ab+1-b}{n^2}} \right\}.$$

Hence,

$$\lim_{n \rightarrow \infty} \left( \frac{a+n-1}{a+n-1+b} \right)^n = \exp \{-b\}.$$

Therefore, the difference in the Bayesian LOO estimates  $\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1}$  as  $n \rightarrow \infty$  is given by:

$$\lim_{n \rightarrow \infty} \Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = b.$$

### 13.B Derivation Example 2 – Chance

The difference in the LOO estimates can be written as

$$\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = \log \left[ \left( \frac{a+b+n-1}{2a+n-2} \right)^{\frac{n}{2}} \right] + \log \left[ \left( \frac{a+b+n-1}{2b+n-2} \right)^{\frac{n}{2}} \right].$$

To investigate how this difference behaves as the number of observations goes to infinity, one can consider the limit of  $\left( \frac{a+b+n-1}{2a+n-2} \right)^{\frac{n}{2}}$  and of  $\left( \frac{a+b+n-1}{2b+n-2} \right)^{\frac{n}{2}}$  as  $n \rightarrow \infty$ . We first introduce a new variable  $m$  so that  $n = 2m$ , where  $m = 1, 2, 3, \dots$ , which ensures that the number of observation is even, and then consider the limits as  $m \rightarrow \infty$ . The limit of the first expression is given by

$$\lim_{m \rightarrow \infty} \left( \frac{a+b+2m-1}{2a+2m-2} \right)^m = \exp \left\{ \lim_{m \rightarrow \infty} \frac{\log \left( \frac{a+b+2m-1}{2a+2m-2} \right)}{\frac{1}{m}} \right\}.$$

The limit of the denominator is 0 and it is also straightforward to show that the limit of the numerator is 0. Hence, L'Hôpital's rule can be applied which yields

$$\lim_{m \rightarrow \infty} \left( \frac{a + b + 2m - 1}{2a + 2m - 2} \right)^m = \exp \left\{ \frac{b - a + 1}{2} \right\}.$$

Next, we consider the limit of the expressions in the second logarithm as  $m \rightarrow \infty$ :

$$\lim_{m \rightarrow \infty} \left( \frac{a + b + 2m - 1}{2b + 2m - 2} \right)^m = \exp \left\{ \lim_{m \rightarrow \infty} \frac{\log \left( \frac{a + b + 2m - 1}{2b + 2m - 2} \right)}{\frac{1}{m}} \right\}.$$

The limit of the denominator is 0 and it is also straightforward to show that the limit of the numerator is 0. Hence, L'Hôpital's rule can be applied which yields

$$\lim_{m \rightarrow \infty} \left( \frac{a + b + 2m - 1}{2b + 2m - 2} \right)^m = \exp \left\{ \frac{a - b + 1}{2} \right\}.$$

Therefore, the difference in the LOO of the two models as  $m \rightarrow \infty$  is given by:

$$\begin{aligned} \lim_{m \rightarrow \infty} [\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1}] &= \frac{b - a + 1}{2} + \frac{a - b + 1}{2} \\ &= 1. \end{aligned}$$

### 13.C Derivation Example 3 – Nullity of a Normal Mean

We first show how to obtain the expression for the difference in the LOO estimates. Note that the LOO estimate under  $\mathcal{H}_1$  can be written as:

$$\begin{aligned} \text{elpd}_{\text{loo}}^{\mathcal{H}_1} &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left( \frac{n + \frac{1}{\sigma_0^2}}{n - 1 + \frac{1}{\sigma_0^2}} \right) - \frac{n - 1 + \frac{1}{\sigma_0^2}}{2 \left( n + \frac{1}{\sigma_0^2} \right)} \sum_{i=1}^n y_i^2 \\ &\quad + \frac{n - 1}{n + \frac{1}{\sigma_0^2}} \sum_{i=1}^n y_i \bar{y}_{-i} - \frac{(n - 1)^2}{2 \left( n + \frac{1}{\sigma_0^2} \right) \left( n - 1 + \frac{1}{\sigma_0^2} \right)} \sum_{i=1}^n \bar{y}_{-i}^2. \end{aligned}$$

Since we consider data sets that have a sample mean of exactly 0, we know that  $\sum_{i=1}^n y_i = 0$  so that  $\sum_{j \neq i} y_j = -y_i$ . Furthermore, since the sample variance is exactly one and the sample mean is exactly zero, we know that  $s^2 = 1 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - 0)^2$ , hence,  $\sum_{i=1}^n y_i^2 = n - 1$ . Using these observations, one can show that

$$\begin{aligned} \sum_{i=1}^n y_i \bar{y}_{-i} &= \sum_{i=1}^n y_i \left[ \frac{1}{n - 1} \sum_{j \neq i} y_j \right] \\ &= \sum_{i=1}^n y_i \left[ -\frac{1}{n - 1} y_i \right] \\ &= -\frac{1}{n - 1} \underbrace{\sum_{i=1}^n y_i^2}_{n - 1} \\ &= -1, \end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^n \bar{y}_{-i}^2 &= \sum_{i=1}^n \left[ -\frac{1}{n-1} y_i \right]^2 \\ &= \frac{1}{(n-1)^2} \underbrace{\sum_{i=1}^n y_i^2}_{n-1} \\ &= \frac{1}{n-1}.\end{aligned}$$

Hence, using these results and after some further simplifications, the LOO estimate under  $\mathcal{H}_1$  can be written as

$$\text{elpd}_{\text{loo}}^{\mathcal{H}_1} = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left( \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right) - \frac{(n-1) \left( n + \frac{1}{\sigma_0^2} \right)}{2 \left( n-1 + \frac{1}{\sigma_0^2} \right)}.$$

Therefore, the difference in the LOO estimates can be written as

$$\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1} = \log \left[ \left( \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right)^{\frac{n}{2}} \right] + \frac{n-1}{2 \left( n-1 + \frac{1}{\sigma_0^2} \right)}.$$

To investigate how this difference behaves as the number of observations goes to infinity, we take the limit of each of the terms. The limit of the first term is obtained by taking the limit of the expression in the logarithm:

$$\lim_{n \rightarrow \infty} \left( \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right)^{\frac{n}{2}} = \exp \left\{ \frac{1}{2} \lim_{n \rightarrow \infty} \frac{\log \left[ \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right]}{\frac{1}{n}} \right\}.$$

The limit of the denominator is 0 and it is also straightforward to show that the limit of the numerator is 0. Hence, L'Hôpital's rule can be applied which yields

$$\lim_{n \rightarrow \infty} \left( \frac{n + \frac{1}{\sigma_0^2}}{n-1 + \frac{1}{\sigma_0^2}} \right)^{\frac{n}{2}} = \exp \left\{ \frac{1}{2} \right\}.$$

The limit of the second term is given by:

$$\lim_{n \rightarrow \infty} \frac{n-1}{2 \left( n-1 + \frac{1}{\sigma_0^2} \right)} = \frac{1}{2}.$$

Therefore, the difference in the LOO of the two models as  $n \rightarrow \infty$  is given by:

$$\begin{aligned}\lim_{n \rightarrow \infty} [\Delta \text{elpd}_{\text{loo}}^{\mathcal{H}_0, \mathcal{H}_1}] &= \log \left[ \exp \left\{ \frac{1}{2} \right\} \right] + \frac{1}{2} \\ &= 1.\end{aligned}$$

---

# Rejoinder: More Limitations of Bayesian Leave-One-Out Cross-Validation

---

## Abstract

We recently discussed several limitations of Bayesian leave-one-out cross-validation (LOO) for model selection. Our contribution attracted three thought-provoking commentaries. In this rejoinder, we address each of the commentaries and identify several additional limitations of LOO-based methods such as Bayesian stacking. We focus on differences between LOO-based methods versus approaches that consistently use Bayes' rule for both parameter estimation and model comparison. We conclude that LOO-based methods do not align satisfactorily with the epistemic goal of mathematical psychology.

Bayesian leave-one-out cross-validation (LOO) is increasingly popular for the comparison and selection of quantitative models of cognition and behavior.<sup>1</sup> In a recent article for *Computational Brain & Behavior*, we outlined several limitations of LOO (Gronau & Wagenmakers, 2019). Specifically, three concrete, simple examples illustrated that when a data set of infinite size is perfectly in line with the predictions of a simple model  $\mathcal{M}_S$  and LOO is used to compare  $\mathcal{M}_S$  to a more complex model  $\mathcal{M}_C$ , LOO shows bounded support for  $\mathcal{M}_S$ . As we mentioned, this model selection inconsistency has been known for a long time (e.g., Shao, 1993). We also discussed limitations that were unexpected (at least to us). Concretely,

---

This chapter is published as Gronau, Q. F., & Wagenmakers, E.-J. (2019). Rejoinder: More limitations of Bayesian leave-one-out cross-validation. *Computational Brain & Behavior*, 2, 35–47. doi: <https://doi.org/10.1007/s42113-018-0022-4>. Also available as *PsyArXiv preprint*: <https://psyarxiv.com/38z xu>

<sup>1</sup>Throughout this chapter, we use the terms *model comparison* and *model selection* interchangeably, although it may be argued that there is a subtle difference.

for data perfectly consistent with the simpler model  $\mathcal{M}_S$ , (1) the limiting bound of evidence for  $\mathcal{M}_S$  is often surprisingly modest; (2) the LOO preference for  $\mathcal{M}_S$  may be a non-monotonic function of the number of observations (meaning that additional observations perfectly consistent with  $\mathcal{M}_S$  may in fact *decrease* the LOO-preference for  $\mathcal{M}_S$ ); and (3) contrary to popular belief, the LOO result can depend strongly on the parameter prior distribution, even asymptotically.

Our discussion of the limitations of LOO attracted three commentaries. In the first commentary, Vehtari, Simpson, Yao, and Gelman (2019) claim that we “focus on pathologizing a known and essentially unimportant property of the method; and they fail to discuss the most common issues that arise when using LOO for a real statistical analysis”. Furthermore, Vehtari, Simpson, et al. state that we used a version of LOO that is not best practice and they suggest to use LOO-based Bayesian stacking instead (Yao et al., 2018). Vehtari, Simpson, et al. also criticize us for making the assumption that one of the models under consideration is “true” and use this as a springboard to question the usefulness of Bayes factors (e.g., Jeffreys, 1961; Kass & Raftery, 1995) and Bayesian model averaging (BMA; e.g., Hoeting et al., 1999; Jevons, 1874/1913). Finally, Vehtari, Simpson, et al. point out what they believe are more serious limitations of LOO-based methods. The second commentary is by Navarro (2019) and discusses how the scientific goal of explanation aligns with traditional statistical concerns; Navarro suggests that the model selection literature may focus too heavily on the statistical issues of model choice and too little on the scientific questions of interest. In the third commentary, Shiffrin and Chandramouli (2019) advocate Bayesian inference for non-overlapping model classes. Furthermore, Shiffrin and Chandramouli advocate tests of interval-null hypotheses instead of point-null hypotheses. Finally, Shiffrin and Chandramouli demonstrate that comparing non-overlapping hypotheses (where the null is an interval) eliminates the model selection inconsistency of LOO.

We thank the contributors for a productive discussion. To keep this rejoinder concise, we decided to address only the key points of disagreement. First, however, we will outline what we believe to be the primary goal of mathematical psychology.

## 14.1 Mathematical Psychology: An Epistemic Enterprise

Mathematical psychology is founded on the principle that psychological theories about cognition and behavior ought to be made precise by implementing them as quantitative models. Fum, Del Missier, and Stocco (2007, p. 136) write:

“Verbally expressed statements are sometimes flawed by internal inconsistencies, logical contradictions, theoretical weaknesses and gaps. A running computational model, on the other hand, can be considered as a sufficiency proof of the internal coherence and completeness of the ideas it is based upon.”

There exist different opinions about the role of models. As mentioned by Navarro (2019), Bernardo and Smith (1994, p. 238) state:

“Many authors [...] highlight a distinction between what one might call *scientific* and *technological* approaches to models. The essence of the dichotomy is that scientists are assumed to seek *explanatory* models, which aim at providing insight into and understanding of the “true” mechanisms of the phenomenon under study; whereas technologists are content with *empirical* models, which are not concerned with the “truth” but simply with providing a reliable basis for practical action in predicting and controlling phenomena of interest.”

Bernardo and Smith (1994, p. 238) conclude that when models are evaluated based on their predictions, the distinction is immaterial. In contrast, we believe that the distinction remains crucial. To us, the purpose of mathematical psychology is epistemic: the ultimate goal is to understand phenomena by developing theories, implementing these theories rigorously as quantitative models, and testing these models on observed data. Hence, our view of mathematical psychology aligns with what Bernardo and Smith call the “scientific approach”. In contrast, the main goal of the “technological approach” is the prediction of future data. There is an important distinction between these two approaches since, if the goal is solely prediction, one may be satisfied with models and methods that can be characterized as black-box “prediction devices”. The components and parameters of such prediction devices may not permit a substantive interpretation.

We believe that for many mathematical psychologists predictive adequacy is only a pragmatic means to an epistemic end. Quantitative models of cognition and behavior typically feature parameters that represent latent cognitive processes; these are of interest in and of themselves and do not serve only as tuning knobs of prediction devices. We do not wish to suggest that prediction is unimportant; in fact, we believe that models ought to be compared based on the predictions they made for observed data. However, we feel that the goal in mathematical psychology is virtually always an epistemic one, where models instantiate meaningful theories, and not a predictive one, where predictions are made for their own sake without the goal of developing and employing substantive theory. The following sections demonstrate by example that LOO-based methods have important limitations when the goal is epistemic rather than purely predictive.

## 14.2 Rejoinder to Vehtari, Simpson, Yao, & Gelman

Vehtari, Simpson, et al. (2019, henceforth VSYG) claim that we used a LOO version that is not in line with best practice and conclude that “[.] the claimed “limitations of Bayesian leave-one-out cross-validation” from GW do *not* apply to the version of Bayesian leave-one-out cross-validation that we recommend”. Specifically, (1) VSYG claim that we fail to take into account the empirical variance of the LOO estimate; they recommend doing so by using pseudo-BMA+ weights (Yao et al., 2018); (2) VSYG suggest that it would be even better to use Bayesian stacking (Yao et al., 2018). First, we agree that one should take into account the empirical variance of the LOO estimate in case it is non-zero. However, as VSYG mention “[...] this does not make a difference in their very specialized examples”. Second, since VSYG claim that the limitations we mentioned are well-known and

suggest Bayesian stacking instead, below we outline further limitations of LOO-based methods such as Bayesian stacking. We start by discussing the relevance of the assumption that one of the models under consideration is “true” which VSYG use to question the usefulness of Bayes factors and Bayesian model averaging.

### 14.2.1 LOO Is Motivated by an Illusory Distinction Between $\mathcal{M}$ -Open Tools and $\mathcal{M}$ -Closed Tools

LOO-based methods have been recommended for what is called the  $\mathcal{M}$ -open setting (Bernardo & Smith, 1994). Consider a set of  $M$  candidate models:  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$ .  $\mathcal{M}$ -open refers to a situation where the “true” model is not included in the set of candidate models. This stands in contrast to the  $\mathcal{M}$ -closed setting where one of the models in the set is “true” in the sense that it corresponds to the data-generating process.

In the  $\mathcal{M}$ -closed case it is valid (although not universally recommended; see Gelman, Carlin, et al., 2014, chapter 7.4; Gelman & Shalizi, 2013) to employ model comparison and prediction approaches that consistently use Bayes’ rule, not only to update one’s knowledge about parameters within a model, but also about the models themselves (e.g., by means of BMA, Bayes factors, posterior model probabilities). These approaches assign prior probabilities  $p(\mathcal{M}_k)$ ,  $k = 1, 2, \dots, M$  to a set of  $M$  models.<sup>2</sup>

In the  $\mathcal{M}$ -open case, however, the appropriateness of these supposedly “ $\mathcal{M}$ -closed tools” is often questioned (Bernardo & Smith, 1994, pp. 383–407; Yao et al., 2018). Moreover, George Box’s famous adage “all models are wrong” may then be invoked to question the use of these “ $\mathcal{M}$ -closed tools” in any practical application. For instance, Li and Dunson (in press) argue that “Philosophically, in order to interpret  $\text{pr}(\mathcal{M}_j \mid y^{(n)})$  as a model *probability*, one must rely on the (arguably always flawed) assumption that one of the models in the list  $\mathcal{M}$  is exactly true, known as the  $\mathcal{M}$ -closed case.”

Our objections to this line of reasoning are threefold. First, if we were to accept that these “ $\mathcal{M}$ -closed tools” are unsuitable for practical data analysis, this would similarly disqualify the specification of parameter priors and the computation of posterior predictives. As explained in the next section, individual parameter values or specific parameter ranges can be conceptualized as individual models.

Second, Bayes’ rule does not refer to an underlying ‘truth’ and the prior probability that is assigned across models (or across parameters) quantifies *relative* plausibility. Feldman (2015) has emphatically argued this point:<sup>3</sup>

---

<sup>2</sup>Note that the value of the Bayes factor is independent of the prior model probabilities since it quantifies the *change* from prior to posterior model odds. However, although it is independent of the value of the prior model odds, it assumes that, in principle, these could be specified.

<sup>3</sup>Relatedly, Wasserman (2000, p. 103) argued: “Second, even when all models are wrong, it is useful to consider the relative merits of two models. Newtonian physics and general relativity are both wrong. Yet it makes sense to compare the relative evidence in favor of one or the other. Our conclusion would be: under the tentative working hypothesis that one of these two theories is correct, we find that the evidence strongly favors general relativity. It is understood that the working hypothesis that one of the models is correct is wrong. But it is a useful, tentative hypothesis and, proceeding under that hypothesis, it makes sense to evaluate the relative posterior probabilities of those hypotheses.”

“But such a strong assumption [that one of the candidate models is true] is not really necessary in a Bayesian framework—at least, it is not required or implied by any of the equations. Rather, Bayesian inference only assumes that there is some set  $M$  of possible models under consideration, which are tied to the data via likelihood functions  $p(X | M)$ . Bayes’ rule allows these models to be compared *to each other* in terms of plausibility, but says nothing whatsoever about whether any of the models is true in a larger or absolute sense (see Feldman, 2014). The ‘truth’ of the models (whatever that even means—see remarks above about semantics) never enters into it.” (Feldman, 2015, p. 1524)

Third, Feldman (2013, pp. 17-18) points out, as did Bayesian pioneers Ed Jaynes and Dennis Lindley before him, that the assignment of prior probabilities is always conditional on background knowledge  $\mathcal{K}$ . Hence, when we write  $p(\mathcal{M}_k)$  this is really just a convenient shorthand for the more accurate notation  $p(\mathcal{M}_k | \mathcal{K})$ , a renormalized probability for a subset of relevant models selected by conditioning on the current knowledge  $\mathcal{K}$ . Background knowledge  $\mathcal{K}$  provides the pragmatic filter that allows us to define, from the infinite collection of possible models, a subset of models that pass a certain minimum threshold of plausibility, feasibility, or substantive interest. This conceptualization of prior model probabilities is in line with our epistemic view on mathematical psychology. Given a set of competing theoretical accounts of interest, implemented as quantitative models (i.e., given our background knowledge  $\mathcal{K}$ ), we are interested in quantifying the relative evidence for each of these models based on observed data. Nowhere do we assume any of the models that represent rival theories to be true in an absolute sense.

We do not wish to suggest that the possibility of model-misspecification can be happily ignored; all models necessarily make assumptions and simplifications and it may happen that given a set of models, even the best one fails to provide a satisfactory description of the phenomenon of interest. In our opinion, however, this does *not* suggest that the entire approach of assigning prior probabilities to a set of rival models is flawed from the outset or that it does not make sense. In contrast, the presence of model-misspecification suggests that one ought to refine the models or develop new theories that are able to better capture the relevant aspects of the phenomenon of interest (i.e., expand the background knowledge base  $\mathcal{K}$ ). These new model versions can then be incorporated in the set of models and can be compared to each other based on new data.

### 14.2.2 LOO Depends on an Arbitrary Distinction Between Parameter Estimation and Model Comparison

We do not believe that the distinction between  $\mathcal{M}$ -open and  $\mathcal{M}$ -closed is a valid argument against approaches that consistently use Bayes’ rule for both parameters and models. Those who disagree may feel that assigning model probabilities  $p(\mathcal{M}_k)$  does not make sense in the  $\mathcal{M}$ -open setting; these dissenters would, in our opinion, then also need to object to assigning prior probabilities to parameters and computing quantities such as posterior predictives. The reason is that the distinction between parameter estimation and model comparison can be regarded

as artificial (see also Gelman, 2011, p. 76). It has long been known that estimation can be viewed as a special case of model comparison (also known as ‘testing’):<sup>4</sup>

“We shall not consider the philosophy of Bayesian estimation procedures here. These procedures can be regarded as a special case of Bayesian hypothesis testing since every statement of the form that a vectorial parameter belongs to a region is itself a hypothesis [but estimates are less often formulated before making observations].” (Good, 1983, p. 126)

### 14.2.2.1 Discrete Parameters

The fact that labeling a problem as parameter estimation or model comparison can be regarded as arbitrary is most apparent for discrete parameter models. As a concrete example, consider a scenario inspired by Hammersley (1950, p. 236) about tumor transplantability in mice (see also Choirat & Seri, 2012). For a certain type of mating, the probability of a tumor “taking” when transplanted from one of the grandparents is  $(1/2)^k$ , where  $k$  is an integer that corresponds to the number of genes determining transplantability (all of which must be present for a “take” to occur). Suppose, for illustrative purposes, we know that the number of relevant genes is between 1 and 10 and we deem each number equally likely a priori:  $p(k) = 1/10$ , for all  $k \in \{1, 2, \dots, 10\}$ . The likelihood corresponds to a binomial distribution with success probability  $\theta = (1/2)^k$ . Suppose fictitious data show 1 “take” out of 6 attempts. The resulting posterior distribution for  $k$  is displayed in Figure 14.1. In this example,  $k$  could be regarded as a parameter, so that the distribution in Figure 14.1 is a parameter posterior distribution. However,  $k$  could also be regarded as an index for a set of 10 competing models  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{10}$ , where  $\mathcal{M}_k : \theta = (1/2)^k$ ,  $k = 1, 2, \dots, 10$ . In this case, the distribution in Figure 14.1 visualizes the posterior model probabilities.

After having obtained a posterior over the number of genes  $k$ , one may be interested in predicting new data  $y_{\text{new}}$  given the observed data  $y$  (i.e., 1 “take” out of 6 attempts). This is achieved by marginalizing over  $k$ :

$$p(y_{\text{new}} | y) = \sum_{k=1}^{10} p(y_{\text{new}} | k) p(k | y), \quad (14.1)$$

where  $p(k | y)$  corresponds to the posterior distribution depicted in Figure 14.1. When  $k$  is regarded as a parameter, Equation 14.1 corresponds to the posterior predictive distribution; when  $k$  is regarded as indexing separate models, Equation 14.1 corresponds to the BMA predictive distribution for new data. This shows that the mathematical operation of computing a posterior predictive is identical to that used in Bayesian model averaging.<sup>5</sup> Proponents of LOO-based methods who believe there is an issue with BMA may not appreciate that this issue applies with equal force to posterior predictives, a concept integral to LOO-based

---

<sup>4</sup>See also <http://www.bayesianspectacles.org/bayes-factors-for-those-who-hate-bayes-factors/>.

<sup>5</sup>Appendix A contains a fragment from Jevons (1874/1913) that features another example.

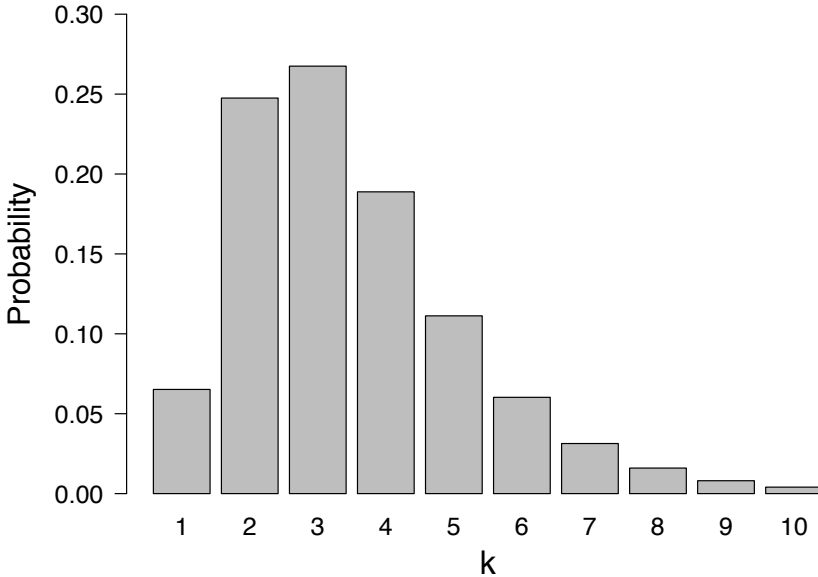


Figure 14.1: Parameter estimation or model comparison? Shown is the posterior distribution for the tumor transplant example based on 1 “take” out of 6 attempts and a uniform prior for  $k$ , the number of genes determining transplantability. Here  $k$  may be regarded as a parameter, such that the depicted distribution is a parameter posterior distribution, or  $k$  may be regarded as indexing separate models, so that the depicted distribution corresponds to posterior model probabilities. Available at <https://tinyurl.com/y94uj4h8> under CC license <https://creativecommons.org/licenses/by/2.0/>.

methods such as Bayesian stacking. When treating  $k$  as a parameter, one could equally ask ‘what if none of the values for  $k$  is ‘true’? How can we define  $p(k)$  in the knowledge that none of these values will perfectly capture the data-generating process?’

As mentioned earlier, one may argue that it *does* make sense to define  $p(k)$ , even when it is not strictly speaking true, because we assume that we operate within a more narrow context, one that is obtained by conditioning on a model  $\mathcal{M}_{\text{Estimation}}$ :<sup>6</sup>  $p(k \mid \mathcal{M}_{\text{Estimation}})$ . We agree and, crucially, this conditioning argument applies to models as well; we should really write  $p(\mathcal{M}_k \mid \mathcal{K})$ , that is, the probability of model  $\mathcal{M}_k$  given background knowledge  $\mathcal{K}$ . Both for parameters and models, plausibility assessments are always part of a subset of possibilities. In other words, regardless of whether we are estimating parameters or comparing

<sup>6</sup>Note that, in contrast to  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{10}$ , the model  $\mathcal{M}_{\text{Estimation}}$  does not fix  $k$  to a single value but allows  $k$  to vary freely.

models, we have to make assumptions and simplifications. When these assumptions are violated this signals a potential problem with the inference, but it does not mean that the entire approach is flawed from the outset. In sum, for predictions from discrete parameter models the proponents of LOO may recommend posterior predictives when the problem is phrased as estimation, whereas they may recommend LOO-based Bayesian stacking when the problem is phrased as model comparison.

#### 14.2.2.2 Continuous Parameters

We have argued that the distinction between parameter estimation and model comparison is purely semantic. Bayes' rule does not care about such labels: the same result is obtained regardless of what is called a parameter or a model. In contrast, LOO-based methods lack this coherence: the distinction between parameters and models is crucial. For instance, BMA yields the same results as Bayesian parameter estimation when the set of models is obtained by partitioning a continuous parameter space into non-overlapping intervals, with prior model probabilities set equal to the prior mass in the respective intervals (see Appendix B for a derivation). As a concrete example, suppose observations  $y_i$ ,  $i = 1, 2, \dots, n$  are assumed to follow a Bernoulli distribution with success probability  $\theta$ . In this scenario, one could assign  $\theta$  a prior distribution  $p(\theta)$  – for concreteness, we assume a uniform prior – and then obtain a posterior for  $\theta$ . Subsequently, one may obtain predictions for a new data point  $y_{\text{new}}$  based on the posterior for  $\theta$ . Alternatively, one could also use BMA for the following three models:  $\mathcal{M}_1 : \theta \in [0, .25]$ ,  $\mathcal{M}_2 : \theta \in [.25, .75]$ , and  $\mathcal{M}_3 : \theta \in (.75, 1]$ . Given a uniform prior on  $\theta$ , BMA and Bayesian parameter estimation yield identical results when (1) the prior for  $\theta$  under each model is a (renormalized) uniform prior, and (2) the prior model probabilities are  $p(\mathcal{M}_1) = .25$ ,  $p(\mathcal{M}_2) = .5$ , and  $p(\mathcal{M}_3) = .25$  (i.e., the probabilities that the uniform prior for  $\theta$  assigns to the three intervals).

The left column of Figure 14.2 displays the BMA results for  $n = 20$  observations, half of which are successes. Panel (1a) depicts the uniform prior distribution for  $\theta$  that is partitioned into three intervals to produce the models  $\mathcal{M}_1$ ,  $\mathcal{M}_2$ , and  $\mathcal{M}_3$ . The displayed prior model probabilities correspond to the mass that the uniform prior for  $\theta$  assigns to each interval. Panel (1b) displays the BMA posterior distribution – it is identical to the posterior obtained when conducting Bayesian parameter estimation for the common model that assigns  $\theta$  a uniform prior from 0 to 1. The weights that BMA uses to average the results of the different models are given by the posterior model probabilities.  $\mathcal{M}_2$  receives almost all posterior model probability:  $p(\mathcal{M}_2 | y) = .99$ , as the observed data are predicted much better by values of  $\theta$  that are inside rather than outside the  $[.25, .75]$  interval. Panel (1c) displays the BMA predictive distribution for a single new observation  $y_{\text{new}}$ . This distribution is identical to the posterior predictive distribution obtained based on Bayesian parameter estimation. In line with the fact that the posterior for  $\theta$  is symmetric around .5,  $y_{\text{new}}$  is predicted to be a success with probability .5.

The right column of Figure 14.2 displays the results obtained when using Bayesian stacking (Yao et al., 2018). Panel (2a) displays again the uniform prior distribution for  $\theta$  that is partitioned into three intervals to produce the models

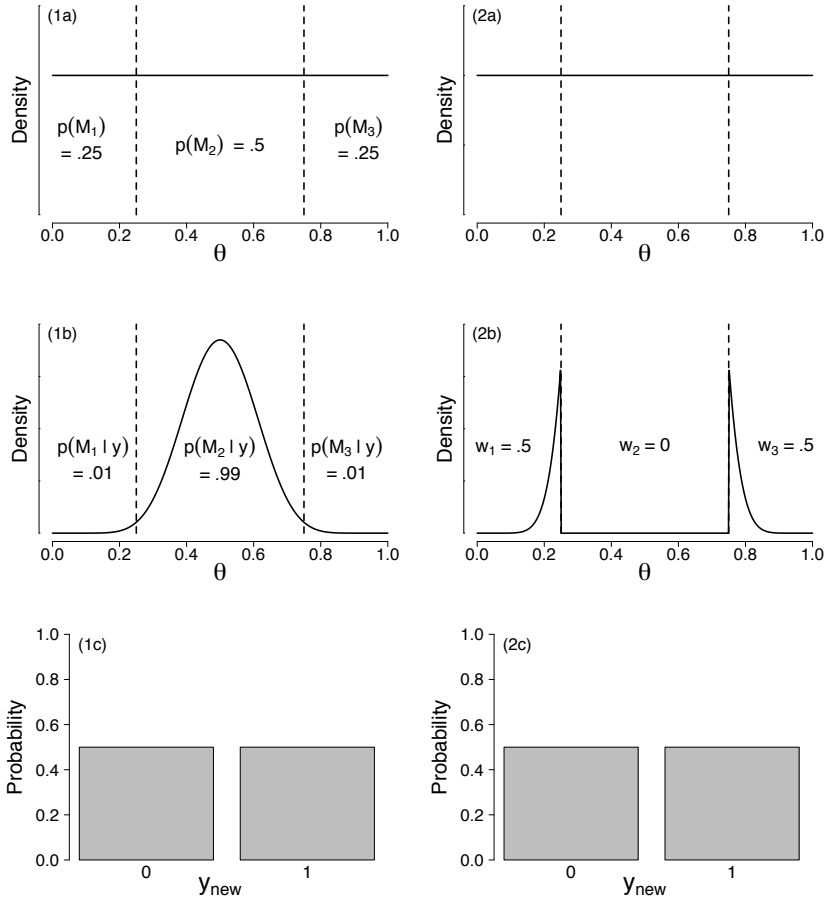


Figure 14.2: BMA (left column) and Bayesian stacking (right column) results for the Bernoulli example based on 10 successes out of  $n = 20$  observations. Panels (1a) and (2a) show the uniform prior distribution for  $\theta$  which is partitioned into three non-overlapping intervals to yield models  $\mathcal{M}_1$ ,  $\mathcal{M}_2$ , and  $\mathcal{M}_3$ . Panel (1a) also displays the prior model probabilities (not used in stacking). Panel (1b) displays the BMA posterior based on using the posterior model probabilities as averaging weights, and panel (2b) displays a model-averaged posterior obtained using the stacking weights. Panel (1c) displays the BMA predictions for a single new observation  $y_{\text{new}}$  and panel (2c) displays the corresponding predictions from stacking. Available at <https://tinyurl.com/yaql2vt4> under CC license <https://creativecommons.org/licenses/by/2.0/>.

$\mathcal{M}_1$ ,  $\mathcal{M}_2$ , and  $\mathcal{M}_3$ . In contrast to BMA, Bayesian stacking does not assign prior probabilities to the different models. Panel (2b) displays a model-averaged pos-

Table 14.1: LOO predictive densities.

Observation	$p(y_i   y_{-i}, \mathcal{M}_1)$	$p(y_i   y_{-i}, \mathcal{M}_2)$	$p(y_i   y_{-i}, \mathcal{M}_3)$
$y_i = 0$	.7758	.4786	.2206
$y_i = 1$	.2206	.4786	.7758

terior distribution and panel (2c) displays the Bayesian stacking predictive distribution; both of these are obtained by combining the different models according to the stacking weights.<sup>7</sup> The stacking-based predictions are indistinguishable from those of BMA and appear very reasonable: it is predicted that the next observation will be a success with probability .5. However, the stacking weights themselves are highly undesirable indicators of the plausibility of the different models in light of the observed data.  $\mathcal{M}_2$ , the model that clearly outpredicts the other two, is in fact decisively ruled out, as its stacking weight is equal to 0. To understand this result, first note that the stacking weights  $w_k$ ,  $k = 1, 2, \dots, M$  are obtained by maximizing the following objective function (subject to the constraint that  $w_k \geq 0$  and  $\sum_{k=1}^M w_k = 1$ ):

$$\frac{1}{n} \sum_{i=1}^n \log \left( \sum_{k=1}^M w_k p(y_i | y_{-i}, \mathcal{M}_k) \right). \quad (14.2)$$

Table 14.1 displays the LOO predictive density values for  $y_i = 0$  and  $y_i = 1$  for the three models under consideration.  $\mathcal{M}_1$  and  $\mathcal{M}_3$  make mirrored predictions whereas the LOO predictive density for  $\mathcal{M}_2$  is identical for  $y_i = 0$  and  $y_i = 1$ . Combining the models' LOO predictive densities according to the stacking weights  $w_1 = .5$ ,  $w_2 = 0$ , and  $w_3 = .5$  yields  $\sum_{k=1}^M w_k p(y_i | y_{-i}, \mathcal{M}_k) \approx .4982$ , for all  $i = 1, 2, \dots, n$ . The objective function thus attains a larger value than when using, for instance,  $w_1 = 0$ ,  $w_2 = 1$ , and  $w_3 = 0$  ( $\sum_{k=1}^M w_k p(y_i | y_{-i}, \mathcal{M}_k) \approx .4786$ ), or when using  $w_1 = 1/3$ ,  $w_2 = 1/3$ , and  $w_3 = 1/3$  ( $\sum_{k=1}^M w_k p(y_i | y_{-i}, \mathcal{M}_k) \approx .4917$ ).

We need to emphasize that Yao et al. do not suggest to use the stacking weights to obtain a model-averaged posterior as in panel (2b); instead, Yao et al. focus purely on predictions. Nevertheless, this distribution highlights the undesirable nature of the stacking weights when used as indicators for the plausibility of different models and parameters. The plot also shows how Bayesian stacking achieves predictions that are indistinguishable from the BMA predictions by combining two models with low plausibility that make mirrored predictions.

Bayesian stacking was designed to make good predictions in the presence of model-misspecification and may be a valuable tool in case prediction is the main goal. However, we believe that mathematical psychology has an epistemic purpose: researchers are typically interested in quantifying the evidence for different models which represent competing theories of cognition and behavior. Our example illustrates that the stacking weights do not appear to align satisfactorily with this goal. This is also highlighted by the fact that, as VSYG mention, the stacking weight for a simple general law model (i.e., Example 1 of Gronau & Wagenmakers,

<sup>7</sup>The stacking weights were obtained using the `loo` package (Vehtari et al., 2018).

2019) is equal to 1 when all observations are in line with the general law, *independent* of the number of observations  $n$ . VSYG state: “The lack of dependence on  $n$  may look suspicious”. Indeed; suppose one is asked whether all swans are white and two white swans are observed. Is it warranted to conclude that the general law is now firmly established? Should predictions about the future disregard the possibility that the general law might fail? Even though VSYG provide an explanation why they believe suspicion is not warranted, we remain doubtful.

In sum, we are skeptical about the usefulness of Bayesian stacking in mathematical psychology where the goal is of an epistemic and not a purely predictive nature.

### 14.2.3 LOO Depends on an Arbitrary Distinction Between Data that Arrive Sequentially or “Simultaneously”

LOO is based on repeatedly leaving out one of the observations and evaluating the prediction for this held-out data point based on the remaining observations. Concretely, given data  $y = (y_1, y_2, \dots, y_n)$ , LOO evaluates the predictive density  $p(y_i | y_{-i})$  for all  $i = 1, 2, \dots, n$ , where  $y_{-i}$  denotes all data points except the  $i$ th one. It is well-known that LOO is theoretically unsatisfactory when applied to time series data since, in this case, LOO uses the future to predict the past, for all  $i \neq n$  (e.g., Bürkner, Vehtari, & Gabry, 2018). As VSYG point out, there exist alternative cross-validation schemes that do not have this property and may be applied in this context (e.g., Bürkner et al., 2018). Therefore, time series data are treated differently from data that do not exhibit a temporal structure. However, we argue that *all* data form a time series. When conducting an experiment, participants come in over time; the data have a temporal order. Consequently, the use of LOO implies that one uses the future to predict the past. It seems unsatisfactory to apply a method that is not recommended for time series to data that have a temporal order, even if that temporal order is disregarded in the analysis because the observations are judged to be exchangeable.

Another consequence of the fact that LOO does not respect the temporal nature of the data is that LOO is inconsistent with what Dawid (1984, p. 278) termed the *prequential approach* which “[...] is founded on the premiss that the purpose of statistical inference is to make sequential probability forecasts for future observations”. In contrast, Bayes factors are consistent with the prequential approach (e.g., Wagenmakers, Grünwald, & Steyvers, 2006). The reason is that the Bayes factor compares two models based on the ratio of their *marginal likelihoods*. The marginal likelihood corresponds to the joint probability of the data given a model. Consequently, it is easy to show that the marginal likelihood of model  $\mathcal{M}_k$  can be conceptualized as an accumulation of one-step-ahead predictions:

$$p(y | \mathcal{M}_k) = p(y_1 | \mathcal{M}_k) p(y_2 | y_1, \mathcal{M}_k) p(y_3 | y_{1:2}, \mathcal{M}_k) \dots p(y_n | y_{1:(n-1)}, \mathcal{M}_k), \quad (14.3)$$

where  $y_{1:i} = (y_1, y_2, \dots, y_i)$  denotes the first  $i$  observations. Each term in Equation 14.3 is obtained by integrating over the model parameters  $\theta$ . For the first observation,  $p(y_1 | \mathcal{M}_k) = \int_{\Theta} p(y_1 | \theta, \mathcal{M}_k) p(\theta | \mathcal{M}_k) d\theta$ , and for  $i > 1$ ,  $p(y_i | y_{1:(i-1)}, \mathcal{M}_k) = \int_{\Theta} p(y_i | \theta, y_{1:(i-1)}, \mathcal{M}_k) p(\theta | y_{1:(i-1)}, \mathcal{M}_k) d\theta$ . Thus, Bayes

factors – but not LOO – produce the same result, regardless of whether the data are analyzed one at a time or all at once.

A common criticism of the Bayes factor is its dependence on the parameter prior distribution since one starts by making predictions based on the prior distribution. There are a number of replies to this concern. First, it may be regarded as desirable that the result depends on the prior information, as this allows one to incorporate existing prior knowledge. In mathematical psychology, parameters typically correspond to psychological variables about which theories exist; the parameter prior can be used to encode these existing psychological theories (e.g., Lee & Vanpaemel, 2018; Vanpaemel, 2010). Second, proponents of LOO who criticize Bayes factors for being prior dependent do not object to generating predictions based on posterior distributions, as this is an integral part of the LOO procedure. However, the prior that one entertains at a certain time may be the posterior based on past observations. Third, as is good practice in parameter inference, concerns about prior sensitivity of the Bayes factor may be alleviated by conducting sensitivity analyses across a range of plausible prior distributions. In many cases, the sensitivity analysis may show that the qualitative conclusions are robust to the exact prior choice. However, when the results change drastically this is also valuable information since it highlights that researchers with different, reasonable prior beliefs may draw quite different conclusions.

In sum, we argue that LOO uses the future to predict the past: all data have a temporal structure, even though the analyst may not have access to it or may choose to ignore it. LOO is therefore inconsistent with Dawid’s prequential approach. In contrast, Bayes factors can be naturally conceptualized as assessing the models’ sequential, probabilistic one-step-ahead predictions and are thus consistent with the prequential approach.

### 14.3 Rejoinder to Navarro

The commentary by Navarro (2019) discusses how the scientific goal of explanation aligns with traditional statistical concerns and suggests that the model selection literature may focus too much on the statistical issues of model choice and too little on the scientific questions of interest.<sup>8</sup> In line with our epistemic view on mathematical psychology, we agree that the starting point should always be meaningful theories that are made precise by implementing them as quantitative models. The models’ plausibilities may then be evaluated based on observed data. In case the data pass what Berkson termed the *interocular traumatic test* – the data are so compelling that the conclusion “hits you straight between the eyes” – no statistical analysis may be required. However, as Edwards, Lindman, and Savage (1963, p. 217) remark: “[...] the enthusiast’s interocular trauma may be the skeptic’s random error. A little arithmetic to verify the extent of the trauma can yield great peace of mind for little cost.” Furthermore, often the data may not yield a clear result at first sight; consequently, we believe it is useful to more formally quantify the evidence for the models, just as it is useful to make

---

<sup>8</sup>One key aspect that is being discussed is the  $\mathcal{M}$ -open versus  $\mathcal{M}$ -closed distinction that we have already addressed in a previous section.

verbal theories precise by implementing them as quantitative models. Of course, researchers should be aware of the assumptions not only of their models but also of their model evaluation metrics. We agree with Lewandowsky and Farrell (2010, p. 10): “Model comparison rests on both quantitative evaluation and intellectual and scholarly judgment.”

Navarro writes: “I am of the view that the behaviour of a selection procedure applied to toy problems is a poor proxy for the inferential problems facing scientists.” First, although the examples we used are simple, we do not regard them as “toy problems”. Our first example dealt with quantifying evidence for a general law of the form “all  $X$ ’s have property  $Y$ ”; this is perhaps the world’s oldest inference problem and has been discussed by a plethora of philosophers, mathematicians, and statisticians (e.g., Laplace, 1829/1902; Polya, 1954a, 1954b; Wrinch & Jeffreys, 1919). Even Aristotle was already concerned with making inference about a general law (Whewell, 1840, p. 294):<sup>9</sup>

“We find that several animals which are deficient in bile are longlived, as man, the horse, and the mule; hence we infer that *all* animals which are deficient in bile are longlived.” (*Analytica Priora*, ii, 23)

Second, although we agree with Navarro that scientists should also consider more complex problems, we still believe that considering simple problems is invaluable for investigating how model evaluation metrics behave. Suppose one considers a simple example and finds that a model evaluation metric of interest exhibits highly undesirable properties. One could proceed to more complex problems in the hope that these undesirable properties will not be manifest; however, to us, it seems questionable whether this hope is warranted and it may be considerably harder to verify this in the more complex case.

Navarro uses an example to showcase how Bayes factors can “misbehave”. A general law model  $\mathcal{M}_1$  that asserts that a Bernoulli probability  $\theta$  equals 1 is compared to an “unknown quantity” model  $\mathcal{M}_2$  that assigns  $\theta$  a uniform prior. For any data set of size  $n$  that consists of only successes with the exception of a single failure, the Bayes factor will decisively rule out the general law model  $\mathcal{M}_1$  in favor of  $\mathcal{M}_2$ .<sup>10</sup> Navarro concludes that the Bayes factor misbehaves since “In real life none of us would choose  $\mathcal{M}_2$  over  $\mathcal{M}_1$  in this situation, because from our point of view the general law model is actually “closer” to the truth than the uninformed model”. Navarro furthermore states: “While there are many people who assert that “a single failure is enough to falsify a theory”, I confess I have not yet encountered anyone willing to truly follow this principle in real life”. Indeed, we believe that a single failure is enough to falsify a general law and so did, for instance, Wrinch and Jeffreys (1919, p. 729):

“[...] if for instance we consider that either Einstein’s or Silberstein’s form of the principle of general relativity is true, a single fact

---

<sup>9</sup>The authors would like to state that they disagree with the conclusion in this particular example.

<sup>10</sup>Note that  $n$  may be infinity.

contradictory to one would amount to a proof of the other in every case.”

Other examples are provided by Polya (1954a) who discussed how mathematical conjectures are “irrevocably exploded” by a single failure. For instance, the famous Goldbach conjecture holds that every even integer greater than two can be expressed as the sum of two prime numbers. The conjecture has been confirmed for all integers up to  $4 \times 10^{18}$ .<sup>11</sup> Yet, the occurrence of a single failure would refute the Goldbach conjecture decisively. Polya (1954a, p. 6) notes how the search for a suitable decomposition of 60 has ended in success ( $60 = 7 + 53$ ) and explains:

“The conjecture has been verified in one more case. The contrary outcome would have settled the fate of Goldbach’s conjecture once and for all. If, trying all primes under a given even number, such as 60, you never arrive at a decomposition into a sum of two primes, you thereby *explode the conjecture irrevocably* [italics ours].”

Finally, suppose the general law of interest states that “all swans are white”. In case one traveled to Australia and observed a single black swan, to us, the only reasonable conclusion to draw would be that the general law does not hold. We speculate that researchers who believe that in this situation  $\mathcal{M}_1$  should be favored do not truly entertain a general law model, but an alternative model  $\mathcal{M}_1^*$  that states “*almost* all  $X$ ’s have property  $Y$ ”. Under  $\mathcal{M}_1^*$ ,  $\theta$  is assigned a prior that is concentrated near 1 but does not completely rule out values very close to 1 (e.g.,  $\theta \sim \text{Beta}(a, 1)$ , with  $a$  large). This showcases that what has been termed a “misbehavior” of the Bayes factor may be due to the implicit invocation of a third model  $\mathcal{M}_1^*$  as a replacement of the general law model  $\mathcal{M}_1$ .

## 14.4 Rejoinder to Shiffrin & Chandramouli

Shiffrin and Chandramouli (2019, henceforth SC) argue in favor of comparing non-overlapping model classes using Bayesian inference. Furthermore, SC advocate focusing on interval-null hypotheses instead of point-null hypotheses. Finally, SC demonstrate that comparing non-overlapping hypotheses (where the null is an interval) eliminates the model selection inconsistency of LOO. We believe it is interesting to see that LOO can be made consistent when the models are defined so that the parameter spaces do not overlap, although – as SC state themselves – the result is not completely unexpected.

SC remark that when testing a point-null versus a hypothesis that assigns a continuous prior distribution to the parameter of interest, the “standard” approach of calculating Bayes factors is identical to SCs proposal to consider non-overlapping models (since a single point has measure zero). Therefore, SCs approach only differs in case one does not consider point-null hypotheses. We believe that it may be of interest to consider interval-hypotheses in certain scenarios; in these cases, we agree that defining the models such that the parameter spaces do not overlap

---

<sup>11</sup><http://sweet.ua.pt/tos/goldbach.html>

can be beneficial (see also Morey & Rouder, 2011). However, we also believe that there are situations where it is useful to test point-null hypotheses.<sup>12</sup>

First, we believe that there are situations in which the point-null is exactly true. SC mention an example of testing ESP with coin flipping and argue that the ‘chance’ point-null hypothesis is never exactly true since coins are never perfect and, consequently, will not produce ‘heads’ with probability exactly .5. However, consider the following alternative experiment for testing ESP: Participants are presented with pictures either on the right or left side of the screen and are asked to indicate on which side the next picture will appear. Suppose that exactly half of the pictures are presented on the right, the other half on the left (and the order is randomly permuted). In this scenario, given that we do not believe in ESP, we believe that the point-null – which states that the probability of a correct response is .5 – is exactly true.

Second, we believe that testing point-null hypotheses is crucial in all stages of cognitive model development, validation, and application. When developing and validating a model, it is important to show that certain experimental manipulations selectively influence only a subset of the model parameters whereas the remaining parameters are unaffected. In applications, cognitive models may be used, for instance, to investigate which subprocesses differ or do not differ in clinical subpopulations (cognitive psychometrics; e.g., Riefer et al., 2002). In these applications researchers are interested in quantifying evidence for a difference (“there is evidence that cognitive process  $X$  is affected”), but, crucially, also for an invariance or, equivalently, point-null hypothesis (“there is evidence that cognitive process  $Y$  is *not* affected”).<sup>13</sup>

Third, even in case one does not believe that the point-null hypothesis can be true exactly, it appears that it is still useful to be able to reject at least this “unreasonable” hypothesis. For instance, if one wants to convince a skeptic that a new research finding works, it seems difficult to do so if one cannot even reject a point-null hypothesis which some people argue is never true exactly.

To use SCs proposal in practice, it appears crucial to be able to detect shared model instances (i.e., parameter settings that predict the same outcome distribution). This may not always be straightforward, especially when the two models are defined on different parameter spaces. Consider the comparison between  $\mathcal{M}_1$  with parameter  $\theta \in \Theta$  and  $\mathcal{M}_2$  with parameter  $\xi \in \Xi$ . Suppose one is told that  $\theta$  corresponds to a Bernoulli success probability and  $\xi = \log(\theta/(1 - \theta))$  denotes the log odds with the restriction that  $\xi > 0$ . In this case, it is straightforward to see that the models share instances (i.e., the restriction  $\xi > 0$  corresponds to  $\theta > .5$ ). Consequently, it appears to us that SC would recommend to eliminate the shared instances and would consider the comparison between  $\mathcal{M}_1^* : \theta \leq .5$  and  $\mathcal{M}_2 : \xi > 0$ . However, in case the models under consideration are more complex

<sup>12</sup>We have detailed our arguments for why we believe it can be useful to test point-null hypotheses in the following blog posts: <https://tinyurl.com/y8org8bt> and <https://tinyurl.com/ya7c13cq>.

<sup>13</sup>Proponents of interval-null hypotheses might argue that the same can be achieved using interval-null hypotheses. However, one would then need to adjust the statement to read “there is evidence that cognitive process  $Y$  is *almost* not affected”.

cognitive models that feature many parameters, it may not be trivial to detect whether the models share instances.

SC write that their commentary is motivated by “the desire to have statistics serve science, not science serve statistics”. However, to us, it seems that their approach imposes certain constraints on how researchers can act which appears to go against the dictum advanced by SC. Suppose there are two researchers, A and B, who have different hypotheses,  $\mathcal{H}_A$  and  $\mathcal{H}_B$ , about a phenomenon of interest. These hypotheses happen to overlap. In line with the fact that “statistics should serve science” we believe that these two researchers should be allowed to compare their hypotheses in their original versions without first altering the hypotheses to the non-overlapping  $\mathcal{H}_A^*$  and  $\mathcal{H}_B^*$  to fit SCs Procrustean bed of model comparison with non-overlapping model classes. Moreover, it appears that researcher A and B would need to change their hypotheses again in case a third hypothesis  $\mathcal{H}_C$  is introduced that partially overlaps with the first two hypotheses.

## 14.5 Concluding Remarks

In this rejoinder to Vehtari, Simpson, et al. (2019), Navarro (2019), and Shiffrin and Chandramouli (2019), we have pointed out further limitations of Bayesian leave-one-out cross-validation. In particular, (1) LOO-based methods such as Bayesian stacking do not align satisfactorily with the epistemic goal of mathematical psychology; (2) LOO-based methods depend on an arbitrary distinction between parameter estimation and model comparison; and (3) LOO-based methods depend on an arbitrary distinction between data that arrive sequentially or “simultaneously”. In line with Lewandowsky and Farrell (2010) we believe that careful model comparison requires both quantitative evaluation and intellectual and scholarly judgment. We personally prefer quantitative evaluation of models based on consistently using Bayes’ rule for both parameters and models (e.g., via the Bayes factor). This approach has the advantage that, in line with the epistemic purpose of mathematical psychology, it enables the quantification of evidence for a set of competing theories that are implemented as quantitative models. Researchers may criticize the specification of an ingredient of Bayes’ rule such as the prior distribution for a particular application. However, once the ingredients have been specified, there is only one optimal way of updating one’s knowledge in light of observed data: the one that is dictated by Bayes’ rule. Alternative methods may be useful in specific circumstances and for specific purposes but – as we illustrated with the case of LOO – they will break down in other settings yielding results that can be surprising, misleading, and incoherent.

R code for reproducing the examples can be found at: <https://osf.io/eydtg/>.

## 14.A Jevons (1874) on Bayesian Model Averaging

Jevons' 1874 masterpiece *The Principles of Science* contains the section "Simple Illustration of the Inverse Problem" that showcases how BMA (for prediction) and posterior prediction are identical operations. For historical interest, and out of respect for the clarity of Jevons' writing, we present the section in full:

"Suppose it to be known that a ballot-box contains only four black or white balls, the ratio of black and white balls being unknown. Four drawings having been made with replacement, and a white ball having appeared on each occasion but one, it is required to determine the probability that a white ball will appear next time. Now the hypotheses which can be made as to the contents of the urn are very limited in number, and are at most the following five:—

4 white and 0 black balls

3	„	„	1	„	„
2	„	„	2	„	„
1	„	„	3	„	„
0	„	„	4	„	„

The actual occurrence of black and white balls in the drawings renders the first and last hypotheses out of the question, so that we have only three left to consider.

If the box contains three white and one black, the probability of drawing a white each time is  $\frac{3}{4}$ , and a black  $\frac{1}{4}$ ; so that the compound event observed, namely, three white and one black, has the probability  $\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4}$ , by the rule already given (p. 233).<sup>14</sup> But as it is indifferent to us in what order the balls are drawn, and the black ball might come first, second, third, or fourth, we must multiply by four, to obtain the probability of three white and one black in any order, thus getting  $\frac{27}{64}$ .

Taking the next hypothesis of two white and two black balls in the urn, we obtain for the same probability the quantity  $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times 4$ , or  $\frac{16}{64}$ , and from the third hypothesis of one white and three black we deduce likewise  $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} \times 4$ , or  $\frac{3}{64}$ . According, then, as we adopt the first, second, or third hypothesis, the probability that the result actually noticed would follow is  $\frac{27}{64}$ ,  $\frac{16}{64}$ , and  $\frac{3}{64}$ . Now it is certain that one or other of these hypotheses must be the true one, and their absolute probabilities are proportional to the probabilities that the

---

<sup>14</sup>The relevant text on p. 233 reads: "When the component events are independent, a simple rule can be given for calculating the probability of the compound event, thus—*Multiply together the fractions expressing the probabilities of the independent component events.*" [italics in original]

observed events would follow from them (see p. 279).<sup>15</sup> All we have to do, then, in order to obtain the absolute probability of each hypothesis, is to alter these fractions in a uniform ratio, so that their sum shall be unity, the expression of certainty. Now since  $27 + 16 + 3 = 46$ , this will be effected by dividing each fraction by 46 and multiplying by 64. Thus the probability of the first, second, and third hypotheses are respectively—

$$\frac{27}{46}, \quad \frac{16}{46}, \quad \frac{3}{46}.$$

The inductive part of the problem is now completed, since we have found that the urn most likely contains three white and one black ball, and have assigned the exact probability of each possible supposition. But we are now in a position to resume deductive reasoning, and infer the probability that the next drawing will yield, say a white ball. For if the box contains three white and one black ball, the probability of drawing a white one is certainly  $\frac{3}{4}$ ; and as the probability of the box being so constituted is  $\frac{27}{46}$ , the compound probability that the box will be so filled and will give a white ball at the next trial, is

$$\frac{27}{46} \times \frac{3}{4} \text{ or } \frac{81}{184}.$$

Again, the probability is  $\frac{16}{46}$  that the box contains two white and two black, and under those conditions the probability is  $\frac{1}{2}$  that a white ball will appear; hence the probability that a white ball will appear in consequence of that condition, is

$$\frac{16}{46} \times \frac{1}{2} \text{ or } \frac{32}{184}.$$

From the third supposition we get in like manner the probability

$$\frac{3}{46} \times \frac{1}{4} \text{ or } \frac{3}{184}.$$

Now since one and not more than one hypothesis can be true, we may add together these separate probabilities, and we find that

$$\frac{81}{184} + \frac{32}{184} + \frac{3}{184} \text{ or } \frac{116}{184}$$

is the complete probability that a white ball will be next drawn under the conditions and data supposed.” (Jevons, 1874/1913, pp. 292-294)

In the next section, *General Solution of the Inverse Problem*, Jevons points out that in order for the procedure to be applied to natural phenomena, an infinite number of hypotheses need to be considered:

---

<sup>15</sup>Note from the authors: this assumes that the hypotheses are equally likely a priori. The relevant text on p. 279 reads: “*If an event can be produced by any one of a certain number of different causes, the probabilities of the existence of these causes as inferred from the event, are proportional to the probabilities of the event as derived from these causes.*” [italics in original]

“When we take the step of supposing the balls within the urn to be infinite in number, the possible proportions of white and black balls also become infinite, and the probability of any one proportion actually existing is infinitely small. Hence the final result that the next ball drawn will be white is really the sum of an infinite number of infinitely small quantities. It might seem, indeed, utterly impossible to calculate out a problem having an infinite number of hypotheses, but the wonderful resources of the integral calculus enable this to be done with far greater facility than if we supposed any large finite number of balls, and then actually computed the results. I will not attempt to describe the processes by which Laplace finally accomplished the complete solution of the problem. They are to be found described in several English works, especially De Morgan’s ‘Treatise on Probabilities,’ in the ‘Encyclopædia Metropolitana,’ and Mr. Todhunter’s ‘History of the Theory of Probability.’ The abbreviating power of mathematical analysis was never more strikingly shown. *But I may add that though the integral calculus is employed as a means of summing infinitely numerous results, we in no way abandon the principles of combinations already treated.*[italics ours]” (Jevons, 1874/1913, p. 296)

## 14.B Coherence of BMA and Bayesian Parameter Inference

Here we show why BMA yields the same results as Bayesian parameter inference when the set of models is obtained by partitioning a continuous parameter space into non-overlapping intervals, with prior model probabilities set equal to the prior mass in the respective intervals. Given observed data  $y$ , a parameter of interest  $\theta$ ,<sup>16</sup> a corresponding prior distribution  $p(\theta)$ , and likelihood  $p(y | \theta)$ , the posterior distribution for  $\theta$  is given by

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{\int_{\Theta} p(y | \theta) p(\theta) d\theta}, \quad (14.4)$$

where  $\Theta$  denotes the parameter space. The posterior predictive distribution for new data  $y_{\text{new}}$  is given by

$$p(y_{\text{new}} | y) = \int_{\Theta} p(y_{\text{new}} | \theta, y) p(\theta | y) d\theta, \quad (14.5)$$

where it is often the case that  $p(y_{\text{new}} | \theta, y) = p(y_{\text{new}} | \theta)$ .

BMA is based on combining the results of different models based on the models’ plausibilities in light of the observed data. We consider the models  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$  that are obtained by partitioning the parameter space  $\Theta$  into  $M$  non-overlapping intervals. We denote these non-overlapping intervals by  $A_1, A_2, \dots, A_M$ . For instance, when  $\theta$  corresponds to a success probability, we

---

<sup>16</sup>Here we focus on the case of a single parameter, however, the results naturally generalize to the case where  $\theta$  is a parameter vector.

could partition  $\Theta = [0, 1]$  into two intervals  $A_1 = [0, .5]$  and  $A_2 = [.5, 1]$ . The prior distribution for  $\theta$  under each model  $\mathcal{M}_k$ ,  $k = 1, 2, \dots, M$  is obtained by considering the part of  $p(\theta)$  that corresponds to the interval  $A_k$  and then renormalizing the prior density by the prior mass in that subinterval:

$$p(\theta \mid \mathcal{M}_k) = \frac{p(\theta)}{C_k} \mathbb{I}(\theta \in A_k), \quad (14.6)$$

where  $C_k = \int_{A_k} p(\theta) d\theta$  and  $\mathbb{I}$  denotes the indicator function. Note that the  $M$  models differ only in the prior distribution for  $\theta$  but not in the likelihood, consequently  $p(y \mid \theta, \mathcal{M}_k) = p(y \mid \theta)$ . Each model's prior probability  $p(\mathcal{M}_k)$  is set equal to the prior mass that  $p(\theta)$  assigns to the interval  $A_k$ :

$$p(\mathcal{M}_k) = \int_{A_k} p(\theta) d\theta = C_k. \quad (14.7)$$

Given this set-up, the posterior probability for model  $\mathcal{M}_k$  corresponds to the posterior mass that the “regular” parameter posterior for  $\theta$  assigns to the interval  $A_k$ :

$$\begin{aligned} p(\mathcal{M}_k \mid y) &= \frac{p(y \mid \mathcal{M}_k) C_k}{\sum_{j=1}^M p(y \mid \mathcal{M}_j) C_j} \\ &= \frac{\int_{A_k} p(y \mid \theta) \frac{p(\theta)}{C_k} d\theta C_k}{\sum_{j=1}^M \int_{A_j} p(y \mid \theta) \frac{p(\theta)}{C_j} d\theta C_j} \\ &= \frac{\int_{A_k} p(y \mid \theta) p(\theta) d\theta}{\int_{\Theta} p(y \mid \theta) p(\theta) d\theta} \\ &= \int_{A_k} p(\theta \mid y) d\theta, \end{aligned} \quad (14.8)$$

where we used – in reverse order – the fact that for  $b_2 \in (b_1, b_3)$ ,  $\int_{b_1}^{b_3} f(x) dx = \int_{b_1}^{b_2} f(x) dx + \int_{b_2}^{b_3} f(x) dx$ .

The model-averaged posterior distribution for  $\theta$  is obtained as follows:

$$\begin{aligned} p_{\text{BMA}}(\theta \mid y) &= \sum_{k=1}^M p(\theta \mid y, \mathcal{M}_k) p(\mathcal{M}_k \mid y) \\ &= \sum_{k=1}^M \underbrace{\frac{p(y \mid \theta) \frac{p(\theta)}{C_k} \mathbb{I}(\theta \in A_k)}{p(y \mid \mathcal{M}_k)}}_{p(\theta \mid y, \mathcal{M}_k)} \underbrace{\frac{p(y \mid \mathcal{M}_k) C_k}{\sum_{j=1}^M p(y \mid \mathcal{M}_j) C_j}}_{p(\mathcal{M}_k \mid y)} \\ &= \frac{p(y \mid \theta) p(\theta)}{\sum_{j=1}^M p(y \mid \mathcal{M}_j) C_j} \sum_{k=1}^M \mathbb{I}(\theta \in A_k) \\ &= \frac{p(y \mid \theta) p(\theta)}{\sum_{j=1}^M \int_{A_j} p(y \mid \theta) \frac{p(\theta)}{C_j} d\theta C_j} \\ &= \frac{p(y \mid \theta) p(\theta)}{\int_{\Theta} p(y \mid \theta) p(\theta) d\theta}, \end{aligned} \quad (14.9)$$

where we used the fact that any given value for  $\theta$  will only fall in one of  $A_k$ ,  $k = 1, 2, \dots, M$ , hence,  $\sum_{k=1}^M \mathbb{I}(\theta \in A_k) = 1$ . This shows that the model-averaged posterior  $p_{\text{BMA}}(\theta | y)$  is identical to the “regular” parameter posterior (i.e., Equation 14.4).

To obtain the model-averaged predictive distribution for new data  $y_{\text{new}}$ , we first note that the predictive distribution for model  $\mathcal{M}_k$  is given by

$$\begin{aligned} p(y_{\text{new}} | y, \mathcal{M}_k) &= \int p(y_{\text{new}} | \theta, y) p(\theta | y, \mathcal{M}_k) d\theta \\ &= \int p(y_{\text{new}} | \theta, y) \underbrace{\frac{p(y|\theta) \frac{p(\theta)}{C_k} \mathbb{I}(\theta \in A_k)}{p(y|\mathcal{M}_k)}}_{p(\theta|y, \mathcal{M}_k)} d\theta \\ &= \frac{\int_{A_k} p(y_{\text{new}} | \theta, y) p(y | \theta) p(\theta) d\theta}{C_k p(y | \mathcal{M}_k)}. \end{aligned} \quad (14.10)$$

The model-averaged predictive distribution is

$$\begin{aligned} p_{\text{BMA}}(y_{\text{new}} | y) &= \sum_{k=1}^M p(y_{\text{new}} | y, \mathcal{M}_k) p(\mathcal{M}_k | y) \\ &= \sum_{k=1}^M \underbrace{\frac{\int_{A_k} p(y_{\text{new}} | \theta, y) p(y | \theta) p(\theta) d\theta}{C_k p(y | \mathcal{M}_k)}}_{p(y_{\text{new}}|y, \mathcal{M}_k)} \underbrace{\frac{p(y | \mathcal{M}_k) C_k}{\sum_{j=1}^M p(y | \mathcal{M}_j) C_j}}_{p(\mathcal{M}_k|y)} \\ &= \frac{\sum_{k=1}^M \int_{A_k} p(y_{\text{new}} | \theta, y) p(y | \theta) p(\theta) d\theta}{\sum_{j=1}^M \int_{A_j} p(y | \theta) \frac{p(\theta)}{C_j} d\theta C_j} \\ &= \frac{\int_{\Theta} p(y_{\text{new}} | \theta, y) p(y | \theta) p(\theta) d\theta}{\int_{\Theta} p(y | \theta) p(\theta) d\theta} \\ &= \int_{\Theta} p(y_{\text{new}} | \theta, y) p(\theta | y) d\theta. \end{aligned} \quad (14.11)$$

This shows that the model-averaged predictive distribution  $p_{\text{BMA}}(y_{\text{new}} | y)$  is identical to the “regular” predictive distribution (i.e., Equation 14.5).



**Part IV**

**Conclusion**



---

## Summary and Future Directions

---

In this dissertation entitled “Bayes Factor Model Comparison for Psychological Science”, rival scientific models were compared by treating them as competing forecasters and assessing their relative predictive adequacy using the Bayes factor. The first part of the dissertation was concerned with bridge sampling, a computational procedure for estimating the marginal likelihood – the key quantity for computing Bayes factors. The second part of the dissertation was concerned with Bayesian methods for meta-analyzing a set of studies. One central concept of this part was the idea to combine several forecasters using Bayesian model averaging (BMA). The third part of the dissertation introduced Bayesian approaches to a number of standard statistical tests. A central idea of this part was the incorporation of prior knowledge into the analyses to make the models’ forecasts more precise. Below, I summarize each chapter and its main conclusions accompanied by potential avenues for future development. This chapter ends with a general conclusion.

### 15.1 Part I: Bridge Sampling

#### 15.1.1 Chapter Summaries and Future Directions

Chapter 2 proposed the use of bridge sampling for estimating the marginal likelihood, the key quantity for comparing the relative predictive adequacy of competing models using the Bayes factor. Obtaining accurate estimates of this quantity is challenging, particularly for complex models that feature many parameters and are implemented in a hierarchical fashion. Using a reinforcement learning example, it was demonstrated that bridge sampling yields reliable and accurate estimates of the marginal likelihood, even for hierarchical versions of the model. Furthermore, bridge sampling is relatively straightforward to implement. Since bridge sampling requires estimating the marginal likelihood for each model separately, when the comparison of interest involves a large number of models, other methods may be more efficient. However, in contrast to bridge sampling these methods typically tend to be problem-specific. Furthermore, it was argued that most applications of

interest in mathematical psychology specifically, and psychology more generally, involve only a limited number of potentially non-nested competing models that may be implemented in a hierarchical fashion. In this setting, bridge sampling is an ideal candidate for obtaining accurate estimates of the marginal likelihood.

Chapter 3 applied Warp-III bridge sampling for comparing hierarchical multinomial processing tree (MPT) models. In contrast to the version of bridge sampling introduced in Chapter 2, Warp-III bridge sampling also takes into account potential skewness of the posterior distribution which makes this version more efficient in case the posterior is indeed asymmetrical. Using a nested and a non-nested example it was demonstrated how this approach enables researchers to address concrete questions of interest. Specifically, the first example applied Bayesian model averaging to assess which MPT parameters differ across trials in a pair-clustering experiment whereas the second example compared two structurally different MPT models concerning the illusory truth effect. One central advantage of Bayesian model comparison is the ability to disentangle *absence of evidence* (i.e., the data are inconclusive) from *evidence of absence* (i.e., the data support an invariance). It was argued that it is crucial in all stages of cognitive model development, validation, and application that one is able to quantify evidence in favor of invariances. When developing and validating a model one key step is to show that certain experimental manipulations affect only a subset of the model parameters, but also to show that the remaining parameters are unaffected (i.e., demonstrate selective influence). Furthermore, in applications, a researcher may wish to make statements of the form “there is evidence that cognitive process  $X$  is not affected” which again corresponds to quantifying evidence for an invariance. An avenue for future development is to apply Warp-III bridge sampling to hierarchical MPT models that feature random effects not only for participants, but also items (Matzke et al., 2015).

Chapter 4 applied Warp-III bridge sampling for computing the marginal likelihood of evidence-accumulation models. Obtaining posterior samples for evidence-accumulation models such as the Linear Ballistic Accumulator (LBA) is challenging and requires specialized sampling algorithms such as differential evolution Markov chain Monte Carlo (DE-MCMC). It was demonstrated that, in combination with DE-MCMC, Warp-III bridge sampling provides precise estimates of the marginal likelihood for both single-participant and hierarchical versions of the LBA. To facilitate the practical application of this Bayesian model comparison approach for evidence-accumulation models, an easy-to-use software implementation has been provided. The chapter concluded with a series of recommendations for applying Warp-III bridge sampling in practical applications. Many of these recommendations were aimed at assessing whether the estimate of the marginal likelihood is precise enough to draw meaningful conclusions. For many of the hierarchical examples in this chapter, the Bayes factors provided overwhelming evidence in favor of one model, so that some fluctuation in the marginal likelihood estimates may not alter the overall conclusions. However, the final example also demonstrated that Warp-III bridge sampling can provide precise estimates of the marginal likelihoods and hence also the Bayes factor when there is about equal evidence in favor of both hierarchical LBA models of interest. In this case, it is crucial that the estimates are precise in order to avoid the erroneous conclusion

that one model is favored when in fact this is an artefact of a variable estimate.

Chapter 5 applied Bayesian methods to infer the appropriate number of dimensions and the metric structure of multidimensional scaling (MDS) models. Priors were defined for making the model identifiable under metrics corresponding to psychologically separable and psychologically integral stimulus domains. Obtaining high quality posterior samples is challenging for MDS models and in this chapter, DE-MCMC was used for this task. Warp-III bridge sampling was applied to identify the appropriate number of dimensions and to infer the appropriate metric of the latent space. A series of examples demonstrated that the procedure provides sensible results for many data sets. However, it was also pointed out that there are a number of challenges that need to be addressed before the method can be applied in a general, straightforward manner. Most importantly, for certain examples, it can be difficult to obtain high quality posterior samples, even using DE-MCMC. When the posterior samples are not of a high quality, this will also negatively affect the precision of the Warp-III bridge sampling estimate. Therefore, in the future it should be investigated whether posterior sampling algorithms other than DE-MCMC are better suited for obtaining high quality samples for MDS models in a reliable manner.

Chapter 6 introduced **bridgesampling**, an R package for estimating the marginal likelihood (or, more generally, normalizing constants) using bridge sampling in a generic and easy-to-use fashion. Specifically, the package enables the computation of the marginal likelihood for any model for which one can provide posterior samples, a function that computes the log of the unnormalized posterior density for a set of model parameters, the data, and lower and upper bounds for the parameters. When the model of interest is implemented in **Stan** (Carpenter et al., 2017), the computation of the marginal likelihood is automatic: one simply needs to pass the object with the posterior samples to the **bridgesampling** package. Thus, the package makes it possible to obtain marginal likelihood estimates for any model that can be implemented in **Stan** (in a way that retains the constants). Adding support for a similar automatic computation of the marginal likelihood for other Bayesian sampling software such as **JAGS** (Plummer, 2003) is a worthwhile future task. In fact, a similar support has already been added for the **nimble** package (de Valpine et al., 2017).

### 15.1.2 Discussion

One reason why Bayesian model comparison approaches have not been applied more widely so far may be that it can be difficult to compute the quantities of interest. This is particularly true when the comparison involves hierarchical models. Bridge sampling provides a computational resolution to this challenge that yields precise estimates of the marginal likelihood, even for hierarchical models. One benefit of bridge sampling is that it is relatively generic and simple to apply in different applications. These characteristics enabled the development of the **bridgesampling** package that can provide an automatic estimate of the marginal likelihood. The package implements two versions of bridge sampling. The first is based on a proposal distribution that matches the mean and covariance of the posterior samples, the second, Warp-III bridge sampling, additionally takes

into account skewness. These two version of bridge sampling can provide precise estimates for many applications of interest. However, in case the posterior samples exhibit multiple modes, both of these versions may be inefficient. Developing efficient bridge sampling versions for these challenging cases is the topic of ongoing research (e.g., Nott, Kohn, & Fielding, 2008; L. Wang & Meng, 2016).

Bridge sampling provides an estimate of the marginal likelihood. To use this estimate in practice, it is important to confirm that the estimate is precise enough to address the question of interest. There exist quick approximate methods for assessing the variability of the estimate (Frühwirth-Schnatter, 2004), however, they do not always appear suitable, particularly not for Warp-III bridge sampling. Therefore, the best way of assessing the variability is repeating the bridge sampling procedure multiple times, a process that can be time-consuming. Developing reliable error estimates that are quick to compute, also for Warp-III bridge sampling, is an avenue for future development.

Bridge sampling provides a solution to the difficult task of computing the marginal likelihood and therefore allows researchers to spend more time on the conceptual challenge of specifying prior distributions that are robust or meaningful. Specifically, it is important that bridge sampling is not applied blindly but that researchers carefully consider their prior choices which directly affect what data the models can predict. There are a number of approaches to specifying sensible prior distributions (e.g., Bayarri et al., 2012; Lee & Vanpaemel, 2018). One way of alleviating concerns about the influence of particular choices is to simply conduct a prior sensitivity analysis as has been done, for instance, in Chapter 3.

Finally, it could be argued that it might be dangerous to provide researchers with tools that could be used in an incorrect way. It is true that researchers could use the `bridgesampling` package to conduct meaningless model comparisons in case the priors are pathological, for instance, in case they do not correspond to what the researchers actually want to test. However, it could also be argued that developing such tools is crucial since they allow researchers to spend more time on thinking about the more conceptual rather than the computational challenges. Furthermore, only when one can compute the marginal likelihood for a model of interest one can check how much different prior choices actually affect the results.

## 15.2 Part II: Multi-Model Meta-Analysis

### 15.2.1 Chapter Summaries and Future Directions

Chapter 7 proposed a two-component Bayesian mixture model for meta-analyzing the distribution of significant  $p$  values of a set of studies. One component corresponds to the null hypothesis of no effect (i.e.,  $\mathcal{H}_0$ ), the other component corresponds to the alternative hypothesis which states that an effect is present (i.e.,  $\mathcal{H}_1$ ). This mixture model allows researchers to estimate the proportion of significant results that originate from  $\mathcal{H}_0$ . Additionally, the mixture model provides an estimate of the probability that each specific  $p$  value originates from  $\mathcal{H}_0$ . The procedure was demonstrated using two examples and a web application has been provided to facilitate the application of the method in practice. It was pointed out that even with many  $p$  values available, the results can be affected by the choice

of the prior distributions. Therefore, it was recommended to conduct a prior sensitivity analysis to assess the effect of different prior choices on the results. It was also noted that the distribution of significant  $p$  values originating from  $\mathcal{H}_1$  is represented by a simple parametric distribution. The distribution of significant  $p$  values originating from  $\mathcal{H}_1$  can be complex so that this parametric representation could be too simplistic, at least for certain examples. Therefore, a potential avenue for future development is to explore alternative non-parametric approaches for representing the  $p$  value distribution under  $\mathcal{H}_1$ . One such non-parametric approach (i.e., a Dirichlet process mixture) has already been explored when developing the chapter. However, unfortunately, the specific approach that was explored made the model challenging to estimate and simulations suggested that it cannot easily be applied across sets of  $p$  values with different characteristics.

Chapter 8 introduced Bayesian model-averaged meta-analysis. This procedure implements the idea of combining several forecasters according to their plausibility in light of the observed data to avoid an all-or-none decision between a fixed-effect and a random-effects meta-analysis model. Specifically, four Bayesian meta-analysis models are considered simultaneously: (1) fixed-effect null hypothesis, (2) fixed-effect alternative hypothesis, (3) random-effects null hypothesis, and (4) random-effects alternative hypothesis. This approach allows researchers to address, in a principled manner, the two key questions “Is the overall effect non-zero?” and “Is there between-study variability in effect size?”. The procedure was demonstrated using an example concerning the self-concept maintenance theory. Prior recommendations were provided for standardized mean difference effect sizes. An avenue for future development is to provide prior recommendations for other effect sizes such as Fisher’s  $z$  and log odds ratios. Furthermore, it would be interesting to explore how the idea that subsets of studies might have different latent effect sizes could be incorporated in this Bayesian model averaging framework. One possibility is to specify a latent mixture of normal distributions for the effect sizes (e.g., Moreau & Corballis, 2019). One could then take into account model versions with different numbers of latent components and combine them again for final inference using Bayesian model averaging.

Chapter 9 applied the Bayesian model-averaged meta-analysis introduced in Chapter 8 to analyze six preregistered studies concerning the effect of power posing. Specifically, the meta-analysis focused on the effect of power posing on felt power. There was strong evidence for an effect of power posing on felt power, however, the evidence was only moderate when considering only participants that were unfamiliar with the effect. An avenue for future development is to adjust the Bayesian model-averaged meta-analysis to test this potential participant-level moderator directly.

### 15.2.2 Discussion

The second part of this dissertation presented different Bayesian methods for meta-analyzing a set of studies. Just as with existing meta-analysis procedures, the adage “garbage in, garbage out” also applies to these Bayesian methods. With low quality data, no statistical analysis can provide high quality inference. However, recently it has become more common to conduct preregistered studies, for instance,

in the form of *Registered Reports* (Chambers, 2013; Chambers et al., 2015) or *Registered Replication Reports* (e.g., Wagenmakers, Beek, et al., 2016) which are free from publication bias and unaffected by cherry-picking. These high quality data greatly facilitate the application of the proposed methods. Concretely, when applying the Bayesian mixture model from Chapter 7, one aspect that makes estimating the parameters challenging is that, in its current version, the model considers only significant  $p$  values. Therefore, all statistical action is in the tail of the distribution where it is difficult to disentangle the two mixture components. This has the consequence that a relatively large number of  $p$  values is required to obtain reliable estimates of the model parameters. However, when meta-analyzing a set of preregistered studies one could adjust the model to consider the complete distribution of  $p$  values since publication bias can be ruled out. This may greatly facilitate estimating the model parameters, even based on a smaller number of  $p$  values. Naturally, having available a set of studies free of publication bias also greatly benefits the application of the Bayesian model-averaged meta-analysis presented in Chapter 8 and Chapter 9. One may of course also use the procedure to analyze studies that have not been preregistered; however, the conclusions need to be interpreted with scepticism if it cannot be ruled out that the included studies represent a biased sample of all conducted studies in the field. In contrast, if the included studies have been preregistered, Bayesian model-averaged meta-analysis can be a valuable tool that allows researchers to address key questions of interest in a principled manner.

## 15.3 Part III: Hypothesis Testing

### 15.3.1 Chapter Summaries and Future Directions

Chapter 10 demonstrated how Bayesian inference can be used to quantify the evidence in favor of a general law based on finite data. Specifically, the chapter focused on quantifying evidence in favor of the hypothesis that each digit in the decimal expansions of  $\pi$ ,  $e$ ,  $\sqrt{2}$ , and  $\ln 2$  occurs equally often. The evidence in favor of the general law was overwhelming for all four constants. This analysis provided a concrete demonstration that Bayesian inference can be used sequentially to update one's knowledge as new data become available over time. Importantly, this process can be continued indefinitely without invalidating the results as can be the case when using frequentist inference based on  $p$  values (e.g., Berger & Wolpert, 1988; Rouder, 2014). Furthermore, the chapter illustrated that what can be called the *second derivative of belief* – the change in the Bayes factor as a result of new data – becomes insensitive to the prior specification as the number of observations grows large. Therefore, although the results based on different priors were not identical quantitatively (but they were qualitatively), the evidential trajectories for all prior choices suggested that the evidence for the general law increases as more digits become available.

Chapter 11 proposed the use of a flexible  $t$ -prior for effect size in the Bayesian  $t$ -test. This approach enables researchers to incorporate prior knowledge into the analysis to make their predictions more precise. The proposed prior specification contains previous subjective Bayesian  $t$ -test versions, but also objective ones. It

can therefore be used to incorporate prior knowledge, but also to conduct a default Bayesian  $t$ -test in case strong prior knowledge is absent. Two measures for informed prior distributions were proposed that quantify the departure from the objective Bayes factor desiderata of predictive matching and information consistency. One possible application of these departure measures is issuing recommendations for researchers who would like to incorporate expert knowledge into the prior specification, but would also like to retain Jeffreys's desiderata as much as possible. For instance, specifying a  $t$ -prior with one degree of freedom (i.e., a Cauchy distribution) ensures that information consistency holds. Crucially, this is also the case if this Cauchy prior is centered on a value other than zero which enables one to incorporate expert knowledge about effect size. Researchers who want to retain predictive matching should specify the prior to be centered on zero, but can freely choose the scale parameter and the degrees of freedom. The proposed Bayesian  $t$ -test using a flexible  $t$ -prior was demonstrated using an example concerning the facial feedback hypothesis that featured an expert prior elicitation effort.

Chapter 12 introduced **abtest**, an R package for conducting Bayesian A/B tests based on work by Kass and Vaidyanathan (1992). This Bayesian hypothesis testing approach comes with the well-known advantages of allowing researchers to (1) obtain evidence in favor of the null hypothesis that the treatment is ineffective, (2) monitor the evidence as the data accumulate, (3) take into account expert prior knowledge. Specifically, the implemented Bayesian A/B test enables one to monitor the evidence for the hypotheses that the treatment has either a positive effect, a negative effect, or, crucially, no effect. Furthermore, this method also allows one to incorporate expert knowledge about the relative prior plausibility of the rival hypotheses as well as about the expected size of the effect. An avenue for future development is to extend the package functionality so that it is possible to simultaneously compare more than two groups.

Chapter 13 discussed Bayesian leave-one-out cross-validation (LOO), an alternative method for comparing competing models. Using three concrete examples, it was illustrated that when a data set of infinite size is perfectly in line with the predictions of a simple model and this model is compared to a more complex model, LOO shows bounded support for the simple model. Importantly, (1) this limiting bound of support is often surprisingly low, (2) the LOO-preference for the simple model may be a non-monotonic function of the number of observations (i.e., additional observations perfectly in line with the simple model may in fact *decrease* the LOO-preference for the simple model), (3) the LOO result can depend strongly on the prior choice, even asymptotically. Therefore, it was concluded that LOO is not a panacea for model selection.

Chapter 14 addressed three commentaries on Chapter 13 and identified additional limitations of methods that are based on LOO (such as Bayesian stacking). Specifically, (1) LOO-based methods such as Bayesian stacking do not align satisfactorily with the epistemic goal of mathematical psychology, (2) LOO-based methods depend on an arbitrary distinction between parameter estimation and model comparison, and (3) LOO-based methods depend on an arbitrary distinction between data that arrive sequentially or “simultaneously”. It was argued in favor of using Bayes' rule consistently, for both parameter estimation and model

comparison. Alternative methods such as LOO may be useful in specific circumstances and for specific purposes but, as illustrated, will yield results that are surprising, misleading, and incoherent in other settings.

### 15.3.2 Discussion

The third part of this dissertation introduced Bayesian hypothesis testing approaches to a number of standard statistical tests and discussed Bayesian leave-one-out cross-validation, an alternative method for comparing competing models. The introduced approaches come with the well-known pragmatic benefits of Bayesian hypothesis testing (e.g., Wagenmakers, Marsman, et al., 2018; Wagenmakers, Morey, & Lee, 2016) such as enabling researchers to (1) obtain evidence in favor of the null hypothesis, (2) monitor the evidence sequentially as the data accumulate, (3) incorporate existing prior knowledge into the analyses. The advantages of the last point were already illustrated in the introduction of this dissertation: specifying informed prior distributions can result in more precise predictions which will be rewarded in case they turn out to be true. When designing and conducting experiments, researchers rely on prior knowledge to make the best possible choices. A natural next step is to apply the idea of cumulative scientific learning also on the level of the statistical analysis and incorporate existing knowledge in the form of an informed prior distribution.

There exist alternative approaches to Bayesian model comparison that do not rely on computing Bayes factors. One of these is Bayesian leave-one-out cross-validation (LOO). Applying this approach effectively means that parameter estimation is approached in a very different way than model comparison. Specifically, to estimate the parameters one simply applies Bayes' rule whereas to compare models Bayesian cross-validation is conducted. However, the distinction between parameter estimation and model comparison is in some sense artificial (consider, e.g., models with discrete parameters) and treating them differently results in certain inconsistencies, as has been demonstrated. These inconsistencies can be avoided by approaching parameter estimation and model comparison in the same way – by systematically applying Bayes' rule.

## 15.4 General Conclusion

To conclude this dissertation, a concrete example is used to illustrate how the different parts of this dissertation can be applied in combination to address practical questions of interest. The example is a reanalysis of the Registered Replication Report concerning the facial feedback hypothesis (Wagenmakers, Beek, et al., 2016) and has been presented in Hinne et al. (2020). The facial feedback hypothesis states that affective responses can be influenced by one's facial expression even when that facial expression is not the result of an emotional experience. Strack et al. (1988) reported that participants who held a pen between their teeth (inducing a facial expression similar to a smile; see Figure 15.1, left panel) rated cartoons as more funny on a 10-point Likert scale ranging from 0-9 than participants who held a pen with their lips (inducing a facial expression similar to a

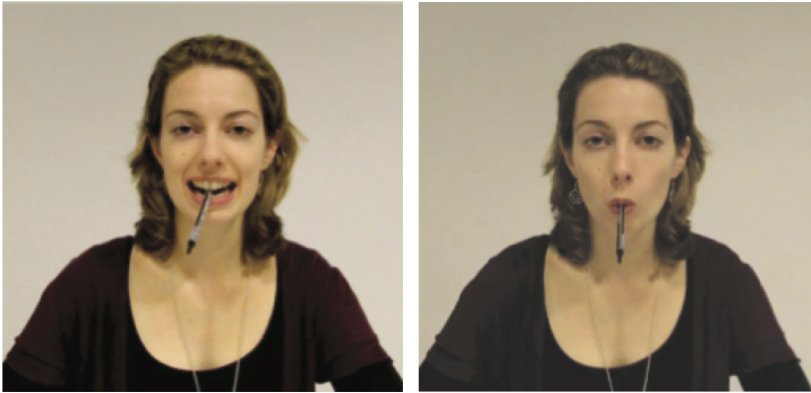


Figure 15.1: Illustration of two ways of holding a pen in a facial feedback study (see also Wagenmakers, Beek, et al., 2016). Left panel: the pen is held with the teeth, inducing a facial expression similar to a smile. Right panel: the pen is held with the lips, inducing a facial expression similar to a pout. Available at <http://tinyurl.com/zm7p917> under CC license <https://creativecommons.org/licenses/by/2.0/>.

pout; see Figure 15.1, right panel). In a Registered Replication Report that featured data from 17 labs, Wagenmakers, Beek, et al. (2016) reported a classical random-effects meta-analysis estimate of the mean difference between the “smile” and “pout” condition equal to 0.03 [95% CI: -0.11, 0.16]. Furthermore, for all labs individually, default independent samples *t*-test Bayes factors indicated evidence in favor of the null hypothesis and for 13 out of the 17 labs, the Bayes factor in favor of the null hypothesis was larger than 3.

Figure 15.2 displays the results of a Bayesian reanalysis of these replication data (see also Hinne et al., 2020). Based on work presented in the third part of this dissertation (i.e., Chapter 11), for each of the 17 labs separately, an informed independent samples *t*-test Bayes factor is displayed, accompanied by an estimate of the lab’s effect size plus 95% Bayesian uncertainty interval.<sup>1</sup> Specifically, the informed “Oosterwijk prior” from Chapter 11 was used for effect size: a *t* distribution with location 0.35, scale 0.102, and 3 degrees of freedom. For 15 out of the 17 labs, the informed *t*-test Bayes factor indicates evidence in favor of the null hypothesis and for 10 out of the 17 labs the Bayes factor in favor of the null hypothesis is larger than 3. For both labs for which the Bayes factor does not indicate evidence in favor of the null hypothesis, the Bayes factor in favor of an effect is smaller than 2.

The lower part of Figure 15.2 displays the results of a Bayesian model-averaged meta-analysis as has been introduced in the second part of this dissertation. Specifically, below the results of a Bayesian fixed-effect and a Bayesian random-effects

<sup>1</sup>For testing, the prior on effect size was truncated below at zero, whereas for estimation this truncation was removed.

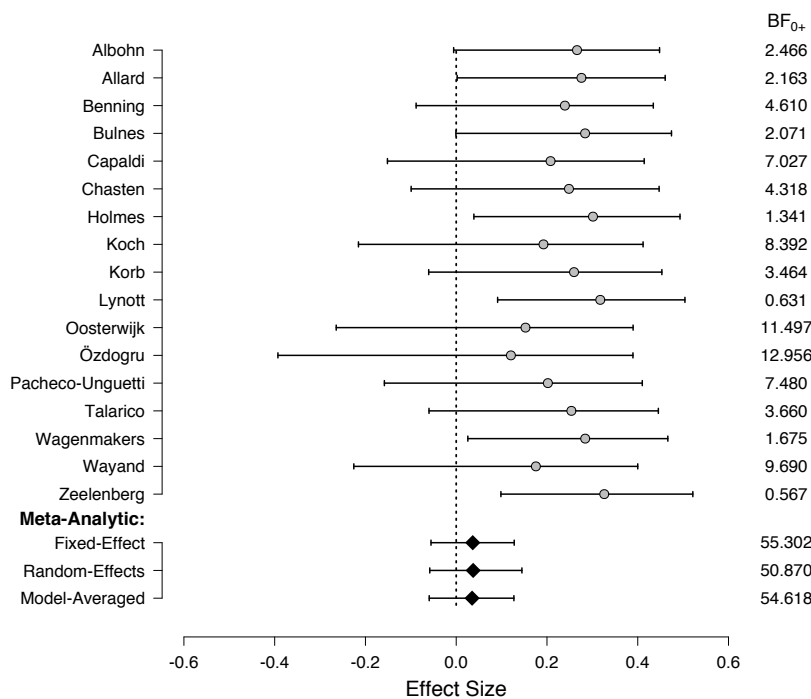


Figure 15.2: Results of the reanalysis of the facial feedback Registered Replication Report data (Wagenmakers, Beek, et al., 2016; see also Hinne et al., 2020). The upper part displays for each lab separately an informed Bayesian independent samples *t*-test Bayes factor, accompanied by an estimate of the lab’s effect size plus 95% Bayesian uncertainty interval. The lower part displays the results of a Bayesian fixed-effect, a Bayesian random-effects, and a Bayesian model-averaged meta-analysis. Available at <https://tinyurl.com/y4ocdpjf> under CC license <https://creativecommons.org/licenses/by/2.0/>.

meta-analysis, the model-averaged results are displayed which combine the results of the Bayesian fixed-effect and Bayesian random-effects meta-analysis according to their plausibility in light of the observed data. To obtain the meta-analytic Bayes factors of interest, one needs to compute the marginal likelihood for all meta-analysis models under consideration. This can be challenging, particularly, for the hierarchical random-effects model. One computational resolution is to use bridge sampling, as has been presented in the first part of this dissertation. Furthermore, when implementing the Bayesian meta-analysis models one needs to choose a prior for the between-study heterogeneity parameter  $\tau$ . As has been recommended in Chapter 8, an informed Inverse-Gamma(1, 0.15) prior is used that is based on the distribution of non-zero between-study standard deviation estimates for standardized mean difference effect sizes from meta-analyses reported in *Psy-*

*chological Bulletin* in the years 1990–2013 (van Erp et al., 2017). The Bayesian model-averaged meta-analytic Bayes factor in favor of the null hypothesis is equal to 54.618, indicating very strong evidence in favor of the facial feedback effect being absent.

This example demonstrated how ideas and methods from the three parts of this dissertation can be used to address practical questions of interest. Furthermore, the example also provided a concrete demonstration of a few general advantages of Bayesian hypothesis testing using Bayes factors. Specifically, it demonstrated that Bayes factors can be used to quantify evidence in favor of the null hypothesis of no effect and that existing prior knowledge can be incorporated into the statistical analysis. Additionally, in case new facial feedback replications become available in the future, the presented results can be updated based on these new data which highlights that Bayesian inference can be applied sequentially to update one's knowledge as new information becomes available.

In sum, in this dissertation, I hope to have provided tools for researchers interested in applying Bayesian inference to their own data that facilitate addressing questions of interest in a principled and easy-to-use manner.



---

## References

---

- Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, 32, 1376–1402.  
36, 37
- Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., & Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, 4, 95–110.  
15, 41
- Aitkin, M. (2001). Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, 1, 287–304.  
79
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.  
56, 302
- Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, S., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered Replication Report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9, 556–578.  
183
- Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94, 443–458.  
45
- Ando, T. (2010). *Bayesian model selection and statistical modeling*. Boca Raton, FL: CRC Press.  
302
- Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, 66, 1–7.  
13
- Apgar, J. F., Witmer, D. K., White, F. M., & Tidor, B. (2010). Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, 6, 1890–1900.

- 82
- Aragón Artacho, F. J. A., Bailey, D. H., Borwein, J. M., & Borwein, P. B. (2012). Walking on real numbers. *The Mathematical Intelligencer*, 35, 42–60.
- 236
- Armitage, P. (1960). *Sequential medical trials*. Springfield (IL): Thomas.
- 264
- Arnold, J. B. (1971). A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology*, 90, 349–372.
- 121
- Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology*, 63, 516–556.
- 115
- Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2017). Could a woman be superman? Gender and the embodiment of power postures. *Comprehensive Results in Social Psychology*, 2, 6–27.
- 214
- Bailey, D. H., & Borwein, J. M. (2009). Experimental mathematics and computational statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1, 12–24.
- 233, 236, 239
- Bailey, D. H., Borwein, J. M., Calude, C. S., Dinneen, M. J., Dumitrescu, M., & Yee, A. (2012). An empirical approach to the normality of  $\pi$ . *Experimental Mathematics*, 21, 375–384.
- 236
- Bailey, D. H., & Crandall, R. E. (2001). On the random character of fundamental constant expansions. *Experimental Mathematics*, 10, 175–190.
- 236
- Bakker, R., & Poole, K. T. (2013). Bayesian metric multidimensional scaling. *Political Analysis*, 21, 125–140.
- 119, 120
- Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. New York: Pergamon Press Inc.
- 183
- Bark, R., Dieckmann, S., Bogerts, B., & Northoff, G. (2005). Deficit in decision making in catatonic schizophrenia: An exploratory study. *Psychiatry Research*, 134, 131–141.
- 36
- Barker, R. J., & Link, W. A. (2013). Bayesian multimodel inference by RJMCMC: A Gibbs sampling approach. *The American Statistician*, 67, 150–156.
- 90, 109, 110
- Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44, 533–534.
- 268
- Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, 87, 375–397.
- 58

- Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, 97, 548–564.  
56
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86.  
56
- Bateman, H., Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G., Bertin, D., ... Stampfel, R. (1953). *Higher transcendental functions* (Vol. 2). McGraw-Hill New York.  
255
- Bateman, H., Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G., Bertin, D., ... Stampfel, R. (1954). *Tables of integral transforms* (Vol. 1). McGraw-Hill.  
255
- Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72, 90–103.  
13
- Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40, 1550–1577.  
77, 84, 92, 166, 252, 256, 261, 303, 346
- Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7–15.  
16, 35, 36, 45
- Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience*, 19, 5473–5481.  
35
- Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation of working memory from decision making within the human prefrontal cortex. *Journal of Neuroscience*, 18, 428–437.  
35
- Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275, 1293–1295.  
36
- Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, 123, 2189–2202.  
35
- Becker, B. J. (1991). Small-sample accuracy of approximate distributions of functions of observed probabilities from *t* tests. *Journal of Educational Statistics*, 16, 345–369.  
172, 173

- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.  
308
- Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22, 245–268.  
13, 16
- Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 1 (2nd ed.) (pp. 378–386). Hoboken, NJ: Wiley.  
15
- Berger, J. O., & Berry, D. A. (1988a). The relevance of stopping rules in statistical inference. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics: Vol. 4* (pp. 29–72). New York: Springer Verlag.  
235
- Berger, J. O., & Berry, D. A. (1988b). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.  
235
- Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2, 317–352.  
265
- Berger, J. O., & Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59, 3–15.  
14
- Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In P. Lahiri (Ed.), *Model selection* (pp. 135–207). Beachwood, OH: Institute of Mathematical Statistics Lecture Notes—Monograph Series, volume 38.  
252
- Berger, J. O., Pericchi, L. R., & Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, 60, 307–321.  
252
- Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward (CA): Institute of Mathematical Statistics.  
205, 265, 348
- Berman, R., Pekelis, L., Scott, A., & Van den Bulte, C. (2018). p-hacking and false discovery in A/B testing. *SSRN*. Retrieved from <http://dx.doi.org/10.2139/ssrn.3204791>  
264
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.  
320, 321, 322
- Bishop, Y. M., Fienberg, S., & Holland, P. (Eds.). (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.  
56
- Blair, R. J. R., Colledge, E., & Mitchell, D. G. V. (2001). Somatic markers and response reversal: Is there orbitofrontal cortex dysfunction in boys with

- psychopathic tendencies? *Journal of Abnormal Child Psychology*, 29, 499–511.  
36
- Böckenholt, U. (2012a). The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika*, 77, 388–399.  
56
- Böckenholt, U. (2012b). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17, 665–678.  
56
- Boehm, U., Hawkins, G. E., Brown, S. D., van Rijn, H., & Wagenmakers, E.-J. (2016). Of monkeys and men: Impatience in perceptual decision-making. *Psychonomic Bulletin & Review*, 23, 738–749.  
73
- Boehm, U., Steingroever, H., & Wagenmakers, E.-J. (2018). Using Bayesian regression to test hypotheses about relationships between parameters and covariates in cognitive models. *Behavior Research Methods*, 50, 1248–1269.  
80
- Bombardi, D., Schmid Mast, M., & Pulfrey, C. (2017). Real and imagined power poses: Is the physical experience necessary after all? *Comprehensive Results in Social Psychology*, 2, 44–54.  
214
- Borel, E. (1909). Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27, 247–271.  
236
- Borel, E. (Ed.). (1965). *Elements of the theory of probability*. Englewood Cliffs, NJ: Prentice-Hall.  
234
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.  
190
- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.  
114, 118, 121, 122, 129, 133
- Bortz, J. (1974). Kritische Bemerkungen über den Einsatz nichteuklidischer Metriken im Rahmen der multidimensionalen Skalierung. *Archiv für Psychologie*, 126, 196–212.  
121, 134
- Borwein, J. M., Bailey, D. H., & Bailey, D. (2004). *Mathematics by experiment: Plausible reasoning in the 21st century*. Natick, MA: AK Peters.  
236
- Brooks, S. B., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.  
93

- Brown, C. E. (1998). Coefficient of variation. In *Applied multivariate statistics in geohydrology and related sciences* (pp. 155–157). Berlin, Heidelberg: Springer.  
35
- Brown, K. S., & Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68, 021904.  
80
- Brown, P. E., & Zhou, L. (2018). glmmBUGS: Generalised linear mixed models and spatial models with WinBUGS, JAGS, and OpenBUGS [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=glmmBUGS> (R package version 2.4.2)  
166
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, 57, 153–178.  
80, 82, 103
- Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44, 108–132.  
84, 302
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80, 1–28.  
280
- Bürkner, P.-C., Vehtari, A., & Gabry, J. (2018). *Approximate leave-future-out cross-validation for time series models*. Retrieved from <http://mc-stan.org/loo/articles/loo2-lfo.html>  
329
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.)*. New York: Springer Verlag.  
56, 302, 304, 314
- Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, 14, 253–262.  
16, 35, 36, 37, 38, 40, 41, 43, 45
- Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57, 473–484.  
16, 46, 143
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief non-verbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21, 1363–1368.  
213, 214
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and summary of research on the embodied effects of expansive (vs. contractive) nonverbal displays. *Psychological Science*, 26, 657–663.  
214

- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.  
8, 145, 155, 280, 345, 396
- Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), *Multidimensional scaling: Theory and application in the behavioral sciences* (pp. 105–155). New York: Seminar Press.  
136
- Carroll, J. D., & Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of ‘Eckart–Young’ decomposition. *Psychometrika*, 35, 283–319.  
136
- Carroll, J. D., & Wish, M. (1974). Multidimensional perceptual models and measurement methods. *Handbook of Perception*, 2, 391–447.  
129
- Cavedini, P., Riboldi, G., D’Annuncci, A., Belotti, P., Cisima, M., & Bellodi, L. (2002). Decision-making heterogeneity in obsessive-compulsive disorder: Ventromedial prefrontal cortex function predicts different treatment outcomes. *Neuropsychologia*, 40, 205–211.  
36
- Cavedini, P., Riboldi, G., Keller, R., D’Annuncci, A., & Bellodi, L. (2002). Frontal lobe dysfunction in pathological gambling patients. *Biological Psychiatry*, 51, 334–341.  
36
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49, 609–610.  
79, 172, 183, 348
- Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction*, 110, 10–11.  
79, 172
- Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, 66, A1–A2.  
183, 348
- Chambers, C. D., Munafo, M., & et al. (2013). Trust in science would be improved by study pre-registration. *The Guardian*.  
189
- Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*®, 39, 860–864.  
268
- Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2002). *Monte Carlo methods in Bayesian computation*. New York: Springer.  
16
- Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, S., ... Yon, J. C. (2016). Registered Replication Report: Study 1 from

- Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, 11, 750–764.  
183
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313–1321.  
143
- Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96, 270–281.  
47
- Choirat, C., & Seri, R. (2012). Estimation in discrete parameter models. *Statistical Science*, 27, 278–293.  
324
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.  
302
- Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, 29, 1165–1175.  
127, 128, 138
- Coomarasamy, A., Devall, A. J., Cheed, V., Harb, H., Middleton, L. J., Gallos, I. D., ... Jurkovic, D. (2019). A randomized trial of progesterone in women with bleeding in early pregnancy. *New England Journal of Medicine*, 380, 1815–1824.  
298
- Corter, J. E. (1996). *Tree models of similarity and association*. Thousand Oaks, CA: Sage.  
135
- Cox, T. F., & Cox, M. A. A. (1991). Multidimensional scaling on a sphere. *Communications in Statistics: Theory and Methods*, 20, 2943–2953.  
135
- Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. London: Chapman and Hall.  
114
- Culpepper, S. A. (2014). If at first you don't succeed, try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement*, 38, 632–644.  
56
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.  
266
- Cushny, A. R., & Peebles, A. R. (1905). The action of optical isomers: II hyoscines. *The Journal of Physiology*, 32, 501–510.  
151, 152
- Dai, J., Kerestes, R., Upton, D. J., Busemeyer, J. R., & Stout, J. C. (2015). An improved cognitive model of the Iowa and Soochow gambling tasks with regard to model fitting performance and tests of parameter consistency. *Frontiers in Psychology*, 6:229.

- 36
- Dalal, S. R., & Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45, 278–286.  
243, 245
- Dawid, A. P. (1970). On the limiting normality of posterior distributions. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 67, pp. 625–633).  
79
- Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, 147, 278–292.  
3, 329, 330
- de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26, 403–413.  
144, 345
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48, 1–28.  
56
- Deng, A., Lu, J., & Chen, S. (2016). Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *2016 IEEE international conference on data science and advanced analytics* (pp. 243–252).  
265
- Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71, 1–25.  
144
- DiCiccio, T. J., Kass, R. E., Raftery, A. E., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92, 903–915.  
16, 24, 25, 45, 58, 63
- Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42, 204–223.  
35, 43, 46
- Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, 41, 214–226.  
35, 43, 46, 57, 71, 85, 109, 166, 250
- Didelot, X., Everitt, R. G., Johansen, A. M., & Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6, 49–76.  
14
- Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 363–375.  
175

- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5:781.  
264
- Donkin, C., & Brown, S. D. (2018). Response times and decision making. In *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 349–377). John Wiley & Sons, Inc.  
82
- Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, 16, 1129–1135.  
91
- Doxas, I., Dennis, S., & Oliver, W. L. (2010). The dimensionality of discourse. *Proceedings of the National Academy of Sciences*, 107, 4866–4871.  
303
- Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., & Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40, 1–18.  
70, 78
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.  
330
- Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., ... Prenoveau, J. M. (2016). Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, 11, 158–171.  
183
- Efron, B. (2012). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99, 96–104.  
173
- Eich, E. (2014). Business not as usual. *Psychological Science*, 25, 3–6.  
172
- Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, 38, 467–474.  
130
- Ennis, D. M. (1992). Modeling similarity and identification when there are momentary fluctuations in psychological magnitudes. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 279–298). Hillsdale, NJ: Erlbaum.  
136
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie*, 217, 108–124.  
56
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, 53, 134–140.  
118
- Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32, 313–329.

- 2, 15, 57, 84, 102, 142, 197, 252, 305
- Evans, N. J., & Annis, J. (2019). Thermodynamic integration via differential evolution: A method for estimating marginal likelihoods. *Behavior Research Methods*, *51*, 930–947.
- 85
- Evans, N. J., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, *50*, 589–603.
- 85, 86, 93, 95
- Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge: Cambridge University Press.
- 82
- Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*, 1209–1217.
- 82
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002.
- 56, 73, 74, 75
- Fechner, G. T. (1966 [1860]). (H. E. Adler, Trans.). In D. H. Howes & E. G. Borning (Eds.), *Elements of psychophysics [Elemente der Psychophysik]* (Vol. 1). New York: Holt, Rinehart and Winston.
- 129
- Feldman, J. (2013). Tuning your priors to the world. *Topics in Cognitive Science*, *5*, 13–34.
- 323
- Feldman, J. (2015). Bayesian inference and “truth”: A comment on Hoffman, Singh, and Prakash. *Psychonomic Bulletin & Review*, *22*, 1523–1525.
- 322, 323
- Feller, W. (1940). Statistical aspects of ESP. *Journal of Parapsychology*, *4*, 271–298.
- 264
- Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). Edinburgh: Oliver and Boyd.
- 194, 265
- Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.
- 172
- Frank, L. E. (2006). *Feature network models for proximity data: Statistical inference, model selection, network representations and links with related models* (Doctoral dissertation, Leiden University). Retrieved from <http://hdl.handle.net/1887/4560>
- 134
- Frey, J. (2009). An exact multinomial test for equivalence. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *37*, 47–59.
- 235, 236

- Frühwirth–Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal*, 7, 143–167.  
16, 24, 27, 28, 29, 34, 35, 45, 79, 105, 143, 145, 147, 150, 165, 166, 346
- Frühwirth–Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.  
173, 175, 243
- Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, 8, 135–142.  
320
- Gail, M., & Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 361–372.  
208
- Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.  
14, 16, 20, 62, 124, 142, 176
- Ganz, R. E. (2014). The decimal expansion of  $\pi$  is not statistically random. *Experimental Mathematics*, 23, 99–104.  
236
- Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.  
115
- Garrison, K. E., Tang, D., & Schmeichel, B. J. (2016). Embodying power: A preregistered replication and extension of the power pose effect. *Social Psychological and Personality Science*, 7, 623–630.  
214
- Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70, 320–328.  
303
- Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74, 153–160.  
303, 304
- Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 501–514.  
24, 304
- Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 147–167). Oxford: Oxford University Press.  
303
- Gelling, N., Schofield, M. R., & Barker, R. J. (2017). rjmc: Reversible-jump MCMC using post-processing [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=rjmc> (R package version 0.3.2)  
145, 166

- Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, 2, 67–78.  
324
- Gelman, A. (2013). Two simple examples for understanding posterior  $p$ -values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7, 2595–2602.  
56
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd ed.)*. Boca Raton (FL): Chapman & Hall/CRC.  
79, 322
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.  
56, 60, 62, 82
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.  
303, 304
- Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13, 163–185.  
14, 46, 63, 141, 143
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–472.  
69, 176
- Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, 25, 165–173.  
79, 266
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 8–38.  
322
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60, 328–331.  
227
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC.  
82
- Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach* (1st ed.). Boca Raton (FL): CRC Press.  
60
- Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample  $t$  test. *The American Statistician*, 59, 252–257.  
252, 253, 254, 260, 261
- Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2019). Comparing objective and subjective Bayes factors for the two-sample comparison: The classification theorem in action. *The American Statistician*, 73, 22–31.

- 253, 255
- Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.
- 324
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.
- 16, 46, 80, 85, 90, 109, 143, 166
- Groenen, P. J., Heiser, W. J., & Meulman, J. J. (1998). City-block scaling: Smoothing strategies for avoiding local minima. In *Classification, data analysis, and data highways* (pp. 46–53). Springer.
- 116
- Gronau, Q. F., Heathcote, A., & Matzke, D. (2020). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods*, 52, 918–937.
- 150, 164
- Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician*, 74, 137–143.
- 151, 194, 203, 216, 218
- Gronau, Q. F., Raj K. N., A., & Wagenmakers, E.-J. (2019). Informed Bayesian inference for the A/B test. *Manuscript submitted for publication*. Retrieved from <http://arxiv.org/abs/1905.02068>
- 298
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–97. Retrieved from <https://doi.org/10.1016/j.jmp.2017.09.005>
- 58, 63, 64, 65, 80, 87, 89, 125, 143, 145, 146, 147
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92.
- 45, 127, 270, 280
- Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, 2, 123–138.
- 190, 193, 194, 195
- Gronau, Q. F., & Wagenmakers, E.-J. (2018). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*, 27, 277–286.
- 308
- Gronau, Q. F., & Wagenmakers, E.-J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior*, 2, 1–11.
- 319, 328
- Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2019). A simple method for comparing complex models: Bayesian model comparison

- for hierarchical multinomial processing tree models using Warp-III bridge sampling. *Psychometrika*, 84, 261–284.  
87, 107, 127, 149, 150, 164
- Grünwald, P. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.  
302
- Grünwald, P., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.  
302
- Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, 23, 74–86.  
172
- Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3, e189.  
82
- Haaf, J. M., Hoogeveen, S., Berkhout, S., Gronau, Q. F., & Wagenmakers, E.-J. (2020). A Bayesian multiverse analysis of Many Labs 4: Quantifying the evidence against mortality salience. *PsyArXiv*. Retrieved from <https://psyarxiv.com/cb9er/>  
190, 195
- Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature*, 567, 461.  
194, 265
- Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33.  
166
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., ... Zwieneberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546–573.  
189
- Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28, 55–61.  
217
- Hammersley, J. M. (1950). On estimating restricted parameters (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 12, 192–240.  
324
- Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods*. London: Methuen.  
19
- Hankin, R. K. S. (2007). Very large numbers in R: Introducing package Brobdignag. *R News*, 7.  
146

- Hastie, T., Tibshirani, R., Friedman, J., & Vetterling, W. (2008). *The elements of statistical learning (2nd ed.)*. New York: Springer.  
303
- Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2018). Dynamic models of choice. *Behavior Research Methods*, 51, 961–985.  
82, 83, 86, 93, 102, 103, 124
- Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review*, 122, 376–410.  
83
- Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, 3.  
90, 103
- Hebart, M., Zheng, C. Y., Pereira, F., & Baker, C. (2020). *Revealing the multidimensional mental representations of natural objects underlying human similarity judgments*. Retrieved from <https://doi.org/10.31234/osf.io/7wrgh>  
134
- Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, 50, 264–284.  
60, 61, 80
- Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, 23, 1440–1465.  
58
- Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, 83, 893–918.  
58
- Heck, D. W., Gronau, Q. F., & Wagenmakers, E.-J. (2019). metaBMA: Bayesian model averaging for random and fixed effects meta-analysis [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=metaBMA> (R package version 0.6.1)  
191, 218, 220
- Heck, D. W., & Wagenmakers, E.-J. (2016). Adjusted priors for Bayes factors involving reparameterized order constraints. *Journal of Mathematical Psychology*, 73, 110–116.  
79
- Helm, C. E. (1964). Multidimensional ratio scaling analysis of perceived color relations. *Journal of the Optical Society of America*, 54, 256–262.  
129
- Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society A*, 172, 137–159.  
191

- Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3, 200–215.  
190, 194, 198, 350, 351, 352
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.  
6, 15, 57, 72, 102, 142, 198, 217, 320
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623.  
155
- Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The piecewise linear ballistic accumulator model. *Cognitive Psychology*, 85, 1–29.  
104
- Hoogeveen, S., Wagenmakers, E.-J., Kay, A. C., & Elk, M. V. (2018). Compensatory control and religious beliefs: A Registered Replication Report across two countries. *Comprehensive Results in Social Psychology*, 3, 240–265.  
190, 194
- Howard, J. V. (1998). The  $2 \times 2$  table: A discussion from a Bayesian viewpoint. *Statistical Science*, 13, 351–367.  
265, 268, 273, 294
- Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, 45, 603–634.  
58
- Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59, 21–47.  
56
- Hubert, L., Arabie, P., & Hesson-McInnis, M. (1992). Multidimensional scaling in the city-block metric: A combinatorial approach. *Journal of Classification*, 9, 211–236.  
116
- Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, 27, 116–159.  
56
- Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17, 295–311.  
21
- Jackson, B., Nault, K., Smart Richman, L., LaBelle, O., & Rohleder, N. (2017). Does that pose becomes you? Testing the effect of body postures on self-concept. *Comprehensive Results in Social Psychology*, 2, 81–105.  
214, 216
- Jaditz, T. (2000). Are the digits of  $\pi$  an independent and identically distributed sequence? *The American Statistician*, 54, 12–16.  
236
- Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, 52, 297–303.

- 116
- Jamil, T., Marsman, M., Ly, A., Morey, R. D., & Wagenmakers, E.-J. (2017). What are the odds? Modern relevance and Bayes factor solutions for MacAlister's problem from the 1881 *Educational Times*. *Educational and Psychological Measurement*, 77, 819–830.
- 265
- Janzen, F. J., Tucker, J. K., & Paukstis, G. L. (2000). Experimental analysis of an early life-history stage: Selection on size of hatchling turtles. *Ecology*, 81, 2290–2304.
- 157, 158
- JASP Team. (2020). *JASP (Version 0.12.2)[Computer software]*. Retrieved from <https://jasp-stats.org/>
- 84, 191, 261, 280
- Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80, 64–72.
- 57, 102, 142
- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 31, 203–222.
- 197, 268
- Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.
- 84, 155, 193, 194, 265, 266, 269, 277, 295, 298, 314
- Jeffreys, H. (1942). On the significance tests for the introduction of new functions to represent measures. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 180, 256–268.
- 256
- Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford, UK: Oxford University Press.
- 252, 253, 260
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
- 2, 15, 27, 57, 84, 102, 142, 151, 166, 197, 216, 217, 234, 238, 240, 246, 252, 302, 320
- Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method*. London: MacMillan.
- 320, 324, 336, 337
- Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1517–1525). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/3097983.3097992>
- 265
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23, 524–532.
- 172

- Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19313–19317.  
178, 179, 183, 184
- Kadane, J. B., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 47, 3–19.  
257
- Kahneman, D. (2011). *Thinking, fast and slow*. London: Allen Lane.  
179
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.  
2, 14, 27, 57, 84, 102, 114, 142, 197, 216, 235, 252, 266, 269, 302, 304, 320
- Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B*, 54, 129–144.  
9, 85, 263, 265, 266, 268, 270, 298, 349, 397
- Kellen, D., Singmann, H., & Klauer, K. C. (2014). Modeling source-memory overdistribution. *Journal of Memory and Language*, 76, 216–236.  
56
- Keller, V. N., Johnson, D. J., & Harder, J. A. (2017). Meeting your inner super(woman): Are power poses effective when taught? *Comprehensive Results in Social Psychology*, 2, 106–122.  
214
- Klaschinski, L., Schröder-Abé, M., & Schnabel, K. (2017). Benefits of power posing: Effects on dominance and social sensitivity. *Comprehensive Results in Social Psychology*, 2, 55–67.  
214
- Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, 71, 7–31.  
72
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75, 70–98.  
56, 60, 61, 62, 83
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1, 443–490.  
189
- Knaus, J. (2015). snowfall: Easier cluster computing (based on snow) [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=snowfall> (R package version 1.84-6.1)  
146
- Kooperberg, C. (2016). logspline: Logspline density estimation routines [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=logspline> (R package version 2.1.9)  
109, 111

- Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5, 3–36.  
130, 131, 134, 136
- Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.  
114
- Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146, 451–479.  
189, 195
- Laplace, P.-S. (1829/1902). *A philosophical essay on probabilities*. London: Chapman & Hall.  
331
- Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55, 195–207.  
143
- Latu, I. M., Duffy, S., Pardal, V., & Alger, M. (2017). Power vs. persuasion: Can open body postures embody openness to persuasion? *Comprehensive Results in Social Psychology*, 2, 68–80.  
214
- Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, 45, 149–166.  
114, 116, 118, 131
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15, 1–15.  
14, 116, 118, 131
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55, 1–7.  
60, 86
- Lee, M. D. (2014). *Applications of Bayesian graphical modeling to psychophysics*. Paper presented at 30th Annual Meeting of the International Society of Psychophysics. Retrieved from <http://fechnerday.com/fd2014/pdfs/FD14.Proceedings.book.pdf>  
136
- Lee, M. D. (2018). Bayesian methods in cognitive modeling. In E.-J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience: Vol. 5. Methodology (4th ed.)* (pp. 37–84). New York: Wiley.  
92, 134, 136
- Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9, 43–58.  
131, 133

- Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, 47, 32–46.  
118
- Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, 25, 114–127.  
79, 104, 117, 166, 303, 330, 346
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.  
3, 60, 68, 72, 82, 86, 175
- Lee, M. D., & Wetzels, R. (2010). Individual differences in attention during category learning. In R. Catrambone & S. Ohlsson (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 387–392). Austin, TX: Cognitive Science Society.  
134
- Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, 72, 246–273.  
103
- Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.  
331, 334
- Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, 92, 648–655.  
15, 102
- Li, M., & Dunson, D. B. (in press). Comparing and weighting imperfect models using D-probabilities. *Journal of the American Statistical Association*. Retrieved from <https://arxiv.org/abs/1611.01241>  
322
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of  $g$  priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423.  
252, 253
- Lin, Y.-S., & Heathcote, A. (2017). ggdmc: Dynamic models of choice with parallel computation, and C++ capabilities [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=ggdmc> (R package version 0.1.6.5)  
103
- Lin, Y.-S., & Heathcote, A. (2019). Parallel probability density approximation. *Behavior Research Methods*, 51, 2777–2799.  
104
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187–192.  
193, 235, 238, 268
- Lindley, D. V. (1985). *Making decisions* (2nd ed.). London: Wiley.  
280, 301

- Lindman, H., & Caelli, T. (1978). Constant curvature Riemannian scaling. *Journal of Mathematical Psychology*, 17, 89–109.  
135
- Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician*, 43, 283–288.  
264
- Liu, S., & Trenkler, G. (2008). Hadamard, Khatri-Rao, Kronecker and other matrix products. *International Journal of Information and Systems Sciences*, 4, 160–177.  
61
- Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, 55, 331–347.  
46, 143
- Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review*, 121, 66–95.  
103
- Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14, 41–67.  
160, 161, 163, 164
- Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.  
37
- Lunn, D. J., Best, N., & Whittaker, J. C. (2009). Generic reversible jump MCMC using graphical models. *Statistics and Computing*, 19, 395–408.  
46
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.  
18
- Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., & Matzke, D. (in press). A flexible and efficient hierarchical Bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior*. Wiley Blackwell.  
58
- Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson’s correlation coefficient. *Statistica Neerlandica*, 72, 4–13.  
255
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55.  
15, 57, 84, 102, 252
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.  
15, 151, 152, 166, 216, 252, 280, 302

- Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535–1546.  
103
- Mair, P., & Leeuw, J. D. (2014). Unidimensional scaling. *Wiley StatsRef: Statistics Reference Online*, 1–3.  
116
- Malek, A., Katariya, S., Chow, Y., & Ghavamzadeh, M. (2017). Sequential multiple hypothesis testing with type I error control. In *Proceedings of the 20th international conference on artificial intelligence and statistics* (pp. 1468–1476).  
264
- Marsaglia, G. (2005). On the randomness of pi and other decimal expansions. *Interstat*, 5.  
236
- Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of "replications of important results in social psychology". *Royal Society Open Science*, 4, 160426.  
216
- Martino, D. J., Bucay, D., Butman, J. T., & Allegri, R. F. (2007). Neuropsychological frontal impairments and negative symptoms in schizophrenia. *Psychiatry Research*, 152, 121–128.  
36
- Maruyama, Y., & George, E. I. (2011). Fully Bayes factors with a generalized  $g$ -prior. *The Annals of Statistics*, 39, 2740–2765.  
253
- Matzke, D., Boehm, U., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part III: Parameter estimation in nonstandard models. *Psychonomic Bulletin & Review*, 25, 77–101.  
83
- Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, 80, 205–235.  
14, 56, 58, 60, 62, 69, 80, 83, 86, 344
- Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, 142, 1047–1073.  
86, 103
- Matzke, D., Hughes, M., Badcock, J. C., Michie, P., & Heathcote, A. (2017). Failures of cognitive control or attention? The case of stop-signal deficits in schizophrenia. *Attention, Perception, & Psychophysics*, 79, 1078–1086.  
83
- Matzke, D., Love, J., & Heathcote, A. (2017). A Bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior Research Methods*, 49, 267–281.  
80, 103

- Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, 16, 798–817.  
14, 104
- Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and Goodness-of-Fit testing in  $2^n$  contingency tables. *Journal of the American Statistical Association*, 100, 1009–1020.  
56
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45, 633–644.  
201
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1, 108–141.  
102
- Meng, X.-L. (1994). Posterior predictive  $p$ -values. *The Annals of Statistics*, 22, 1142–1160.  
56
- Meng, X.-L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11, 552–586.  
28, 47, 57, 63, 85, 87, 95, 102, 125, 141, 143, 148, 149, 165
- Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6, 831–860.  
13, 16, 28, 30, 45, 57, 63, 64, 65, 85, 87, 88, 89, 125, 141, 143, 145, 146, 165, 270
- Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85.  
166
- Mira, A., & Nicholls, G. K. (2004). Bridge estimation of the probability density at a point. *Statistica Sinica*, 14, 603–612.  
47
- Moreau, D., & Corballis, M. C. (2019). When averaging goes wrong: The case for mixture model estimation in psychological science. *Journal of Experimental Psychology: General*, 148, 1615–1627.  
208, 347
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16, 406–419.  
333
- Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.11-1*. Comprehensive R Archive Network. Retrieved from <http://cran.r-project.org/web/packages/BayesFactor/index.html>  
84, 152, 193, 218, 257, 280
- Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, 92, 121–124.

- 258
- Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52, 1–4.
- 258
- Mulder, J., & Wagenmakers, E.-J. (2016). Editor’s introduction to the special issue on “Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments”. *Journal of Mathematical Psychology*, 72, 1–5.
- 14, 302
- Mulder, M. J., Van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences—a model-based review. *Neuroscience*, 277, 872–884.
- 82
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44, 190–204.
- 302
- Myung, I. J., Forster, M. R., & Browne, M. W. (2000a). Guest editors’ introduction: Special issue on model selection. *Journal of Mathematical Psychology*, 44, 1–2.
- 14
- Myung, I. J., Forster, M. R., & Browne, M. W. (2000b). Model selection [Special issue]. *Journal of Mathematical Psychology*, 44(1–2).
- 302
- Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50, 167–179.
- 302
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam’s razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4, 79–95.
- 57, 102, 142, 302
- Nathoo, F. S., & Masson, M. E. J. (2016). Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology*, 72, 144–157.
- 302
- Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, 2, 28–34.
- 320, 330, 331, 334
- Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50, 101–122.
- 15
- Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 59–66). Cambridge, MA: MIT Press.
- 135

- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11, 125–139.  
21
- Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 3–48.  
20, 24
- Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, 14, 1105–1107.  
227
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348, 1422–1425.  
172
- Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, 38, 415–432.  
122
- Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43, 25–53.  
114
- Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, 50, 530–556.  
134
- Nott, D. J., Kohn, R. J., & Fielding, M. (2008). Approximating the marginal likelihood using copula. *arXiv preprint arXiv:0810.5474*.  
346
- Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.  
14, 16, 20
- Nuijten, M. B., Hartgerink, C. H., Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, 48, 1205–1226.  
251
- Oh, M.-S. (2012). A simple and efficient Bayesian procedure for selecting dimensionality in multidimensional scaling. *Journal of Multivariate Analysis*, 107, 200–209.  
114, 116, 118
- Oh, M.-S., & Raftery, A. E. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, 96, 1031–1044.  
114, 116, 118
- O’Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73, 69–81.  
264, 265

- O'Hagan, A., & Forster, J. J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.)*. London: Arnold.  
237, 314
- Okada, K. (2012). A Bayesian approach to asymmetric multidimensional scaling. *Behaviormetrika*, 39, 49–62.  
118
- Okada, K., & Mayekawa, S.-i. (2018). Post-processing of Markov chain Monte Carlo output in Bayesian latent variable models with application to multidimensional scaling. *Computational Statistics*, 33, 1457–1473.  
118, 119
- Okada, K., & Shigemasu, K. (2010). Bayesian multidimensional scaling for the estimation of a Minkowski exponent. *Behavior Research Methods*, 42, 899–905.  
116
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.  
172
- Osth, A., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with a combined model of retrieval and decision making. *Cognitive Psychology*, 104, 106–142.  
83
- Overstall, A. M. (2010). *Default Bayesian model determination for generalised linear mixed models* (Doctoral dissertation, University of Southampton). Retrieved from <https://eprints.soton.ac.uk/170229/>  
63, 67, 107, 149
- Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54, 3269–3288.  
25, 27, 28, 30, 52, 63, 65, 70, 78, 87, 95, 103, 106, 145, 148, 157
- Owen, A., & Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95, 135–143.  
21, 147
- Pajor, A. (2017). Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis*, 12, 261–287.  
20
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528–530.  
172
- Pham-Gia, T., Van Thin, N., & Doan, P. P. (2017). Inferences on the difference of two proportions: A Bayesian approach. *Open Journal of Statistics*, 7, 1–15.  
265
- Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, 113, 57–83.  
116

- Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109, 472–491.  
14
- Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, 53, 633–654.  
56
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria.  
18, 41, 62, 144, 176, 345
- Plummer, M. (2016). rjags: Bayesian graphical models using MCMC [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=rjags> (R package version 4-6)  
143
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6, 7–11.  
35, 64, 89, 143
- Poirier, D. J. (2006). The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1, 969–980.  
13
- Polya, G. (1941). Heuristic reasoning and the theory of probability. *The American Mathematical Monthly*, 48, 450–465.  
234
- Polya, G. (1954a). *Mathematics and plausible reasoning: Vol. I. Induction and analogy in mathematics*. Princeton, NJ: Princeton University Press.  
331, 332
- Polya, G. (1954b). *Mathematics and plausible reasoning: Vol. II. Patterns of plausible inference*. Princeton, NJ: Princeton University Press.  
331
- Portman, F. (2019). bayesAB: Fast Bayesian methods for AB testing [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=bayesAB> (R package version 1.1.2)  
265
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>  
16, 70, 84, 103, 143, 191, 254, 258, 264, 265
- Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1226–1243.  
83
- Raftery, A. E., & Banfield, J. D. (1991). Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics (Bayesian image restora-

- tion, with two applications in spatial statistics)–(discussion). *Annals of the Institute of Statistical Mathematics*, 43, 32–43.
- 19
- Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Kegan Paul.
- 234
- Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26, 653–656.
- 214
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- 82, 103
- Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, 2, 237–279.
- 82
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20, 873–922.
- 82, 103
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20, 260–281.
- 82
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- 37
- Rhodes, K. M., Turner, R. M., & Higgins, J. P. T. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology*, 68, 52–60.
- 215, 226
- Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95, 318–339.
- 55, 58
- Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, 14, 184–201.
- 68, 69, 70, 71, 72, 78, 333
- Rissanen, J. (2007). *Information and complexity in statistical modeling*. New York: Springer.
- 302
- Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72, 33–37.

- 15  
Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer.
- 176  
Robert, C. P., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. New York: Springer-Verlag.
- 272  
Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability revisited. *Statistical Science*, *24*, 141–172.
- 252  
Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of  $p$  values in composite null models. *Journal of the American Statistical Association*, *95*, 1143–1156.
- 56  
Ronay, R., Tybur, J. M., van Huijstee, D., & Morssinkhof, M. (2017). Embodied power, testosterone, and overconfidence as a causal pathway to risk taking. *Comprehensive Results in Social Psychology*, *2*, 28–43.
- 214  
Rosenthal, R. (1979). An introduction to the file drawer problem. *Psychological Bulletin*, *86*, 638–641.
- 172  
Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.
- 205, 265, 348  
Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological Methods*, *24*, 606–621.
- 191, 208  
Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.
- 14, 15, 60  
Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.
- 15, 69  
Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223.
- 14, 15  
Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.
- 14  
Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689.
- 191

- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877–903.  
84
- Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73, 186–190.  
197
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56, 356–374.  
84, 302
- Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, 8, 520–547.  
261
- Rouder, J. N., Morey, R. D., Verhagen, A. J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, 22, 304–321.  
57, 72
- Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, 2, 1–12.  
261
- Rouder, J. N., Province, J. M., Morey, R. D., Gómez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80, 491–513.  
103
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian  $t$  tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237.  
151, 152, 216, 252, 253, 254, 260, 280
- Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science*, 28, 1698–1701.  
190, 194, 195, 217, 227
- Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, 22, 391–407.  
15
- Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling*. New York, NY: Academic Press.  
114
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.  
46, 56, 302
- Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38, 2587–2619.

- 80, 261
- Severini, T. A., Mukerjee, R., & Ghosh, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika*, 89, 952–957.
- 253
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 286–292.
- 77, 84, 303, 314, 319
- Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22, 325–345.
- 114, 115
- Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27, 125–140.
- 114, 130
- Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39, 373–422.
- 114, 121, 134, 135
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 214, 390–398.
- 135
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317–1323.
- 114, 115, 135
- Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. R. Pomerantz & G. L. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53–71). Washington, DC: American Psychological Association.
- 115, 116, 135
- Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86, 87–123.
- 135
- Shiffrin, R. M., & Chandramouli, S. H. (2019). Commentary on Gronau and Wagenmakers. *Computational Brain & Behavior*, 2, 12–21.
- 320, 332, 334
- Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248–1284.
- 3, 4, 5, 14, 15, 82, 86
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.
- 172
- Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, 9, 552–555.

- 190
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.  
171, 172, 178, 179, 181, 182, 183
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681.  
172
- Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575.  
56
- Singmann, H., Kellen, D., & Klauer, K. C. (2013). Investigating the other-race effect of Germans towards Turks and Arabs using multinomial processing tree models. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1330–1335). Austin, TX: Cognitive Science Society.  
56
- Sinharay, S., & Stern, H. S. (2005). An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *14*, 415–435.  
34, 67, 157, 158, 159
- Skorski, M. (2019). Bounds on Bayes factors for binomial A/B testing. *arXiv preprint arXiv:1903.00049*. Retrieved from <https://arxiv.org/abs/1903.00049>  
265
- Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183.  
56, 72, 73, 80, 83
- Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, *14*, 2685–2699.  
191, 192
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*, 583–639.  
45, 56, 84, 103
- Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society A*, *157*, 357–416.  
237, 247
- Stan Development Team. (2016). *RStan: The R interface to Stan*. Retrieved from <http://mc-stan.org/> (R package version 2.14.1)  
18, 62, 143, 145, 155, 280
- Stan Development Team. (2017). Stan modeling language users guide and reference manual, version 2.16.0 [Computer software manual]. Retrieved from <http://mc-stan.org>

- 155, 157, 164
- Stangl, D., & Berry, D. A. (2000). *Meta-analysis in medicine and health policy*. New York: Marcel Dekker.
- 192
- Steingroever, H., Pachur, T., Šmíra, M., & Lee, M. D. (2018). Bayesian techniques for analyzing group differences in the Iowa gambling task: A case study of intuitive and deliberate decision makers. *Psychonomic Bulletin & Review*, 25, 951–970.
- 35, 36
- Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., & Wagenmakers, E.-J. (2013). Performance of healthy participants on the Iowa gambling task. *Psychological Assessment*, 25, 180–193.
- 35
- Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013a). A comparison of reinforcement-learning models for the Iowa gambling task using parameter space partitioning. *The Journal of Problem Solving*, 5, Article 2.
- 35, 36
- Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013b). Validating the PVL-Delta model for the Iowa gambling task. *Frontiers in Psychology*, 4:898.
- 35
- Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, 1, 161–183.
- 35, 36, 38, 75
- Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2016). Bayes factors for reinforcement-learning models of the Iowa Gambling Task. *Decision*, 3, 115–131.
- 35, 36, 38, 39, 40, 41
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54, 30–34.
- 172
- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, 49, 108–112.
- 172
- Steyvers, M. (2006). Multidimensional scaling. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. Wiley Online Library.
- 114
- Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, 25, 1371–1470.
- 43
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, 36, 111–147.
- 302

- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B*, 39, 44–47.  
77, 314
- Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, 768–777.  
257, 350
- Strickland, L., Loft, S., Remington, R., & Heathcote, A. (2018). Racing to remember: A theory of decision control in event-based prospective memory. *Psychological Review*, 125, 851–887.  
83, 84
- Stucchio, C. (2015). *Bayesian A/B testing at VWO* (Tech. Rep.). VWO. Retrieved from [https://www.chrisstucchio.com/pubs/VWO\\_SmartStats\\_technical\\_whitepaper.pdf](https://www.chrisstucchio.com/pubs/VWO_SmartStats_technical_whitepaper.pdf)  
265
- Su, Y.-S., & Yajima, M. (2015). R2jags: Using R to run JAGS [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=R2jags> (R package version 0.5-7)  
143, 153
- Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10, 277–303.  
191, 217
- Tamhane, A. C., & Shi, J. (2009). Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of FDR. *Lecture Notes-Monograph Series*, 57, 304–325.  
173
- Team, S. D. (2016). *rstanarm: Bayesian applied regression modeling via Stan*. Retrieved from <http://mc-stan.org/> (R package version 2.13.1)  
144
- Tilman, G., Osth, A., van Ravenzwaaij, D., & Heathcote, A. (2017). A diffusion decision model analysis of evidence variability in the lexical decision task. *Psychonomic Bulletin & Review*, 24, 1949–1956.  
83
- Tilman, G., Strayer, D., Eidels, A., & Heathcote, A. (2017). Modeling cognitive load effects of conversation between a passenger and driver. *Attention, Perception, & Psychophysics*, 79, 1795–1803.  
83, 103
- Tran, N.-H. (2018). *Empirical priors for sequential sampling models* (Unpublished master's thesis). University of Amsterdam, The Netherlands.  
82
- Treat, T. A., McFall, R. M., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, 13, 549–565.  
131, 132
- Tu, S.-J., & Fischbach, E. (2005). A study on the randomness of the digits of  $\pi$ . *International Journal of Modern Physics C*, 16, 281–294.

- 236
- Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, 21, 227–250.
- 104
- Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, 18, 368–384.
- 83, 124
- Turner, B. M., Wang, T., & Merkle, E. C. (2017). Factor analysis linking functions for simultaneously modeling neural and behavioral data. *NeuroImage*, 153, 28–48.
- 80
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- 135
- Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, 89, 123–154.
- 115, 135
- van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20, 293–309.
- 172
- Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, 60, 58–71.
- 80
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.
- 21, 57, 80, 142, 302
- Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, 25, 1–4.
- 190
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16, 44–62.
- 82, 87, 102
- van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., ... Wagenmakers, E.-J. (in press). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. Retrieved from <https://psyarxiv.com/yqxfr>
- 206
- van Erp, S., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, 5.
- 193, 219, 220, 228, 353

- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54, 491–498.  
79, 330
- Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, 72, 183–190.  
13
- Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. Retrieved from <https://CRAN.R-project.org/package=loo> (R package version 2.0.0)  
303, 328
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413–1432.  
56, 84, 303, 304
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved  $\hat{R}$  for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*. Retrieved from <https://arxiv.org/abs/1903.08008>  
93
- Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6, 142–228.  
303
- Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2019). Limitations of “Limitations of Bayesian leave-one-out cross-validation for model selection”. *Computational Brain & Behavior*, 2, 22–27.  
320, 321, 334
- Venn, J. (1888). *The logic of chance* (3rd ed.). New York: MacMillan.  
236
- Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, 90, 614–618.  
109
- Verhagen, J., Levy, R., Millsap, R. E., & Fox, J.-P. (2015). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*, 72, 171–182.  
13
- Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., ... Yıldız, E. (2018). Registered Replication Report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, 1, 299–317.  
201, 202, 203, 204, 207, 210, 212
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., ..., & Wagenmakers, E.-J. (under review). *A multi-site, preregistered, paradigmatic test of the ego depletion effect*. (University of Minnesota, Minneapolis MN.)  
190, 195

- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of  $p$  values. *Psychonomic Bulletin & Review*, 14, 779–804.  
83
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R., ... Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11, 917–928.  
183, 189, 218, 257, 348, 350, 351, 352
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11, 192–196.  
56, 304, 314
- Wagenmakers, E.-J., Gronau, Q. F., & Vandekerckhove, J. (2018). Five Bayesian intuitions for the stopping rule principle. *Manuscript submitted for publication*. Retrieved from <https://psyarxiv.com/5ntkd>  
205
- Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50, 149–166.  
329
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60, 158–189.  
35, 43, 71, 85, 109, 166
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian statistical inference for psychological science. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25, 35–57.  
92, 102, 207, 265, 350
- Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25, 169–176.  
207, 350
- Wagenmakers, E.-J., & Waldorp, L. (2006a). Editors' introduction. *Journal of Mathematical Psychology*, 50, 99–100.  
14
- Wagenmakers, E.-J., & Waldorp, L. (2006b). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, 50(2).  
302
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 627–633.  
172
- Wang, L., & Meng, X.-L. (2016). Warp bridge sampling: The next generation. *arXiv preprint arXiv:1609.07690*.  
28, 79, 166, 346
- Wang, M., & Liu, G. (2016). A simple two-sample Bayesian  $t$ -test for hypothesis testing. *The American Statistician*, 70, 195–201.

- 253, 254, 255, 260, 261
- Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statistical Science*, 4, 298–340.
- 264
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92–107.
- 322
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.
- 56
- West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models* (2nd ed.). New York: Springer-Verlag.
- 160, 161
- Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio test. *Computational Statistics & Data Analysis*, 54, 2094–2102.
- 35, 250
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, 6, 291–298.
- 178, 251
- Wetzels, R., Tutschkow, D., Dolan, C., van der Sluis, S., Dutilh, G., & Wagenmakers, E.-J. (2016). A Bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, 72, 200–209.
- 13
- Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, 54, 14–27.
- 15, 41
- Whewell, W. (1840). *The philosophy of the inductive sciences, founded upon their history* (Vol. II). London: John W. Parker.
- 331
- Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift–diffusion model in Python. *Frontiers in Neuroinformatics*, 7:14.
- 82, 87, 102
- Wood, S. (2016). Just another Gibbs additive modeler: Interfacing JAGS and mgcv. *Journal of Statistical Software*, 75(7), 1–15. Retrieved from <https://www.jstatsoft.org/v075/i07> doi: 10.18637/jss.v075.i07
- 166
- Worthy, D. A., & Maddox, W. T. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W. K. Estes. *Journal of Mathematical Psychology*, 59, 41–49.
- 37

- Worthy, D. A., Pang, B., & Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, 4.  
37
- Wrench Jr, J. W. (1960). The evolution of extended decimal approximations to  $\pi$ . *The Mathematics Teacher*, 53, 644–650.  
235
- Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, 38, 715–731.  
331
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42, 369–390.  
197, 235, 252, 302, 305
- Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13, 917–1007.  
304, 320, 321, 322, 326, 328
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12, 1100–1122.  
303
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233–243). Amsterdam: North-Holland.  
253

---

# Nederlandse Samenvatting

---

In dit proefschrift, getiteld “*Bayes Factor Model Comparison for Psychological Science*”, werden rivaliserende wetenschappelijke modellen vergeleken door ze te zien als voorspellers, en de kwaliteit van hun voorspellingen te beoordelen aan de hand van de *Bayes factor*. Het eerste deel van het proefschrift behandelde *bridge sampling*, een computationele methode voor het schatten van de marginale waarschijnlijkheid – de hoofdcomponent voor het berekenen van *Bayes factors*. Het tweede deel van het proefschrift behandelde Bayesiaanse methoden voor meta-analyse van meerdere studies. Een centraal concept in dit deel was het combineren van verschillende voorspellers middels *Bayesian model averaging (BMA)*. Het derde deel van dit proefschrift introduceerde Bayesiaanse equivalenten van een aantal standaard statistische toetsen. Een centraal concept van dit deel was het meenemen van voorkennis in de analyses, zodat de voorspellingen van de statistische modellen accurater worden. Hieronder geef ik een korte samenvatting van elk hoofdstuk.

## Deel I: Bridge Sampling

Hoofdstuk 2 gaf een tutorial over *bridge sampling*, waarmee de marginale waarschijnlijkheid geschat kan worden, de hoofdcomponent voor het vergelijken van de kwaliteit van de voorspellingen van verschillende statistische modellen door middel van de *Bayes factor*. De methode werd geïntroduceerd door de methode te vergelijken met drie andere *Monte Carlo sampling* methoden die gebruikt worden voor het schatten van de marginale waarschijnlijkheid in een simpel beta-binomiaal voorbeeld. De praktische haalbaarheid van de methode werd aangetoond aan de hand van *single-participant* en hiërarchische versies van het *reinforcement learning* model. Hieruit werd geconcludeerd dat *bridge sampling* een aantrekkelijke methode is voor het vergelijken van modellen in mathematische psychologie. In dit veld zijn onderzoekers vaak geïnteresseerd in het vergelijken van een kleine set van hiërarchische modellen, waar er veel parameters geschat moeten worden.

Hoofdstuk 3 liet een toepassing zien van een geavanceerde versie van *bridge sampling (Warp-III)* voor het vergelijken van *hierarchical multinomial processing tree (MPT)* modellen. Deze versie van *bridge sampling* houdt rekening met moge-

lijke *skewing* van de posterior verdeling en kan daardoor een preciezere schatting geven van de marginale waarschijnlijkheid van de modellen. Het eerste voorbeeld liet zien hoe deze methode gebruikt kan worden voor het beoordelen welke model parameters verschillen tussen trials. Dit wordt bewerkstelligd door middel van het combineren van de voorspellingen van verschillende voorspellers (oftewel *Bayesian model averaging*). Het tweede voorbeeld was het opnieuw analyseren van data die werd gebruikt om twee niet-geneste *MPT* modellen te vergelijken, om het *illusory truth effect* te bestuderen.

Hoofdstuk 4 liet een toepassing zien van *Warp-III bridge sampling* voor het berekenen van de marginale waarschijnlijkheid van *evidence-accumulation* modellen. Het *Linear Ballistic Accumulator (LBA)* model werd gebruikt om te demonstrenen dat het combineren van *differential evolution Markov chain Monte Carlo (DE-MCMC)* en *Warp-III bridge sampling* een precieze schatting geeft van de marginale waarschijnlijkheid van zowel *single-participant*, als hiërarchische versies van het *LBA* model. Door de methode te koppelen met een gebruiksvriendelijke software implementatie kunnen onderzoekers gemakkelijk de marginale waarschijnlijkheid schatten van verscheidene *evidence-accumulation* modellen. Dit hoofdstuk concludeerde met een serie van aanbevelingen voor het toepassen van *Warp-III bridge sampling*.

Hoofdstuk 5 liet een toepassing zien van *multidimensional scaling (MDS)* modellen voor het achterhalen van het optimale aantal dimensies, en de metrische structuur van de ruimte waarin gemeten wordt. In de toepassing werd voorkennis meegenomen om het model identificeerbaar te maken wanneer er zowel psychologisch gescheiden, als psychologisch geïntegreerde, stimuli gebruikt worden. *DE-MCMC* werd gebruikt samen met *Warp-III bridge sampling* om conclusies te trekken over de model parameters, om het optimale aantal dimensies te achterhalen, en om de gepaste metriek te vinden voor de latente ruimte. Aan de hand van vijf bestaande datasets werd aangetoond dat de methode zinnige resultaten geeft in deze settings. Het hoofdstuk besprak ook een aantal onopgeloste uitdagingen die opgelost dienen te worden, alvorens de methode in het algemeen toegepast kan worden.

Hoofdstuk 6 introduceerde een R pakket genaamd **bridgesampling**. Dit pakket kan gebruikt worden om de marginale waarschijnlijkheid (of, algemener, de normaliserende constante) te schatten door middel van *bridge sampling*, op een algemene en gebruiksvriendelijke manier. Gecombineerd met de Bayesiaanse sampling software **Stan** (Carpenter et al., 2017), kan het R pakket automatisch de marginale waarschijnlijkheid schatten. Het gebruik van het pakket werd gedemonstreerd aan de hand van drie data voorbeelden.

## Deel II: Multi-Model Meta-Analyse

Hoofdstuk 7 stelde voor om meta-analyse van de verdeling van significante  $p$ -waarden uit te voeren aan de hand van een Bayesiaans gemengd model. Het gemengde model schat de proportie van de significante resultaten die voortkomen uit de nulhypothese, die geen effect veronderstelt, en het model geeft voor elke  $p$ -waarde een schatting van de kans dat deze voortkomt uit de nulhypothese. De

procedure werd gedemonstreerd aan de hand van twee voorbeelden. Door middel van een webapplicatie kunnen onderzoekers de methode toepassen om  $p$ -waarden te analyseren op een gebruiksvriendelijke manier.

Hoofdstuk 8 zette een Bayesiaanse *model-averaged* meta-analyse uiteen. De procedure is gebaseerd op het combineren van verschillende voorspellers, zodat voorkomen kan worden dat een alles-of-niets beslissing genomen wordt tussen een *fixed-effect* model en een *random-effects* meta-analyse model. Deze procedure combineert vier Bayesiaanse meta-analyse modellen op basis van hun plausibiliteit onder de geobserveerde data: (1) *fixed-effect* nulhypothese, (2) *fixed-effect* alternatieve hypothese, (3) *random-effects* nulhypothese, en (4) *random-effects* alternatieve hypothese. Deze procedure stelt de onderzoeker in staat om twee belangrijke vragen te beantwoorden: “Is het algemene effect verschillend van nul?” en “Zijn er verschillen in *effect size* tussen de studies?”. De methode werd gedemonstreerd aan de hand van een data voorbeeld.

Hoofdstuk 9 liet een toepassing zien van de Bayesiaanse *model-averaged* meta-analyse geïntroduceerd in Hoofdstuk 8, op de resultaten van zes gepreregistreerde studies over het effect van *power posing*. De analyse ging hoofdzakelijk over het effect van *power posing* op *felt power*. De meta-analyse resulteerde in sterk bewijs voor het effect van *power posing* op *felt power*. Echter was het bewijs enkel gematigd wanneer participanten werden uitgesloten die al bekend waren met het effect.

## Deel III: Hypothese Toetsing

Hoofdstuk 10 demonstreerde hoe Bayesiaanse statistiek gebruikt kan worden om het bewijs te kwantificeren voor een algemene wet, gebaseerd op een eindige dataset. Het hoofdstuk beschreef hoe het kwantificeren van bewijs voor de hypothese dat fundamentele constanten (zoals  $\pi$ ,  $e$ ,  $\sqrt{2}$ , en  $\ln 2$ ) normaal zijn. Bayesiaanse statistiek werd gebruikt om de restrictieve hypothese te toetsen dat elk cijfer even vaak voorkomt in de decimalen van de constanten. Voor alle vier de constanten werd bewijs gevonden voor de algemene wet.

Hoofdstuk 11 stelde voor om een flexibele  $t$ -prior te gebruiken voor de effect size in de Bayesiaanse  $t$ -toets. Deze prior stelt wetenschappers in staat om voorkennis mee te nemen in hun analyse, en zo een preciezere voorspelling te maken. Deze prior verdeling bevat vorige subjectieve, maar ook objectieve versies van de Bayesiaanse  $t$ -toets. Twee graadmeters werden voorgesteld om te kwantificeren tot in hoeverre werd voldaan aan de desiderata van de objectieve *Bayes factor*: voorspellende passendheid en informatie consistentie. De methode werd gedemonstreerd aan de hand van een voorbeeld over de *facial feedback* hypothese, waarbij de priors werden gekozen door een expert.

Hoofdstuk 12 introduceerde **abtest**, een R pakket voor het uitvoeren van Bayesiaanse A/B toetsen. De geïmplementeerde aanpak is gebaseerd op het werk van Kass en Vaidyanathan (1992) en stelt onderzoekers in staat om het de opstapeling van bewijs te volgen voor de hypotheses dat de behandeling een positief effect, negatief effect, of geen effect heeft. Met deze methode is het ook mogelijk om

kennis mee te nemen van experts voor het berekenen van de relatieve a priori waarschijnlijkheid van zowel hypothesen, als de verwachte waarde van het effect.

Hoofdstuk 13 besprak Bayesiaanse *leave-one-out cross-validation (LOO)*, een alternatieve methode voor het vergelijken van rivaliserende modellen. Verschillende tekortkomingen van deze benadering werden aangetoond aan de hand van concrete data voorbeelden. Hieruit werd de conclusie getrokken dat *LOO* geen wondermiddel is voor modelselectie.

Hoofdstuk 14 biedt een weerwoord aan drie commentaren op Hoofdstuk 13. Elk van de commentaren werd behandeld, en aanvullende tekortkomingen van methoden gebaseerd op *LOO* (zoals *Bayesian stacking*) werden aangetoond. Deze methoden werden vergeleken met methoden die wel consistent zijn in hun gebruik van de regel van Bayes voor zowel parameter schatting als model vergelijking. De conclusie was dat methoden die gebaseerd zijn op *LOO*, niet op één lijn liggen met het epistemische doel van mathematische psychologie.

---

## Acknowledgements

---

I would like to thank my supervisor E.J., my co-supervisors Dora and Maarten, my paranympths Alexandra and Johnny, my co-authors and colleagues, my friends, and my family. Danke.



---

## Publications

---

1. Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., Ly, A., Verhagen, A. J., Selker, R., Sasiadek, A., **Gronau, Q. F.**, Love, J., & Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology: Cognition*, 6:494.
2. van Elk, M., Matzke, D., **Gronau, Q. F.**, Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology: Personality and Social Psychology*, 6:1365.
3. Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C. **Gronau, Q. F.**, Hafenbrack, A. C., Liao, E. Y., Ly, A., Marsman, M., Murase, T., Qureshi, I., Schaerer, M., Thornley, N., Tworek, C. M., Wagenmakers, E.-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Kennedy, D. M., Legate, N., Luoma, T. P., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A.-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei, C., Wetter, E., Wilson, C., Darroux, S.-C., & Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66, 55–67.
4. Tierney, W., Schweinsberg, M., Jordan, J., Kennedy, D., Qureshi, I., Sommer, S. A., Thornley, N., Madan, N., Vianello, M., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C., du Plessis, C., **Gronau, Q. F.**, Hafenbrack,

- A., Liao, E., Ly, A., Marsman, M., Murase, T., Schaerer, M., Tworek, C., Wagenmakers, E.-J., Wong, L., Anderson, T., Bauman, C., Bedwell, W., Brescoll, V., Canavan, A., Chandler, J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P., Amell, A., Frick, S., Gamez-Djokic, M., Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B., Hartwich, N., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Legate, N., Luoma, T., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J., Sackett, A., Sellier, A.-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., van Bavel, J., Washburn, A., Wei, C., Wetter, E., Wilson, C., Darroux, S.-C., & Uhlmann, E. L. (2016). Data from a pre-publication independent replication initiative examining ten moral judgement effects. *Scientific Data*, 3, 160082. <https://www.nature.com/articles/sdata201682>
5. Wagenmakers, E.-J., Beek, T., Dijkhoff, L., **Gronau, Q. F.**, Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupianez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Oezdogru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11, 917–928.
  6. **Gronau, Q. F.**, van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, 2, 123–138. <https://psyarxiv.com/9z8ch/>
  7. Scheibehenne, B., **Gronau, Q. F.**, Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science*, 28, 1698–1701. <https://osf.io/preprints/bitss/8ucf6/>
  8. **Gronau, Q. F.**, Duizer, M., Bakker, M., & Wagenmakers, E.-J. (2017). Bayesian mixture modeling of significant  $p$  values: A meta-analytic method to estimate the degree of contamination from  $\mathcal{H}_0$ . *Journal of Experimental Psychology: General*, 146, 1223–1233. <https://osf.io/mysbp/>
  9. **Gronau, Q. F.**, Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–97. <https://arxiv.org/abs/1703.05984>

10. Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., **Gronau, Q. F.**, Smíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25, 35–57. <https://osf.io/m6bi8/>
11. Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Selker, R., **Gronau, Q. F.**, Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Smíra, M., Epskamp, S., Etz, A., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76. <https://osf.io/m6bi8/>
12. Etz, A., **Gronau, Q. F.**, Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, 25, 219–234. <https://nicebrain.files.wordpress.com/2016/02/etz-et-al-preprint-how-to-become-a-bayesian.pdf>
13. Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., **Gronau, Q. F.**, van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, 1, 357–366. <https://psyarxiv.com/zqkyt>
14. Ly, A., Raj, A., Etz, A., Marsman, M., **Gronau, Q. F.**, & Wagenmakers, E.-J. (2018). Bayesian reanalyses from summary statistics: A guide for academic consumers. *Advances in Methods and Practices in Psychological Science*, 1, 367–374. <https://osf.io/7dzmk/>
15. **Gronau, Q. F.**, & Wagenmakers, E.-J. (2018). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*, 27, 277–286. <http://www.tandfonline.com/doi/abs/10.1080/10586458.2016.1256006>
16. **Gronau, Q. F.**, Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2019). A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using Warp-III bridge sampling. *Psychometrika*, 84, 261–284. <https://psyarxiv.com/yxhfm>
17. **Gronau, Q. F.**, & Wagenmakers, E.-J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior*, 2, 1–11. <https://psyarxiv.com/at7cx/>
18. **Gronau, Q. F.**, & Wagenmakers, E.-J. (2019). Rejoinder: More limitations of Bayesian leave-one-out cross-validation. *Computational Brain & Behavior*, 2, 35–47. <https://psyarxiv.com/38zxx>
19. Boffo, M., Zerhouni, O., **Gronau, Q. F.**, van Beek, R. J. J., Nikolaou, K., Marsman, M., & Wiers, R. W. (2019). Cognitive bias modification for

behavior change in alcohol and smoking addiction: A Bayesian meta-analysis of individual participant data. *Neuropsychology Review*, 29, 52–78.

20. van Dongen, N. N. N., van Doorn, J. B., **Gronau, Q. F.**, van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, 73, 328–339. <https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1565553>
21. Heck, D. W., Overstall, A. M., **Gronau, Q. F.**, & Wagenmakers, E.-J. (2019). Quantifying uncertainty in transdimensional Markov chain Monte Carlo using discrete Markov models. *Statistics and Computing*, 29, 631–643. <https://arxiv.org/abs/1703.10364>
22. Stefan, A. M., **Gronau, Q. F.**, Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis with informed priors. *Behavior Research Methods*, 51, 1042–1058. <https://psyarxiv.com/aqr79>
23. Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A., **Gronau, Q. F.**, Smíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2019). JASP – graphical statistical software for common statistical designs. *Journal of Statistical Software*, 88. <https://www.jstatsoft.org/index.php/jss/article/view/v088i02/v088i02.pdf>
24. **Gronau, Q. F.**, Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92. <https://www.jstatsoft.org/article/view/v092i10>
25. van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., **Gronau, Q. F.**, Kucharsky, S., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology*, 120, 73–96. <https://www.cairn.info/revue-l-annee-psychologique-2020-1-page-73.htm>
26. **Gronau, Q. F.**, Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician*, 74, 137–143. <https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1562983>
27. **Gronau, Q. F.**, Heathcote, A., & Matzke, D. (2020). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods*, 52, 918–937. <https://link.springer.com/article/10.3758/s13428-019-01290-6>
28. Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharsky, S., Derks, K., **Gronau, Q. F.**, Raj, A., Boehm,

- U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.-J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the  $p$  value hypothesis test. *Computational Brain & Behavior*, 3, 153–161. <https://link.springer.com/content/pdf/10.1007/s42113-019-00070-x.pdf>
29. Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., **Gronau, Q. F.**, Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., Cheung, F., Cimpian, A., Dohle, S., Donnellan, M. B., Hahn, A., Hall, M. P., Jiménez-Leal, W., Johnson, D. J., Lucas, R. E., Monin, B., Montealegre, A., Mullen, E., Pang, J., Ray, J., Reinero, D. A., Reynolds, J., Sowden, W., Storage, D., Su, R., Tworek, C. M., Van Bavel, J. J., Walco, D., Wills, J., Xu, X., Yam, K. C., Yang, X., Cunningham, W. A., Schweinsberg, M., Urwitz, M., The Crowdsourcing Hypothesis Tests Collaboration, & Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146, 451–479. <https://osf.io/fgepx/>
  30. **Gronau, Q. F.**, & Lee, M. D. (2020). Bayesian inference for multidimensional scaling representations with psychologically interpretable metrics. *Computational Brain & Behavior*, 3, 322–340. <https://psyarxiv.com/5zmep/>
  31. Hinne, M., **Gronau, Q. F.**, van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3, 200–215. <https://journals.sagepub.com/doi/full/10.1177/2515245919898657>
  32. Wagenmakers, E.-J., **Gronau, Q. F.**, Dablander, F., & Etz, A. (in press). The support interval. *Erkenntnis*. <https://psyarxiv.com/zwnxb/>
  33. van Doorn, J., van den Bergh, D., Boehm, U., Dablander, F., Derks, K., Draws, T., Evans, N. J., **Gronau, Q. F.**, Hinne, M., Kucharsky, S., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (in press). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. <https://psyarxiv.com/yqxfr>
  34. Wagenmakers, E.-J., **Gronau, Q. F.**, & Vandekerckhove, J. (2019). Five Bayesian intuitions for the stopping rule principle. Manuscript submitted for publication. <https://psyarxiv.com/5ntkd>
  35. **Gronau, Q. F.**, Raj K. N., A., & Wagenmakers, E.-J. (2019). Informed Bayesian inference for the A/B test. Manuscript submitted for publication. <https://arxiv.org/abs/1905.02068>
  36. **Gronau, Q. F.**, Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2020). A primer on Bayesian model-averaged meta-analysis. Manuscript submitted for publication. <https://psyarxiv.com/97qup>

37. Hulme, O. J., Wagenmakers, E.-J., Damkier, P., Madelung, C. F., Siebner, H. R., Helweg-Larsen, J., **Gronau, Q. F.**, Benfield, T., & Madsen, K. H. (2020). Reply to Gautret et al. 2020: A Bayesian reanalysis of the effects of hydroxychloroquine and azithromycin on viral carriage in patients with COVID-19. Manuscript submitted for publication. <https://osf.io/7ax9w/>
38. van den Bergh, D., Clyde, M. A., Raj K. N., A., de Jong, T., **Gronau, Q. F.**, Marsman, M., Ly, A., & Wagenmakers, E.-J. (2020). A tutorial on Bayesian multi-model linear regression with BAS and JASP. Manuscript submitted for publication. <https://psyarxiv.com/pqju6/>
39. Sarafoglou, A., Haaf, J. M., Ly, A., **Gronau, Q. F.**, Wagenmakers, E.-J., & Marsman, M. (2020). Evaluating multinomial order restrictions with bridge sampling. Manuscript submitted for publication. <https://psyarxiv.com/bux7p/>