Evaluating Multinomial Order Restrictions with Bridge Sampling

Alexandra Sarafoglou[1], Julia M. Haaf[1], Alexander Ly[1,2], Quentin F. Gronau[1], Eric-Jan Wagenmakers[1], & Maarten Marsman[1]

[1]Department of Psychology, Psychological Methods, University of Amsterdam, The Netherlands

[2]Centrum Wiskunde & Informatica, Machine Learning Group, The Netherlands

Correspondence concerning this article should be addressed to: Alexandra Sarafoglou, Department of Psychology, PO Box 15906, 1001 NK Amsterdam, The Netherlands, E-mail: alexandra.sarafoglou@gmail.com.

Abstract

Hypotheses concerning the distribution of multinomial proportions typically entail exact equality constraints that can be evaluated using standard tests. Whenever researchers formulate inequality constrained hypotheses, however, they must rely on sampling-based methods that are relatively inefficient and computationally expensive. To address this problem we developed a bridge sampling routine that allows an efficient evaluation of multinomial inequality constraints. An empirical application showcases that bridge sampling outperforms current Bayesian methods, especially when relatively little posterior mass falls in the restricted parameter space. The method is extended to mixtures between equality and inequality constrained hypotheses.

Keywords: Bayes factors, model selection, inequality constraints, Savage-Dickey density ratio

Introduction

In many scientific fields the analysis of categorical variables is of major importance. Applications range from the analysis of declared numeric values in forensic accounting, auditing, and fraud detection (Nigrini, 2012; Rauch, Göttsche, Brähler, & Engel, 2011), the analysis of descriptive measures in survey studies (e.g., Haberman, 1978; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016; Sedransk, Monahan, & Chiu, 1985; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014), the analysis of gut microbiome composition (Song, Zhao, & Wang, 2020), to the validation of model assumptions and axioms in the field of psychometrics (see e.g., Cavagnaro & Davis-Stober, 2014; Davis-Stober, 2009; Guo & Regenwetter, 2014; Myung, Karabatsos, & Iverson, 2005; Regenwetter, Dana, & Davis-Stober, 2011; Regenwetter et al., 2018; Tijmstra, Hoijtink, & Sijtsma, 2015). The breadth and depth of these examples underscore the importance of having efficient tools for their analysis readily available.

In each of the examples above, researchers are interested in quantifying evidence for hypotheses that impose certain restrictions on the underlying category proportions. These hypotheses often predict that all category proportions are exactly equal (e.g., the prevalence for a statistical reporting error is equal across different psychological journals; Veldkamp et al., 2014), or that they are fixed and follow a specific pattern (e.g., the digit proportions in non-fraudulent auditing data conform to Benford's law; Benford, 1938; Nigrini, 2012). However, research hypotheses also often stipulate ordinal expectations among category proportions (e.g., students with higher abilities have a higher chance to solve any particular item correctly; Grayson, 1988), or a mix of equality and inequality parameter constraints (e.g., according to the recognition heuristic, when laypeople predict which sports team will win a tournament they assign a higher probability of winning to more familiar teams and equal but lower probabilities to unknown teams; Goldstein & Gigerenzer, 2002).

Ordinal expectations about underlying category proportions are a regular occurrence in scientific theories. However, the evaluation of hypotheses that go beyond exact equality constraints is not very popular, particularly among researchers who use frequentist statis-

tics (Iverson, 2006). As motivating example, consider the study conducted by Uhlenhuth, Lipman, Balter, and Stern (1974), who surveyed 735 adults to investigate the association between symptoms of mental disorders and experienced life stress. To measure participants' life stress, the authors asked them to indicate, out of a list of negative life events, life stresses, and illnesses, which event they had experienced during the last 18 months prior to the interview. A subset of these data was reanalyzed by Haberman (1978, p. 3). Haberman noted that retrospective surveys tend to fall prey to the fallibility of human memory, causing participants to report primarily those negative events that happened most recently. He therefore investigated the 147 participants who reported only one negative life event over this time span and tested whether the frequency of the reported events was equally distributed over the 18 month period. However, Haberman did not directly test the ordinal pattern implied by his assumption of forgetting, namely that the number of reported negative life events decreases as a function of the time passed. Figure 1 shows the frequency of reported negative life events in Haberman's sample.

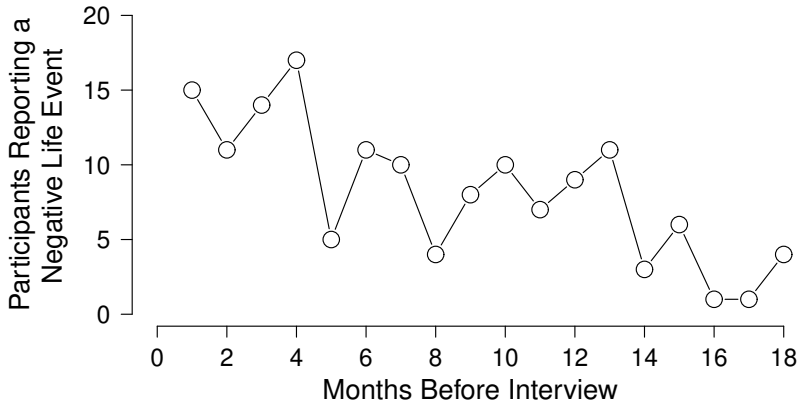


Figure 1. Frequency of reported negative life events over the course of the 18 months prior to the interview for Haberman's (1978) sample of the data collected by Uhlenhuth et al. (1974).

To evaluate ordinal multinomial patterns such as the one hypothesized by Haberman (1978) we focus on Bayesian methods. In the Bayesian statistical framework, researchers

may quantify the evidence for or against a specific restriction on the model parameters using the Bayes factor (Jeffreys, 1935; Kass & Raftery, 1995), that is, the relative predictive performance of the models with and without the restriction. For the usual scenario of equal or fixed underlying category proportions, the Bayes factor is available analytically. This is not the case, unfortunately, when inequality constraints are in play. In these cases, the Bayes factor can be approximated using popular methods such as the encompassing prior approach (Gu, Mulder, Deković, & Hoijtink, 2014; Klugkist, Kato, & Hoijtink, 2005; Hoijtink, Klugkist, & Boelen, 2008; Hoijtink, 2011) or the conditioning method (Mulder et al., 2009, 2009; Mulder, 2014, 2016). As we will show in the following sections, these methods estimate the Bayes factor by approximating the relative mass of the restricted parameter space through samples from the unrestricted distribution. Thus, they allow for convenient computations of Bayes factors. However, both methods have in common that the approximation of the Bayes factor becomes harder (i.e., more time-consuming and less accurate) as researchers are interested in a smaller part of the parameter space. The problem in these cases is that the probability of samples falling within the parameter space of the restricted distribution is very low, making it practically impossible to obtain accurate estimates of the Bayes factor by sampling from the unrestricted distribution. To illustrate this problem, consider a model with $K = 9$ categories, as is used, for example, in Benford tests to assess whether or not observed frequencies of leading digits (e.g., in auditing data) are manipulated or otherwise of poor quality (Nigrini, 2012). The hypothesis that the numbers are not manipulated predicts a decreasing trend in the proportions of the leading digits in accordance with Benford's law, that is $\theta_9 < \theta_8 < \cdots < \theta_1$. When drawing samples from a uniform prior for the $K = 9$ category probabilities, only 1 in $362,880$ samples will obey the restriction, since $\theta_9 < \theta_8 < \cdots < \theta_1$ is one of 9! ways in which 9 proportions can be ordered. When the prior proportion consistent with the restriction is so low, posterior samples in line with the restriction have a large impact on the Bayes factor. When we take $362,880$ posterior draws, the Bayes factor equals the number of samples that fall in the restricted area; thus, 5 posterior samples in the restricted area yields a Bayes factor of 5

in favor of the ordinal pattern of Benford's law, whereas 10 posterior samples in that area yield a Bayes factor of 10. This illustrates that in order to obtain a precise estimate of the Bayes factor, researchers need to draw millions of samples from the posterior. These ratios become increasingly problematic as the models become more complex (Sedransk et al., 1985; Mulder et al., 2009).

To overcome the above limitation we present a bridge sampling routine (e.g., Gronau et al., 2017; Meng & Wong, 1996) to estimate the Bayes factor for multinomial inequality constraints. The advantage of the bridge sampling routine is that its efficiency does not suffer when the size of the restricted parameter space decreases. The resulting Bayes factor estimates are relatively unbiased and precise. In addition, the bridge sampling approach has a fixed cost in terms of runtime, which makes it appealing for the implementation in standard statistical software packages.

The outline of this paper is as follows. First, we introduce the basic theoretical concepts of Bayesian parameter estimation and the computation of Bayes factors for the multinomial model featuring equality constrained hypotheses. We then extend these concepts to inequality constrained hypotheses. Third, we show how the bridge sampling approach compares to the established methods such as the encompassing prior approach and the conditioning method in terms of precision and efficiency by applying the methods to our motivating example. The last section contains a short discussion and the appendix generalizes the proposed methodology to a mixture of equality and inequality constrained hypotheses.

Bayesian Analysis of Multinomial Variables

This section introduces the theoretical concepts of Bayesian inference for the multinomial model, that is, Bayesian parameter estimation using posterior distributions and Bayesian hypothesis testing using Bayes factors. We denote the number of observations in a category k with x_k, and the total number of observations with $N = \sum_{k=1}^{K} x_k$. The multinomial distribution is a generalization of the binomial distribution to variables that can take values

in $K \geq 2$ categories, and it assigns the following probabilities to the different ways that N observations distribute across the K categories,

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = p(x_1, x_2, \ldots, x_K \mid \theta_1, \theta_2, \ldots, \theta_K) = \binom{N}{x_1, x_2, \ldots, x_K} \prod_{k=1}^{K} \theta_k^{x_k},$$

where the first factor in the likelihood denotes an extension of the binomial coefficient known as the multinomial coefficient. The parameters of the multinomial model, θ_k, reflect the probability of observing a value in a particular category, and need to sum to one. Note that due to the sum-to-one constraint, the K-th parameter is sometimes expressed as $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$.

Bayesian Parameter Estimation Without Inequality Constraints

Bayesian parameter estimation concerns the expression of a posterior distribution for model parameters capturing *a priori* information and information from the data (i.e., the likelihood). For the vector of probability parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, we choose a Dirichlet distribution with concentration parameters $(\alpha_1, \alpha_2, \ldots, \alpha_K)$, where each element in $\boldsymbol{\alpha}$ is larger than zero:

$$p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, \ldots, \theta_K) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}.$$

The concentration parameters α_k of the Dirichlet distribution have an intuitive interpretation: they may be interpreted as *a priori* category counts, and their exact values determine both the relative values of category probabilities and their variability. For the problem at hand, the posterior is also a Dirichlet distribution of the form

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(x_k + \alpha_k)} \prod_{k=1}^{K} \theta_k^{x_k + \alpha_k - 1},$$

with the updated concentration parameters $\alpha'_k = x_k + \alpha_k$ (O'Hagan & Forster, 2004). The concentration parameters of the posterior Dirichlet distribution can be interpreted as *a*

posteriori category counts, the sum of the prior and observed category counts.

Bayes Factor Hypothesis Testing Without Inequality Constraints

When stipulating exact equality constraints on the parameters of interest, researchers formulate a point null hypothesis $\mathcal{H}_0$ that assigns expected values $\mathbf{c}$ to the underlying category proportions $\boldsymbol{\theta}$, that is $\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{c}$. We first consider the Bayes factor

$$\text{BF}_{0e} = \frac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)},$$

where the hypothesis $\mathcal{H}_0$ stipulates exact values for all of the model parameters, i.e., $\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{c}$. In the standard multinomial test the null hypothesis states that all model parameters are exactly equal, thus, all elements in $\mathbf{c}$ are set equal to $1/K$. We test the null hypothesis against the encompassing hypotheses which states that all category proportions are free to vary without any ordinal restrictions. We call this hypothesis the encompassing hypothesis $\mathcal{H}_e$, since it encompasses all possible orders of the parameters. The parameter space of the encompassing hypothesis is denoted as $\mathcal{R}_e$. When stipulating exact equality constraints, it is assumed that there is no prior uncertainty about the model parameters, and the marginal likelihood of the null hypothesis is simply a multinomial distribution. Due to the conjugacy of the Dirichlet distribution to the parameters of the multinomial model, the marginal likelihood for the encompassing hypothesis has a simple, closed-form solution. Thus, if all model parameters of the null hypothesis are *a priori* specified, the Bayes factor BF_{0e} is equal to

$$\text{BF}_{0e} = \prod_{k=1}^{K} c_k^{x_k} \times \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \times \frac{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k + x_k)},$$

as derived already by Good (1967). There is another way to express the Bayes factor, which relates the Bayes factor to Bayesian parameter estimation. By rearranging Bayes' rule the marginal likelihood of the encompassing hypothesis can be expressed as:

$$\underbrace{p(\mathbf{x} \mid \mathcal{H}_e)}_{\substack{\text{marginal}\\ \text{likelihood}\\ \text{of } \mathcal{H}_e}} = \frac{\overbrace{p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{H}_e)}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\theta} \mid \mathcal{H}_e)}^{\substack{\text{prior}\\ \text{density}}}}{\underbrace{p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_e)}_{\substack{\text{posterior}\\ \text{density}}}},$$

which is known as Chib's identity (Chib, 1995). Chib's identity allows us to arrive at an alternative characterization of the Bayes factor that only requires the prior and posterior distribution under the alternative hypothesis at $\mathbf{c}$:

$$\text{BF}_{0e} = \frac{p(\mathbf{x} \mid \mathcal{H}_0)}{p(\mathbf{x} \mid \mathcal{H}_e)} = \frac{p(\mathbf{x} \mid \boldsymbol{\theta} = \mathbf{c}, \mathcal{H}_e)}{\frac{p(\mathbf{x} \mid \boldsymbol{\theta} = \mathbf{c}, \mathcal{H}_e)\, p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathbf{x}, \mathcal{H}_e)}} = \frac{\overbrace{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathbf{x}, \mathcal{H}_e)}^{\substack{\text{Height of posterior density of } \mathcal{H}_e\\ \text{at } \theta = c}}}{\underbrace{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)}_{\substack{\text{Height of prior density of } \mathcal{H}_e\\ \text{at } \theta = c}}}.$$

This expression is known as the Savage-Dickey density ratio (Dickey & Lientz, 1970; O'Hagan & Forster, 2004; Dickey, 1971; Verdinelli & Wasserman, 1995). The underlying principle of the Savage-Dickey density ratio is to compute the Bayes factor by dividing the height of the posterior density under $\mathcal{H}_e$ at the point of interest (i.e., $\mathbf{c}$) by the height of the prior density under $\mathcal{H}_e$ at the same point.

For concreteness, we will demonstrate the Bayesian multinomial test for exact equality constraints by reanalyzing the research question of Habermann (1978). The null hypothesis entails that the probability of reporting a negative life event is equally distributed over the 18 months prior to the interview. In particular, the expected category proportions under $\mathcal{H}_0$ are

$$\mathbf{c} : \theta_1, \theta_2, \cdots, \theta_K = 1/K.$$

Assuming that every parameter value is equally likely before we see any data, we assign a uniform prior distribution across the parameter vector $\boldsymbol{\theta}$, such that, $p(\boldsymbol{\theta} \mid \mathcal{H}_e) \sim \text{Dirichlet}(\boldsymbol{\alpha})$ with all concentration parameters set to 1. Using the observed frequencies from Haberman

(1978), that is,

$$\mathbf{x} = (15, 11, 14, 17, 5, 11, 10, 4, 8, 10, 7, 9, 11, 3, 6, 1, 1, 4)',$$

the Bayes factor in favor for the encompassing hypothesis is:

$$\text{BF}_{0e} = \frac{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)} = \frac{\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k + x_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k + x_k)} \prod_{k=1}^{K} \theta_k^{x_k + \alpha_k - 1}}{\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}} = \frac{1}{27.1}.$$

This result indicated that the data are about 27 times more likely under $\mathcal{H}_e$ (in which the parameters are free to vary) than under $\mathcal{H}_0$ (in which the parameters are constrained to be equal).

We have now outlined how to express the prior and posterior distribution for the multinomial model using a Dirichlet prior and expressed the Bayes factor in terms of the change of belief about the parameter value. A related expression for the Bayes factor can be derived in the case of inequality constraints, to which we turn next.

Bayesian Parameter Estimation With Inequality Constraints

When stipulating inequality-constrained hypotheses we can predict, for instance, an increasing trend of the first two categories, $\mathcal{H}_r : \theta_1 < \theta_2$. We refer to such inequality-constrained hypotheses as $\mathcal{H}_r$. Here, the parameter space, $\mathcal{R}_r$ is a subset of $\mathcal{R}_e$ by restrictions imposed on $\boldsymbol{\theta}$, that is, $\mathcal{R}_r = \{\boldsymbol{\theta} \in \mathcal{R}_e \,;\, \mathcal{H}_r\}$. The prior and posterior distributions of the parameters subject to an inequality-constrained hypothesis $\mathcal{H}_r$ thus take the following form:

$$p(\boldsymbol{\theta} \mid \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \tag{1}$$

$$p(\boldsymbol{\theta} \mid \mathbf{x},\, \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathbf{x},\, \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)}, \tag{2}$$

where $\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$ is an indicator function that is one for parameter values $\boldsymbol{\theta}$ in the restricted space $\mathcal{R}_r$ and zero otherwise. As apparent from the equations above, the prior and posterior distributions under an inequality-constrained hypothesis are proportional to their unconstrained counterparts. In principle, whenever the concentration parameters in the Dirichlet distribution are natural numbers for every k, thus, $\alpha_k \in \mathbb{N}$, we are able to achieve an exact result for the normalizing constants for the restricted prior and posterior distribution (see our online appendix for a description of the exact procedure). However, for general $\boldsymbol{\alpha}$ an exact result is not expected. Furthermore, the exact procedure is far more inefficient than sampling-based methods, especially as the number of categories in the model and the number of observations for a fixed K increases. Here, we were only able to obtain exact results for simple cases, involving models with no more than $K = 6$ categories and no more than $N = 63$ observations. For this reason, in the following we limit our descriptions of Bayesian parameter estimation and the computation of Bayes factors to sampling-based procedures.

In general researchers rely on Monte Carlo sampling methods to compute the normalizing constants of the restricted prior and posterior distribution. In the simplest case we can use rejection sampling to simulate values from the unconstrained prior and posterior distributions and only keep those values that conform to the restrictions. The proportion of the retained samples to the total number of samples is then an approximation for the normalizing constant of the restricted distribution. Unfortunately, when many inequality constraints are proposed, the approach outlined above, although intuitive, can be terribly inefficient. For instance, in the Haberman example, when drawing from a uniform prior only 1 in over $18! = 6.4 \times 10^{15}$ samples will obey the restriction. As an alternative, we can use a Markov chain Monte Carlo (MCMC) approach, that allows us through random variable transformation to simulate the values directly from the constrained distribution. Devroye (1986, p. 594), for instance, shows that one can simulate values from a Dirichlet distribution by first simulating K independent random variables γ_k with a Gamma(α_k, 1)

density, for $k = 1, \ldots, K$, and then setting

$$\theta_k = \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k}.$$

The variables θ_k that are generated in this way follow the desired Dirichlet($\boldsymbol{\alpha}$) distribution. Note that with the transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\gamma}$ the sum-to-one constraint is conveniently removed. Additionally, this MCMC method is suitable for drawing values from the restricted distribution because the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is order-preserving. Thus, an inequality-constrained hypothesis $\mathcal{H}_r : \theta_1 < \theta_2$ on the category probabilities translates into the inequality-constrained hypothesis $\mathcal{H}_r : \gamma_1 < \gamma_2$ on the gamma variables. If we simulate the gamma variables consistent with the order restrictions imposed by $\mathcal{H}_r$, that is, $p(\boldsymbol{\gamma} \mid \mathcal{H}_r)$, the transformed gamma variables then generate Dirichlet variables that are consistent with $\mathcal{H}_r$, that is, $p(\boldsymbol{\theta} \mid \mathcal{H}_r)$.

To draw gamma variables that obey the order imposed by the inequality-constrained hypotheses we use the Gibbs sampling algorithm proposed by Damien and Walker (2001). Their Gibbs sampling algorithm assumes fixed upper and lower bounds for each parameter. However, the algorithm can easily be generalized to cases where we wish to draw from gamma variables whose upper and lower bounds are not known, but are itself random variables (as it is the case for inequality-constrained hypotheses).

Instead of simulating values directly from the multivariate distribution of gamma variables that are subject to inequality constraints —$p(\boldsymbol{\gamma} \mid \mathcal{H}_r)$—, the Gibbs sampler operates by iteratively simulating values from the full-conditional posterior distributions, that is, the distribution of one gamma variable given the remaining gamma variables and inequality constraints —$p(\gamma_k \mid \boldsymbol{\gamma}^{(k)}, \mathcal{H}_r)$, where $\boldsymbol{\gamma}^{(k)}$ refers to the vector of gamma variables with the kth parameter removed. If there is no constraint on a gamma variable γ_k then the full conditional is simply the regular Gamma($\alpha_k, 1$) density. However, if γ_k is subject to a constraint, for instance, $\gamma_j < \gamma_k < \gamma_q$, then the gamma variable γ_k has the bounded support $[\gamma_j, \gamma_q]$ instead of $[0, \infty)$. This implies that the full conditional distribution of γ_k

subject to an inequality constraint is a truncated gamma distribution:

$$p(\gamma_k \mid \boldsymbol{\gamma}^{(k)}, \mathcal{H}_r) = p(\gamma_k \mid \gamma_j < \gamma_k < \gamma_q) = \frac{\frac{1}{\Gamma(\gamma_k)}\, \gamma_k^{\alpha_k - 1}\, e^{-\gamma_k}\, \mathbb{I}(\gamma_k \in [\gamma_j,\, \gamma_q])}{p(\gamma_k \in [\gamma_j,\, \gamma_q])}.$$

For gamma variables with bounded support $[\gamma_j,\, \gamma_q]$, the bounds at iteration t are calculated using the current values of the parameters. After the gamma variables have been simulated in this manner, they can be transformed back into category probabilities to obtain samples from the Dirichlet distribution. We provide the `R` code for the implementation of the Gibbs sampling algorithm in Appendix C.

Bayes Factor Hypothesis Testing for Inequality Constraints

We consider the Bayes factor

$$\text{BF}_{re} = \frac{p(\mathbf{x} \mid \mathcal{H}_r)}{p(\mathbf{x} \mid \mathcal{H}_e)},$$

where the hypothesis $\mathcal{H}_r$ stipulates inequality constraints on the model parameters, for instance,

$$\mathcal{H}_r : \theta_1 < \cdots < \theta_K.$$

In order to obtain the marginal likelihood of the inequality-constrained hypothesis we need to integrate over the restricted parameter space $\mathcal{R}_r$, which makes the Bayes factor BF_{re} difficult to compute:

$$p(\mathbf{x} \mid \mathcal{H}_r) = \int_{\mathcal{R}_e} p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{H}_r)\, \mathrm{d}\boldsymbol{\theta}.$$

It is nevertheless possible to arrive at an intuitive expression of the Bayes factor. This expression is a generalization of the Savage-Dickey density ratio mentioned above and follows from an alternative characterization of $p(\mathbf{x} \mid \mathcal{H}_r)$:

$$
\begin{aligned}
p(\mathbf{x} \mid \mathcal{H}_r) &= \int_{\mathcal{R}_e} p(\mathbf{x} \mid \boldsymbol{\theta},\, \mathcal{H}_e)\, p(\boldsymbol{\theta} \mid \mathcal{H}_r)\, \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\mathcal{R}_e} p(\mathbf{x} \mid \boldsymbol{\theta},\, \mathcal{H}_e)\, \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}\, \mathrm{d}\boldsymbol{\theta},
\end{aligned}
$$

where $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ does not depend on $\boldsymbol{\theta}$. Since $p(\mathbf{x} \mid \boldsymbol{\theta},\, \mathcal{H}_e)p(\boldsymbol{\theta} \mid \mathcal{H}_e) = p(\boldsymbol{\theta} \mid \mathbf{x},\, \mathcal{H}_e)p(\mathbf{x} \mid \mathcal{H}_e)$, we obtain the following result:

$$
\begin{aligned}
p(\mathbf{x} \mid \mathcal{H}_r) &= \int_{\mathcal{R}_e} p(\boldsymbol{\theta} \mid \mathbf{x},\, \mathcal{H}_e)\, p(\mathbf{x} \mid \mathcal{H}_e) \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)\, \mathrm{d}\boldsymbol{\theta}\, \frac{1}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \\
&= \int_{\mathcal{R}_e} p(\boldsymbol{\theta} \mid \mathbf{x},\, \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)\, \mathrm{d}\boldsymbol{\theta}\, \frac{p(\mathbf{x} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \\
&= \frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)\, p(\mathbf{x} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)},
\end{aligned}
$$

which was derived in Klugkist et al. (2005). With this characterization the Bayes factor amounts to

$$
\mathrm{BF}_{re} = \frac{\overbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x},\, \mathcal{H}_e)}^{\substack{\text{Proportion of posterior parameter} \\ \text{space consistent with the restriction}}}}{\underbrace{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)}_{\substack{\text{Proportion of prior parameter} \\ \text{space consistent with the restriction}}}}. \tag{3}
$$

Like the Savage-Dickey density ratio, this presents the Bayes factor as the change of belief that the parameters lie in the restricted parameter space $\mathcal{R}_r$ (see also Wetzels, Grasman, & Wagenmakers, 2010 and Mulder et al., 2009). We discuss two established procedures to approximate the Bayes factor BF_{re} in the next section.

Established Procedures to Estimate the Bayes Factor For Inequality-Constraints

One popular method to estimate the Bayes factor for inequality-constrained hypotheses is the encompassing prior approach which relies on simple Monte Carlo estimates (Gelfand, Smith, & Lee, 1992; Klugkist et al., 2005; Sedransk et al., 1985). This method estimates the Bayes factor in Equation 3 by considering the proportion of the prior and posterior distributions of the unrestricted distribution that are in agreement with the constraints. That is, the numerator can be estimated by sampling from the encompassing posterior density and then calculating the proportion of draws in accordance with the restrictions imposed by the inequality-constrained hypothesis. Likewise, the denominator can be estimated by sampling from the encompassing prior density and then calculating the proportion of draws in accordance with the restrictions:

$$\begin{aligned} \text{BF}_{re} &= \frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \\ &\approx \frac{\frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\boldsymbol{\theta}'_s \in \mathcal{R}_r)}{\frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\boldsymbol{\theta}^*_s \in \mathcal{R}_r)}, \end{aligned}$$

where $\boldsymbol{\theta}^*_s$ and $\boldsymbol{\theta}'_s$ denote the s-th sample from the encompassing prior and posterior distribution, respectively, for samples $s = 1, \ldots, S$. Because of the simplicity of the method, numerous applications can be found in the literature, for example in the context of multinomial models (Heck & Davis-Stober, 2019), but also applied to the analysis of contingency tables and to the analysis of variance and covariance models (Hoijtink et al., 2008; Hoijtink, 2011; Klugkist, Laudy, & Hoijtink, 2010), item response theory (Haaf, Merkle, & Rouder, submitted), as well as to Bayesian linear mixed models (Haaf & Rouder, 2017). However, it is also widely recognized that this method is not particularly efficient for models with an increasing number of constraints (Myung, Karabatsos, & Iverson, 2008; Sedransk et al., 1985). The same holds true for models with a small number of constraints that are extremely restrictive or models for which the data do not align with the inequality-constrained hypothesis. This is the case because the efficiency of the method relies on the relative size of

the restricted area: if prior and posterior samples almost never fall inside the area of interest, a large number of samples is required to estimate the proportions accurately (Gelfand et al., 1992; Hoijtink, 2011).

A method that is more stable for larger models is the conditioning method (Mulder et al., 2009; for an application to multinomial models, see Heck & Davis-Stober, 2019). The conditioning method also utilizes the identity in Equation 3. But instead of estimating the normalizing constants of the constrained distribution based on a single set of samples from the encompassing distribution, Mulder et al. (2009) proposed a stepwise approach. For instance, when evaluating a hypotheses concerning $K = 4$ ordered parameters $\mathcal{H}_r : \theta_1 < \theta_2 < \theta_3 < \theta_4$, the proportion of prior parameter space consistent with the restriction can be factored as follows:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = p(\theta_1 < \theta_2 \mid \mathcal{H}_e) \times p(\theta_2 < \theta_3 \mid \theta_1 < \theta_2, \mathcal{H}_e) \times p(\theta_3 < \theta_4 \mid \theta_1 < \theta_2 < \theta_3, \mathcal{H}_e).$$

The proportion of posterior samples consistent with the restriction are estimated in a similar fashion, which yields the Bayes factor:

$$\begin{aligned} \mathrm{BF}_{re} &= \frac{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)} \\ &= \frac{p(\theta_1 < \theta_2 \mid \mathbf{x}, \mathcal{H}_e) \times \cdots \times p(\theta_3 < \theta_4 \mid \theta_1 < \theta_2 < \theta_3, \mathbf{x}, \mathcal{H}_e)}{p(\theta_1 < \theta_2 \mid \mathcal{H}_e) \times \cdots \times p(\theta_3 < \theta_4 \mid \theta_1 < \theta_2 < \theta_3, \mathcal{H}_e)}, \end{aligned}$$

where each Bayes factor is estimated independently. By evaluating the constraints sequentially, the conditioning method yields better results for models featuring larger numbers of constraints (Mulder et al., 2009). A similar method was proposed to evaluate almost-equality constraints: using the transitivity property of the Bayes factor, Laudy (2006, p. 115) and Klugkist (2008) proposed to approximate the Bayes factor for almost-equality constraints by evaluating a series of hypotheses of increasing narrowness, such that for each pair

of parameters $\theta_1 \approx \theta_2$ the distance between the them approaches zero (i.e., $\mid \theta_1 - \theta_2 \mid \rightarrow 0$).[1] However, care must be taken not to set the values for the distance $\mid \theta_1 - \theta_2 \mid$ too small, otherwise there is a risk of restricting the parameter space to such an extent that the efficiency of the method is negatively affected (Klugkist, 2008).

The increased stability of the conditioning method is accompanied by a steep increase in runtime. This increase has three reasons. The first reason follows directly from the sequential evaluation of the individual constraints. To refer again to the Haberman example: Since the associated model features seventeen constraints, seventeen sets of prior and posterior samples must be drawn for the evaluation. The resulting runtime is thus seventeen times higher than that of the encompassing prior approach. The second reason is that even though the method is more stable, there is still the risk that the relative size of the restricted area for each individual restriction is too small to effectively sample from it. The third reason is the implementation of the conditioning method. When evaluating the individual constraints it is not enough to simply draw samples from the encompassing distribution; this is only possible for the first constraint. For each additional constraint, samples are drawn from distributions that are conditional on previous constraints, with a new constraint added at each step. Thus, we need to draw samples from restricted distributions using MCMC methods, that are slower than the standard Monte Carlo methods used in the encompassing prior approach.

A Bridge Sampling Routine to Estimate the Bayes Factor

The main limitations of the encompassing prior approach and the conditioning method–lack of precision, lack of scalability, and long runtimes–come from the effort to estimate the proportion of the encompassing parameter space in accordance with the constraint. In contrast, bridge sampling (Bennett, 1976; Meng & Wong, 1996) estimates normalizing constants using a different approach. The basic principle of bridge sampling is that the ratio between two normalizing constants operating on the same parameter space can be

[1] Wetzels et al. (2010) showed that the proposed almost-equality constraint method approximates the Savage-Dickey density ratio.

estimated by the following identity:

$$\frac{p(\mathbf{x} \mid \mathcal{H}_1)}{p(\mathbf{x} \mid \mathcal{H}_2)} = \frac{\mathbb{E}_{\mathcal{H}_2}\left(p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{H}_1) p(\boldsymbol{\theta} \mid \mathcal{H}_1) h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\mathcal{H}_1}\left(p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{H}_2) p(\boldsymbol{\theta} \mid \mathcal{H}_2) h(\boldsymbol{\theta})\right)},$$

where the term $h(\boldsymbol{\theta})$ refers to an arbitrary bridge function that ensures that the denominator is non-zero. Here we use a slightly modified form of the bridge identity proposed by Overstall and Forster (2010) which estimates not a ratio but a single normalizing constant to further increase the precision of the estimates. The modified form of the bridge identity requires that the second distribution (denoted above as the distribution under $\mathcal{H}_2$) is replaced by a distribution with sufficient overlap to the target distribution and with a known normalizing constant. In the following, we will refer to this distribution as proposal distribution $g(\boldsymbol{\theta})$. The modified identity then becomes:

$$p(\mathbf{x} \mid \mathcal{H}_1) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{H}_1) h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\mathcal{H}_1}\left(g(\boldsymbol{\theta}) h(\boldsymbol{\theta})\right)}, \tag{4}$$

where $p(\mathbf{x} \mid \mathcal{H}_1)$ indicates a normalizing constant we wish to estimate, that is, the normalizing constant of the constrained prior distribution, or the normalizing constant of the constrained posterior distribution, that is, $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ or $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)$, respectively. Since these normalizing constants are of the form

$$\int_{\mathcal{R}_r} p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathrm{d}\boldsymbol{\theta} \quad \text{and} \quad \int_{\mathcal{R}_r} p(\boldsymbol{\theta} \mid \mathbf{x}, \mathcal{H}_e)\, \mathrm{d}\boldsymbol{\theta}$$

the bridge sampler can be used to estimate them, if the support of the proposal $g(\boldsymbol{\theta})$ is $\mathcal{R}_r$. To arrive at the expression for the bridge sampling identity for the normalizing constant of the constrained prior distribution we now simply replace the terms related to

$\mathcal{H}_1$. Specifically, since

$$p(\boldsymbol{\theta} \mid \mathcal{H}_r) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)}{p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)},$$

we can replace the term for the unnormalized density under $\mathcal{H}_1$ in the numerator of Equation 4 (i.e., $p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{H}_1)$) by the corresponding term for the constrained prior distribution, that is, $p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)$. Thus, the resulting bridge sampling identity can be described as follows:

$$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}\left(p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r)h(\boldsymbol{\theta})\right)}{\mathbb{E}_{\text{prior}}\left(g(\boldsymbol{\theta})h(\boldsymbol{\theta})\right)}. \tag{5}$$

The normalizing constant for the constrained posterior distribution can be described similarly. Based on this identity, we can now define the corresponding estimator. We substitute the expectations by sample averages, using N_1 samples from the constrained prior distribution, that is, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta} \mid \mathcal{H}_r)$ and N_2 samples from a suitable proposal distribution, that is $\tilde{\boldsymbol{\theta}} \sim g(\boldsymbol{\theta})$. Then, we can estimate $p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)$ by:

$$\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) \approx \frac{\frac{1}{N_2}\sum_{m=1}^{N_2} p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)h(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\frac{1}{N_1}\sum_{n=1}^{N_1} g(\boldsymbol{\theta}^*_{\boldsymbol{n}})h(\boldsymbol{\theta}^*_{\boldsymbol{n}})}. \tag{6}$$

There are many possible choices for $h(\boldsymbol{\theta})$. Meng and Wong (1996) suggested the use of a bridge function that has been shown to minimize the relative mean square error of the estimate. However, when following this recommendation, the specific choice for $h(\boldsymbol{\theta})$ depends on the unknown normalization constant:

$$h(\boldsymbol{\theta}) = c \times \frac{1}{s_1 p(\boldsymbol{\theta} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta} \in \mathcal{R}_r) + s_2 p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e) g(\boldsymbol{\theta})},$$

where $s_1 = \frac{N_1}{N_2+N_1}$, $s_2 = \frac{N_2}{N_2+N_1}$ and c is a constant that has no influence on the results. To be able to estimate the normalizing constant of the restricted prior distribution we use the iterative scheme proposed by Meng and Wong (1996). Thus, we yield the following formula for the bridge sampling estimator at iteration $t + 1$:

$$\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} \approx \frac{\frac{1}{N_2} \sum_{m=1}^{N_2} \frac{\ell_{2,m}}{s_1 \ell_{2,m} + s_2 p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}}{\frac{1}{N_1} \sum_{n=1}^{N_1} \frac{1}{s_1 \ell_{1,n} + s_2 p(\boldsymbol{\theta}^*_{\boldsymbol{n}} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)}}},$$

where $\ell_{1,n} = \frac{p(\boldsymbol{\theta}^*_{\boldsymbol{n}} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta}^*_{\boldsymbol{n}} \in \mathcal{R}_r)}{g(\boldsymbol{\theta}^*_{\boldsymbol{n}})}$ and $\ell_{2,m} = \frac{p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}$. We then run the iterative scheme until a predefined tolerance criterion is reached. We follow the suggestion by Gronau et al. (2017) to use a tolerance criterion of

$$\frac{\mid \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)} - \hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t)} \mid}{\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(t+1)}} \leq 10^{-10},$$

while setting $\hat{p}(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathcal{H}_e)^{(1)} = 0$ as initial guess.

Transformations To Facilitate Bridge Sampling

Since the bridge function is defined on the common support of the proposal and target distribution, both distributions have to operate on the same parameter space. In addition, the normalizing constant of the proposal distribution must be known, which means that we cannot choose another constrained Dirichlet distribution. To resolve this problem we move the prior and posterior draws from the probability space to the real line using a probit transformation. This transfomation aims to eliminate the constraints inherent to the restricted Dirichlet distribution, namely the sum-to-one constraint and the inequality constraints. Furthermore, the transformation enables us to choose a convenient proposal distribution that is easy to sample from and easy to evaluate, for instance, the multivariate

normal distribution (Overstall & Forster, 2010).

The general idea is as follows: $\boldsymbol{\theta}$ is a probability vector, therefore, its elements must sum to one. As a result, the vector is completely determined by its first $K-1$ elements. For the transformation we therefore only consider the first $K-1$ elements and transform them to $K-1$ elements of a new vector $\boldsymbol{\xi}$ with $\boldsymbol{\xi} \in \mathbb{R}^{K-1}$. To retain the inequality constraints imposed on the parameters, we need to account for the lower bound l_k and the upper bound u_k of each θ_k. These bounds can be determined by adapting a stick-breaking approach (Frigyik, Kapila, & Gupta, 2010; Stan Development Team, 2020). The stick-breaking approach represents $\boldsymbol{\theta}$ as a stick of length one which we subsequently break into K elements. Assuming $\theta_{k-1} < \theta_k$, for $k \in \{1 \cdots, K\}$, the lower bound for any element in $\boldsymbol{\theta}$ is defined as

$$l_k = \begin{cases} 0 & \text{if } k = 1 \\ \theta_{k-1} & \text{if } 1 < k < K. \end{cases} \tag{7}$$

The upper bound is defined as

$$u_k = \begin{cases} \dfrac{1}{K} & \text{if } k = 1 \\ \dfrac{1 - \sum_{j<k} \theta_j}{K + 1 - k} & \text{if } 1 < k < K, \end{cases} \tag{8}$$

where $1 - \sum_{j<k} \theta_j$ represents the length of the remaining stick and $K + 1 - k$ is the number of elements in the remaining stick. Let ϕ denote the density of a normal variable with a mean of zero and a variance of one, Φ its cumulative density function, and Φ^{-1} its inverse cumulative density function. Then, the transformation of $\boldsymbol{\theta}$ is given by:

$$\xi_k = \Phi^{-1}\left(\frac{\theta_k - l_k}{u_k - l_k}\right).$$

$$= \begin{cases} \Phi^{-1}\left(\dfrac{\theta_k}{1/K}\right) & \text{if } k = 1 \\ \Phi^{-1}\left(\dfrac{\theta_k - \theta_{k-1}}{\dfrac{1 - \sum_{j<k}\theta_j}{K+1-k} - \theta_{k-1}}\right) & \text{if } 1 < k < K-1. \end{cases}$$

The inverse transformation is given by:

$$\theta_k = (u_k - l_k)\Phi(\xi_k) + l_k$$

$$= \begin{cases} \dfrac{1}{K}\Phi(\xi_k) & \text{if } k = 1 \\ \left(\dfrac{1 - \sum_{j<k}\theta_j}{K+1-k} - \theta_{k-1}\right)\Phi(\xi_k) + \theta_{k-1} & \text{if } 1 < k < K. \end{cases}$$

In the inverse transformation θ_k depends only on the first k elements of $\boldsymbol{\xi}$. Therefore, we know that the Jacobian matrix will be lower triangular, and the determinant of the Jacobian matrix will be the product of the diagonal entries given by:

$$\frac{\partial \theta_k}{\partial \xi_k} = \begin{cases} \dfrac{1}{K}\phi(\xi_k) & \text{if } k = 1 \\ (u_k - l_k)\,\phi(\xi_k) & \text{if } 1 < k < K. \end{cases}$$

Therefore, the Jacobian of this transformation is:

$$|J| = \frac{1}{K}\phi(\xi_1)\prod_{k=2}^{K-1}\left((u_k - l_k)\,\phi(\xi_k)\right).$$

Taking this transformation into account the bridge sampling estimator computes $\ell_{1,n}$ and $\ell_{2,m}$ as follows:

$$\ell_{1,n} = \frac{p(\boldsymbol{\theta}_{\boldsymbol{n}}^{*} \mid \mathcal{H}_e)\mathbb{I}(\boldsymbol{\theta}_{\boldsymbol{n}}^{*} \in \mathcal{R}_r)}{g(\boldsymbol{\xi}_{\boldsymbol{n}}{}^{*})},$$
$$\ell_{2,m} = \frac{p(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \mid \mathcal{H}_e)\mathbb{I}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} \in \mathcal{R}_r)}{g(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}})},$$

where $\boldsymbol{\xi}_{\boldsymbol{n}}{}^{*} = \Phi^{-1}\left(\frac{\boldsymbol{\theta}_{\boldsymbol{n}}^{*} - \mathbf{l}}{\mathbf{u} - \mathbf{l}}\right)$, and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} = ((\mathbf{u} - \mathbf{l})\Phi(\tilde{\boldsymbol{\xi}}_{\boldsymbol{m}}) + \mathbf{l})\,|J|)$.

Taken together, to apply the proposed bridge sampling routine the following three conditions must be met. First, we need to be able to sample directly from the constrained prior and posterior densities, which can be achieved by using the adapted version of the Gibbs sampling method by Damien and Walker (2001) described above. Second, we need to select a suitable proposal distribution for the bridge sampling algorithm; here we choose a multivariate normal distribution that achieves sufficient overlap with our target distribution by moving the samples from the restricted Dirichlet distribution to the real line. Third, we need to choose a bridge function; here, we have chosen the bridge function proposed in Meng and Wong (1996) which has the favorable property that it minimizes the estimated relative mean-squared error.

Given that bridge sampling only requires draws of the restricted distribution and the proposal distribution, this method is more efficient than the encompassing prior approach (because fewer samples are typically needed) and the conditioning method (because fewer instances of the Gibbs sampler are needed). In addition, the precision of the bridge sampling estimator depends not on the relative size of the restricted parameter space, but on the overlap between the target and proposal distribution; when the proposal distribution resembles the target distribution more closely, the resulting estimates are more accurate

(Meng & Wong, 1996).

Empirical Application: Memory of Negative Life Events

In this section we investigate the precision and efficiency of the estimation methods when applied to a real data set published in Uhlenhuth et al. (1974). Specifically, we conduct a Bayesian reanalysis of Haberman's sample to test whether the reported negative life events decrease over time as a function of forgetting. We test this inequality-constrained hypothesis against the encompassing hypothesis without constraints:

$$\mathcal{H}_r : \theta_1 > \theta_2 > \cdots > \theta_{18}$$
$$\mathcal{H}_e : \theta_1, \theta_2, \cdots, \theta_{18}.$$

Method

We estimated the Bayes factor using the bridge sampling approach, the encompassing approach, and the conditional approach. We computed Bayes factors in favor of $\mathcal{H}_r$ 100 times for the same data set and for each method and recorded the respective values and the runtime to produce a result. We assigned a uniform prior distribution to our parameters of interest, such that we could compute the prior probability of the constraint, $p(\boldsymbol{\theta} \in \mathcal{R} \mid \mathcal{H}_e)$, analytically. For the bridge sampling method, we drew $20,000$ samples from the constrained posterior distribution. For the conditioning method the marginal probabilities of each constraint holding were estimated using $40,000$ draws from the posterior distribution, resulting in a total of $40,000 \times 18$ draws. For the encompassing prior approach, we drew 5 million samples from the unconstrained posterior distribution.

Results

The estimated Bayes factors BF_{re} are displayed in Figure 2. Bayes factors based on the bridge sampling method and the conditioning method are centered around the same value

($M = 168.88$ and $M = 168.55$, respectively); however, the bridge sampling estimates varied far less ($SD = 1.873$) than the estimates produced by the conditioning method ($SD = 22.23$). To understand the reasons for these differences in variability, we investigated the autocorrelation and the influence of chain length on the Bayes factor estimates, but could not identify a consistent pattern.

The encompassing prior approach failed to estimate any Bayes factor, that is, for each iteration none of the 5 million posterior draws were in accordance with the constraint. This is not too surprising; the prior probability of samples obeying the constraint is already 1.3 billion times lower than the number of posterior samples drawn (i.e., $1/18!$). Thus, for the present example the encompassing prior approach can be applied only with great investment of time.

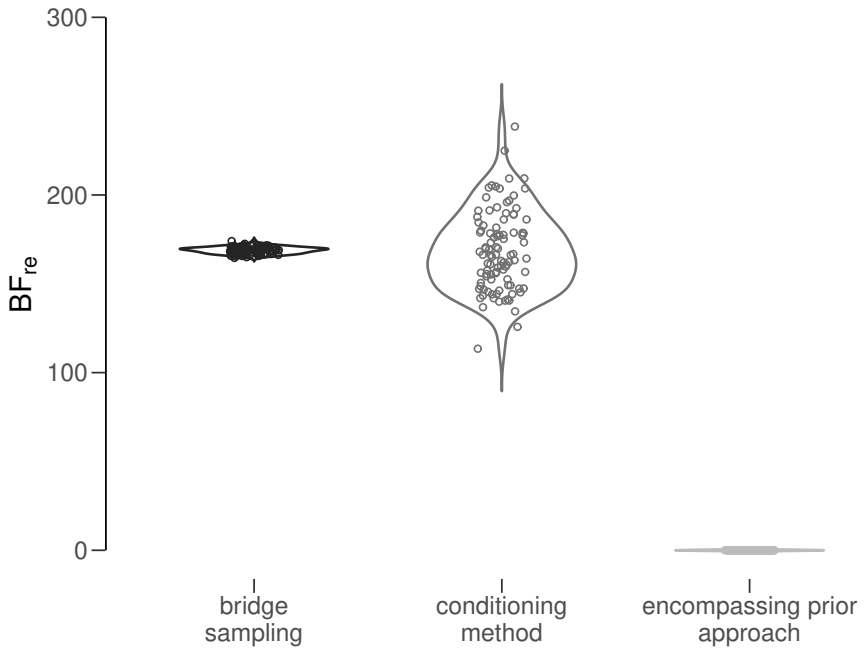


Figure 2. Bayes factors for the bridge sampling method (black), the conditioning method (dark grey), and the encompassing prior approach (light grey) for the test of an order-restriction in Haberman's (1978) data on the reporting of negative life events. Each dot represents one Bayes factor estimate in favor of $\mathcal{H}_r$ obtained by the respective method. The bridge sampling method yields more precise Bayes factor estimates than the conditioning method; the encompassing prior approach fails to estimate any Bayes factor.

The computation times are displayed in Figure 3. Regarding the computational efficiency, the bridge sampling method had the lowest runtimes with a mean of $M = 29.11 (SD = 0.39)$ seconds. The encompassing prior approach had comparable runtimes ($M = 35.89$, $SD = 0.22$). The conditioning method required the most time, with mean runtimes of $M = 375.84$ ($SD = 5.04$) seconds, which is more than 6 minutes to estimate one Bayes factor, compared to less than half a minute for the bridge sampling method.

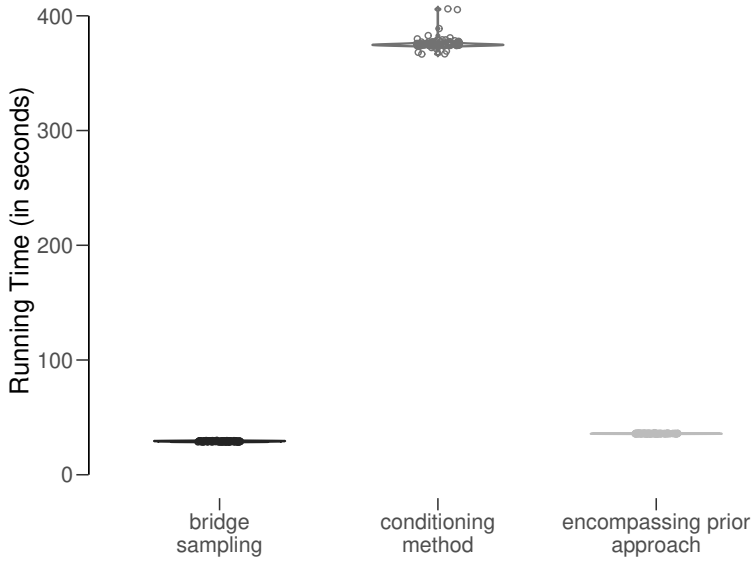


Figure 3. Runtime for the bridge sampling method (black) is similar to that of the encompassing prior approach (light grey), whereas the conditioning method (dark grey) has much higher computational costs. However, even though the runtime for the bridge sampling method and the encompassing prior approach is similar, the latter method failed to estimate any Bayes factors.

In sum, the empirical example demonstrates that the bridge sampling routine outperforms both the conditioning method and the encompassing prior approach. The bridge sampling estimates are considerably more precise than those of the conditioning method, and are obtained more quickly. The encompassing prior approach fails to estimate any Bayes factor altogether.

This example also illustrates how vulnerable the encompassing prior approach is to an increase in model size: even though the data strongly supported the inequality-constrained

hypothesis over the encompassing hypothesis, none of 5 million posterior draws across 100 replications (for a total of 500 million draws) obeyed all of the inequality constraints. Note that, if for any replication a single posterior draw had obeyed the restriction (i.e., 1 out of 5 million) the estimated Bayes factor in favor of the inequality-constrained hypothesis would have been 1.28×10^9 (i.e., a staggering overestimate), as the prior probability of a sample obeying the restriction is minuscule.

Discussion

In this paper we describe a precise, scalable, and efficient bridge sampling routine to estimate Bayes factors for inequality constrained hypotheses on multinomial data. Bridge sampling is a promising alternative to current methods that sample from the unconstrained parameter space and hence may yield imprecise results and long runtimes.

The main reason why the bridge sampling method achieves relatively high precision –even for a model with many categories– is that it does not sample from the unconstrained parameter space. Instead, bridge sampling combines the draws from the restricted target distribution with samples from a proposal distribution to estimate the marginal likelihood efficiently. As a result, the precision of the bridge sampling estimate does not depend on the prior probability of the constraint, but rather depends on the similarity between the proposal distribution and the target distribution. Meng and Schilling (2002, p. 584) note that by using more sophisticated methods (e.g., by using warp bridge sampling) to create more overlap between the proposal distribution and the target distribution "[...] we can achieve better and better estimation efficiency based on the same set of draws, and it seems there is no lower bound on the Monte Carlo error". To achieve sufficient overlap between the two distributions, we applied random variable transformation and used the method of moments to construct a suitable proposal distribution.

Compared to existing methods, the bridge sampling routine requires more effort to implement. As with the conditioning method, researchers who wish to use bridge sampling to evaluate inequality constrained hypotheses need to implement a Gibbs sampling algo-

rithm to draw samples from the constrained prior and posterior distribution. In addition, functions must be implemented to perform the required variable transformations and to apply the bridge sampling algorithm. In order to maximize the accessibility of the proposed method, we therefore provide the `R` code to conduct the analysis in our online appendix. In the near future we plan to make the analysis available in the user-friendly statistical software program JASP (JASP Team, 2020), which does not require any programming experience whatsoever.

The method proposed here is relatively general and may be extended to problems of higher dimension and increasing sophistication. For instance, our bridge sampling routine can be easily generalized to hypotheses that feature a combination of equality and inequality constrained parameters, as well as parameters that are free to vary. We describe this extension in Appendix A. Moreover, the bridge sampling framework could be expanded to multinomial models with complex linear restrictions (e.g., Heck & Davis-Stober, 2019). This would allow researchers to test more complex hypotheses, such as ordinal expectations on the size ratio of the parameters of interest (e.g., $\mathcal{H}_r : \theta_1 > 3 \times \theta_2$), on the differences between category proportions (e.g., $\mathcal{H}_r : (\theta_1 - \theta_2) < (\theta_3 - \theta_4)$), or on odds ratios for data that are summarized in contingency tables (e.g., $\mathcal{H}_r : \frac{\theta_1}{(\theta_1+\theta_2)} < \frac{\theta_3}{(\theta_3+\theta_4)}$). To apply bridge sampling to these models, one needs to adapt the random variable transformations to move $\boldsymbol{\theta}$ to the real line while retaining the imposed constraints. Another generalization of the presented methods concerns the application to hierarchical models, for cases where participants repeatedly choose a response option and therefore category proportions are nested within participants.

Our results demonstrate that bridge sampling offers considerable improvements in precision and efficiency over existing methods. As our empirical application showed, for multinomial models it is common to have a relatively high number of categories (i.e., $K > 10$) which can easily lead to extreme values of the Bayes factors, if the data either speak for or against the restriction. In other disciplines, such as microbiology, we even find multinomial models with up to $K = 46$ categories, as a study of the relationship between gut microbiome

and BMI showed (Song et al., 2020). In these scenarios we believe that the benefit of the bridge sampling routine is particularly apparent. To conclude, the bridge sampling routine of estimating Bayes factors for inequality constraints in multinomial models constitutes a promising tool to evaluate ordinal expectations reliably and efficiently.

Disclosures

Data, and Code

Readers can access the data from the empirical example, and the R code all analyses (including the creation of all figures), in our OSF folder at: `https://osf.io/59tce/`.

Acknowledgements

This research was supported by a Netherlands Organisation for Scientific Research (NWO) grant to AS (406-17-568) and to QFG (406-16-528), a Veni grant from the NWO to MM (451-17-017), as well as a Vici grant from the NWO to EJW (016.Vici.170.083).

References

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*, 245–268.

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, *1*, 102–122.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Damien, P., & Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, *10*, 206–215.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th Conference on Winter Simulation* (pp. 260–265).

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*, 204–223.

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). *Introduction to the Dirichlet distribution and related processes* (Tech. Rep.). Department of Electrical Engineering, University of Washington.

Gelfand, A. E., Smith, A. F., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *29*, 399–431.

Grayson, D. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.

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.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*, 511-527.

Guo, Y., & Regenwetter, M. (2014). Quantitative tests of the perceived relative argument model: Comment on Loomes (2010). *Psychological Review*, *121*, 696–705.

Haaf, J. M., Merkle, E. C., & Rouder, J. N. (submitted). *Do items order? The psychology in IRT models.* Retrieved from `https://psyarxiv.com/xsrfb` (Submitted for publication)

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*, 779–798.

Haberman, S. J. (1978). *Analysis of qualitative data: Introductory topics* (Vol. 1). Academic Press.

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I., & Boelen, P. (Eds.). (2008). *Bayesian evaluation of informative hypotheses.* New York: Springer Verlag.

Iverson, G. J. (2006). An essay on inequalities and order-restricted inference. *Journal of Mathematical Psychology*, *50*, 215–219.

JASP Team. (2020). *JASP (Version 0.13.1.0) [Computer software].* `https://jasp-stats.org/`.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In (Vol. 31, pp. 203–222).

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Klugkist, I. (2008). Encompassing prior based model selection for inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 53–83). New York: Springer Verlag.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, *15*, 281–299.

Laudy, O. (2006). *Bayesian inequality constrained models for categorical data* (Unpublished doctoral dissertation). Utrecht University.

Meng, X.-L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, *11*, 552–586.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463.

Mulder, J. (2016). Bayes factors for testing order–constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530–546.

Mulder, J., Wagenmakers, E.-J., & Marsman, M. (in press). A generalization of the Savage-Dickey density ratio for testing equality and order constrained hypotheses. *The American Statistician*.

Myung, J., Karabatsos, G., & Iverson, G. (2008). A statistician's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131 – 154). Berlin: Springer Verlag.

Myung, J., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). Hoboken, New Jersey: John Wiley & Sons.

Nuijten, M. B., Hartgerink, C. H., van 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.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.)*. London: Arnold.

Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*, 3269–3288.

Pericchi Guerra, L., Liu, G., & Torres, D. (2008). Objective Bayes factors for informative hypotheses: "completing" the informative hypothesis and "splitting" the Bayes factors. In H. Hoijtink,

I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131 – 154). Berlin: Springer Verlag.

Rauch, B., Göttsche, M., Brähler, G., & Engel, S. (2011). Fact and fiction in EU-governmental economic data. *German Economic Review*, *12*, 243–255.

Regenwetter, M., Cavagnaro, D. R., Popova, A., Guo, Y., Zwilling, C., Lim, S. H., & Stevens, J. R. (2018). Heterogeneity and parsimony in intertemporal choice. *Decision*, *5*, 63–94.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.

Robertson, T. (1978). Testing for and against an order restriction on multinomial parameters. *Journal of the American Statistical Association*, *73*, 197–202.

Sedransk, J., Monahan, J., & Chiu, H. (1985). Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *47*, 519–527.

Song, Y., Zhao, H., & Wang, T. (2020). An adaptive independence test for microbiome community data. *Biometrics*, *76*, 414–426.

Stan Development Team. (2020). Stan modeling language user's guide and reference manual, version 2.23.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Tijmstra, J., Hoijtink, H., & Sijtsma, K. (2015). Evaluating manifest monotonicity using Bayes factors. *Psychometrika*, *80*, 880–896.

Uhlenhuth, E. H., Lipman, R. S., Balter, M. B., & Stern, M. (1974). Symptom intensity and life stress in the city. *Archives of General Psychiatry*, *31*, 759–764.

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE*, *9*, e114876.

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.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics & Data Analysis*, *54*, 2094–2102.

Appendix A

Bayes Factors for Mixed Constraints

In addition to pure equality-constrained and pure inequality-constrained hypotheses, researchers may want to specify hypotheses with some parameters that are exactly equal to each other while others can vary freely and again others are ordered (see e.g., Pericchi Guerra, Liu, & Torres, 2008). However, it is not intuitively clear how to compute Bayes factors when parametric constraints are mixed. Without loss of generality, we first consider a mixed hypothesis $\mathcal{H}_m$ where the first j category parameters are constrained to be exactly equal and where the remaining $K - j$ parameters are increasing:

$$\mathcal{H}_m : (\theta_1 = \theta_2 = \cdots = \theta_j) < \theta_{j+1} < \cdots < \theta_K.$$

As shown in Equation (3), the Bayes factor of restricted hypotheses against the encompassing hypothesis can be formulated as

$$\text{BF}_{me} = \frac{p(\boldsymbol{\theta} \in \mathcal{R}_m \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta} \in \mathcal{R}_m \mid \mathcal{H}_e)}.$$

The mixed hypothesis stipulates the following set of constraints

$$\mathcal{R}_m : (\theta_1 = \cdots = \theta_j) \cap (\theta_j < \cdots < \theta_K) = \mathcal{R}_0 \cap \mathcal{R}_r.$$

The first set of constraints, which we denote with $\mathcal{R}_0$, are the equality constraints, and the second set of constraints, which we denote with $\mathcal{R}_r$, are the inequality constraints. Using this notation, the Bayes factor can be reformulated as

$$\text{BF}_{me} = \underbrace{\frac{p(\boldsymbol{\theta}_r \in \mathcal{R}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0, \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta}_r \in \mathcal{R}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0, \mathcal{H}_e)}}_{\text{BF}_{re}} \times \underbrace{\frac{p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathbf{x}, \mathcal{H}_e)}{p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e)}}_{\text{BF}_{0e}},$$

that is, a conditional Bayes factor for the inequality constraints given the equality constraints and a Bayes factor for the equality constraints. The latter is similar to the Savage-

Dickey ratio that we discussed before, but involves a correction for marginalization.

The probabilities above crucially depend on the marginal probabilities $p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e)$ and $p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathbf{x}, \mathcal{H}_e)$, which are derived from the prior and posterior Dirichlet distributions, respectively. Since the derivations and results are the same for the prior and posterior probabilities, we derive it here for the prior distribution. The prior probability is of the form

$$p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \int_{\mathcal{R}_e \setminus \mathcal{R}_0} \theta_j^{\sum_{k=1}^{j} \alpha_k - j} \prod_{k=j+1}^{K-1} \theta_k^{\alpha_k - 1} \left(1 - j\theta_j - \sum_{k=j+1}^{K-1} \theta_k\right)^{\alpha_K - 1} \mathrm{d}\boldsymbol{\theta}_r,$$

and involves a Dirichlet integral, except that the first j probabilities are now collapsed. Here, we have used $\mathcal{R}_e \setminus \mathcal{R}_0$ to denote the unconstrained parameter space for the parameters $\boldsymbol{\theta}_r = (\theta_j, \ldots, \theta_{K-1})^\mathsf{T}$. We introduce a change of variable $\lambda_j = j\theta_j$, and $\lambda_k = \theta_k$, for $k = j+1, \ldots, K-1$, with $|J| = 1/j$, such that

$$\begin{aligned} p(\boldsymbol{\theta}_0 \in \mathcal{R}_0 \mid \mathcal{H}_e) &= \frac{1}{j\mathrm{B}(\boldsymbol{\alpha})} \int_{\mathcal{R}_e \setminus \mathcal{R}_0} \left(\frac{\lambda_j}{j}\right)^{\sum_{k=1}^{j} \alpha_k - j} \prod_{k=j+1}^{K-1} \theta_k^{\alpha_k - 1} \left(1 - \lambda_j - \sum_{k=j+1}^{K-1} \theta_k\right)^{\alpha_K - 1} \mathrm{d}\boldsymbol{\lambda}_r \\ &= \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \left(\frac{1}{j}\right)^{\sum_{k=1}^{j} \alpha_k - j + 1} \mathrm{B}\left(\sum_{k=1}^{j} \alpha_k - j + 1,\, \alpha_{j+1},\, \ldots,\, \alpha_K\right), \end{aligned}$$

which allows us to express the (marginal) Bayes factor for the equality constraints as

$$\mathrm{BF}_{e0} = \frac{\mathrm{B}(\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha} + \mathbf{x})} \left(\frac{1}{j}\right)^{\sum_{k=1}^{j} x_k} \frac{\mathrm{B}\left(\sum_{k=1}^{j}(\alpha_k + x_k) - j + 1,\, \alpha_{j+1} + x_{j+1},\, \ldots,\, \alpha_K + x_K\right)}{\mathrm{B}\left(\sum_{k=1}^{j} \alpha_k - j + 1,\, \alpha_{j+1},\, \ldots,\, \alpha_K\right)},$$

where the latter factor introduces a correction for marginalizing which originates from the marginalization of the remaining free parameters, including the collapsed category parameter. If it is the case that no free parameters are involved, that is, $\mathcal{H}_0$ assigns expected category proportions to the entire parameter vector $\boldsymbol{\theta}$ (such as in the multinomial test), then the Bayes factor for the equality constraints corresponds to the Savage-Dickey density

ratio.[2] It readily follows that the conditional Bayes factor of inequality constraints given the equality constraints now involves expectations over the conditional Dirichlet distributions

$$p(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathcal{H}_e) = \text{Dirichlet}\left(\sum_{k=1}^{j} \alpha_k - j + 1,\, \alpha_{j+1} \ldots,\, \alpha_K\right)$$

and

$$p(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathbf{x},\, \mathcal{H}_e) = \text{Dirichlet}\left(\sum_{k=1}^{j} (\alpha_k + x_k) - j + 1,\, \alpha_{j+1} + x_{j+1} \ldots,\, \alpha_K + x_K\right),$$

which can be computed, as before, using bridge sampling. To generalize the above derivations for any set of mixed constraints, we note that the conditional Dirichlet distribution adds the parameters for the collapsed categories and corrects for the change in degrees of freedom by subtracting the degrees of freedom it lost; $j - 1$ degrees of freedom are lost if j categories are collapsed. Thus, for mixed hypotheses of the form

$$\mathcal{H}_m : \theta_1 < \theta_2 = \theta_3 < \theta_4 = \theta_5 = \theta_6,$$

we find the following conditional Dirichlet distribution $p(\boldsymbol{\theta}_r \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0, \mathcal{H}_e) = \text{Dirichlet}\left(\alpha_1,\, \alpha_2 + \alpha_3 - 1,\, \alpha_4 + \alpha_5 + \alpha_6 - 2\right)$, which has two sets of collapsed categories, and we lose one degree of freedom for the first, and lose two degrees for the second collapsed category.

The marginal probability has two corrections. First, a uniform probability is stipulated for the collapsed categories, i.e., $1/j$ if j categories are collapsed. Its concentration parameter is equal to the sum of the collapsed categories minus the change in degrees of

[2] When stipulating exact equality constraints on all parameters, it is assumed that there is no prior uncertainty about the model parameters, and the likelihood of the constrained hypothesis marginalized over the parameter space is simply a multinomial distribution. This expression follows from the fact that the prior distribution under $\mathcal{H}_0$ is

$$p(\boldsymbol{\theta} \mid \mathcal{H}_0) = \frac{p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} = \mathbf{c})}{\int_{\mathcal{R}_e} p(\boldsymbol{\theta} \mid \mathcal{H}_e)\, \mathbb{I}(\boldsymbol{\theta} = \mathbf{c})\, \mathrm{d}\boldsymbol{\theta}} = \frac{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)}{p(\boldsymbol{\theta} = \mathbf{c} \mid \mathcal{H}_e)} = 1,$$

for $\boldsymbol{\theta} = \mathbf{c}$ and 0 otherwise.

freedom. Second, a multivariate beta function is introduced that incorporates the corrected concentration parameters. For the mixed hypothesis

$$\mathcal{H}_m : \theta_1 < \theta_2 = \theta_3 < \theta_4 = \theta_5 = \theta_6,$$

we readily find the following marginal probability

$$\frac{\mathrm{B}\left(\alpha_1,\, \alpha_2 + \alpha_3 - 1,\, \alpha_4 + \alpha_5 + \alpha_6 - 2\right)}{\mathrm{B}(\boldsymbol{\alpha})} \left(\frac{1}{2}\right)^{\alpha_2+\alpha_3-1} \left(\frac{1}{3}\right)^{\alpha_4+\alpha_5+\alpha_6-2},$$

and marginal Bayes factor,

$$\mathrm{BF}_{e0} = \frac{\mathrm{B}(\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha}')} \left(\frac{1}{2}\right)^{x_2+x_3} \left(\frac{1}{3}\right)^{x_4+x_5+x_6} \frac{\mathrm{B}\left(\alpha_1',\, \alpha_2' + \alpha_3' - 1,\, \alpha_4' + \alpha_5' + \alpha_6' - 2\right)}{\mathrm{B}\left(\alpha_1,\, \alpha_2 + \alpha_3 - 1,\, \alpha_4 + \alpha_5 + \alpha_6 - 2\right)}$$

where we have used $\alpha_k' = \alpha_k + x_k$. Note that this result has also been established for a specific case, albeit for a more general set of hypotheses, in Mulder, Wagenmakers, and Marsman (in press). What the above analysis of the Bayes factor for the mixed hypotheses $\mathcal{H}_m$ shows is that we are, in general, able to factor the hypotheses and associated likelihoods. This factorization is beneficial since it allows us to compute Bayes factors for parametric constraints with the methods described in the main text, even if these constraints are mixed. Intuitively, parameters that vary freely in both hypotheses do not affect the resulting Bayes factor, since the associated part of the marginal likelihood can be split off from both the mixed and encompassing hypotheses.

Example: Investigation of the Mendelian Laws of Inheritance

In order to demonstrate our method, we reanalyze data to test the Mendelian inheritance theory. The data have already been considered in the context of inequality-constrained hypotheses in Robertson (1978) and recently in Mulder et al. (in press). Mendel crossed a plant variety that produced round yellow peas with a plant variety that produced wrinkled green peas. He then classified whether the peas of the crossbreeds fell into one of four

categories; (1) round and yellow, (2) wrinkled and yellow, (3) round and green, and (4) wrinkled and green. From the classical mendelian laws of inheritance, one can formulate the following inequality-constrained hypothesis about the ordering of the probability of peas falling into one of the four categories:

$$\mathcal{H}_m : \theta_1 < \theta_2 = \theta_3 < \theta_4.$$

A total of $N = 556$ peas were examined, that were distributed across the four categories as follows: $\mathbf{x} = (315, 101, 108, 32)'$. When assigning a uniform Dirichlet distribution as prior for $\boldsymbol{\theta}$, we yield the following result:

$$\begin{aligned}
\text{BF}_{me} &= \text{BF}_{0e} \times \text{BF}_{re} \\
&= \left(\frac{1}{2}\right)^{x_2+x_3} \frac{\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma\left(\alpha_2+\alpha_3\right)} \frac{\Gamma\left(\alpha_2+x_2+\alpha_3+x_3\right)}{\Gamma(\alpha_2+x_2)\Gamma(\alpha_3+x_3)} \times \frac{p\left(\theta_1 < \theta_{23} < \theta_4 \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathbf{x},\, \mathcal{H}_e\right)}{p\left(\theta_1 < \theta_{23} < \theta_4 \mid \boldsymbol{\theta}_0 \in \mathcal{R}_0,\, \mathcal{H}_e\right)} \\
&= 10.301 \times 6.0158 \\
&= 61.970,
\end{aligned}$$

where BF_{0e} was computed using the Savage-Dickey density ratio on the second and third category, and BF_{re} was estimated by collapsing the second and third category, correcting the concentration parameters, and then applying the bridge sampling routine.

Appendix B

Simulation Study: Accuracy of Estimation Methods

To illustrate the accuracy of the estimation methods we conducted a simulation study. In this study we used for four different data sets, given in Table B1, for which the exact Bayes factors could be obtained. To apply the exact procedure we used stick-breaking to express the inequality-constrained hypothesis as independent constraints which were then numerically computed one-by-one. A detailed description the exact procedure can be found in our online appendix. The normalizing constant of the restricted prior distribution was readily available as we assigned a uniform Dirichlet prior on the model parameters. The exact Bayes factors were then compared to the estimated Bayes factors from the bridge sampling method, the conditioning method, and the encompassing prior approach.

Methods

The four data sets and exact results are summarized in Table B1. To quantify accuracy, we estimated the Bayes factors 100 times using the bridge sampling method, the conditioning method, and the encompassing prior approach. For all data sets, we estimated the Bayes factor in favor of the inequality-constrained hypothesis $\mathcal{H}_r$ that the probabilities of each category are increasing against the encompassing hypothesis $\mathcal{H}_e$ that allows all probabilities to freely vary:

$$\mathcal{H}_r : \theta_1 < \theta_2 < \cdots < \theta_K$$
$$\mathcal{H}_e : \theta_1, \theta_2, \cdots, \theta_K.$$

For the bridge sampling method, we drew $20,000$ samples from the constrained posterior distribution. For the conditioning method the marginal probabilities of each constraint holding were estimated using $40,000$ draws from the posterior distribution, resulting in a total of $200,000$ draws for $\mathbf{x_1}$ and $\mathbf{x_2}$, and $240,000$ draws for $\mathbf{x_3}$ and $\mathbf{x_4}$. For the en-

Table B1
Data Sets and Corresponding Exact Bayes Factors in Favor of or Against the Inequality-Constrained Hypotheses that the Parameters Are Increasing.

Observations	$p(\boldsymbol{\theta} \in \mathcal{R}_r \mid \mathbf{x}, \mathcal{H}_e)$	BF_{er}	BF_{re}
$\mathbf{x_1} = (3, 6, 9, 12, 15)'$	0.255149	0.03265839	30.62
$\mathbf{x_2} = (3, 6, 9, 6, 3)'$	0.00196566	4.24	0.23588
$\mathbf{x_3} = (3, 6, 9, 12, 15, 18)'$	0.149099	0.00931515	107.352
$\mathbf{x_4} = (18, 15, 12, 9, 6, 3)'$	2.07023×10^{-9}	$452,373$	2.210565×10^{-6}

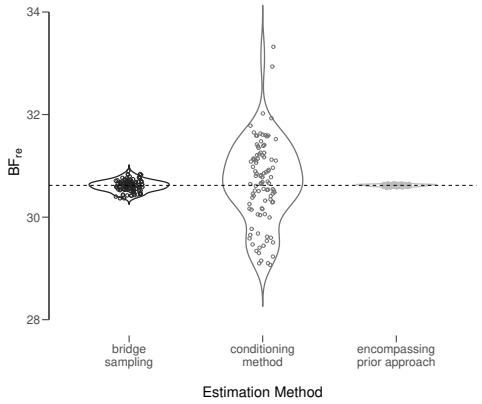
compassing prior approach, we drew 5 million samples from the unconstrained posterior distribution.

Results

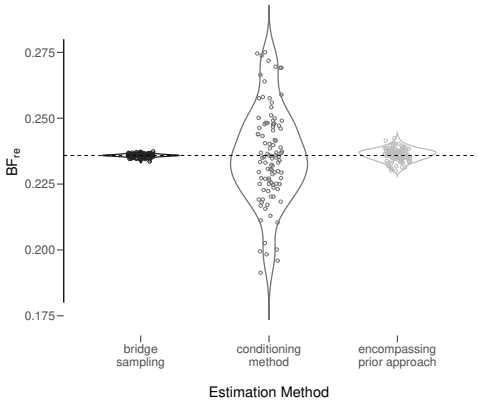
Figure B1 shows violin plots that display the Bayes factors for the three estimation methods for the four data sets (panels a-d). Three results stand out in this simulation: First, for all data sets the variability of the estimates is highest for the conditioning method. Second, the encompassing prior approach is the most accurate for the data sets that show evidence for the restrictive hypothesis (see panels a and c). Third, the advantage of bridge sampling becomes most evident for data sets that show evidence against the restrictive hypothesis (panels b and d). Especially in the 6-category model that shows evidence against the inequality-constrained hypothesis (corresponding to $\mathbf{x_4}$, whose results are displays in Figure B1d), bridge sampling is able to accurately estimate the exact Bayes factor, while the conditioning method is slightly inaccurate and the encompassing prior method fails to estimate any realistic Bayes factor at all: none of the posterior draws were consistent with the restrictive hypothesis, yielding a Bayes factor of 0 for all 100 estimates.

Conclusion

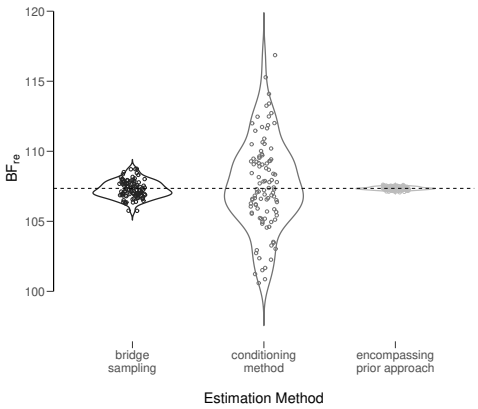
In this simulation study we assessed the accuracy of the bridge sampling method, the conditioning method, and the encompassing prior approach. We computed the exact Bayes factor for four data sets and then estimated the Bayes factor using the three estimation methods.



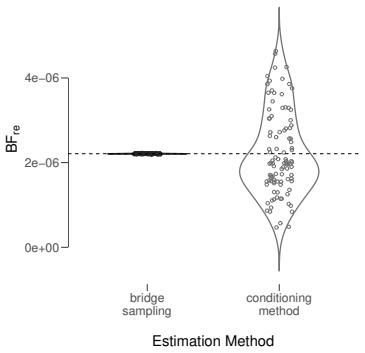
(a) Distribution of Bayes factors for each estimation method for $\mathbf{x_1}$.



(b) Distribution of Bayes factors for each estimation method for $\mathbf{x_2}$.



(c) Distribution of Bayes factors for each estimation method for $\mathbf{x_3}$.



(d) Distribution of Bayes factors for each estimation method for $\mathbf{x_4}$. For this dataset, the encompassing prior was not able to compute any Bayes factor.

Figure B1. Violin plots display the estimated Bayes factors for the bridge sampling method (black), the conditioning method (dark grey), and the encompassing prior approach (light grey). The dashed horizontal line indicates the true Bayes factor. In general, all estimation methods approximate the true Bayes factor, with the conditioning method showing the highest variability. However, for the 6-category model with evidence against the restrictive hypothesis (panel d), the encompassing prior method does not succeed in estimating any Bayes factor, and the conditioning method estimates are less accurate.

The encompassing prior method is most accurate for two of the four data sets. However, the performance of the method depends heavily on the size of the model and quickly deteriorates when the data show evidence against the inequality-constrained hypothesis. In contrast, the conditioning method is more responsive to fluctuations in evidence. To improve the accuracy of the method one could increase the number of samples. However, one should take into account that an increase in the number of samples comes at the expense of runtime, which is already many times higher than the runtime of the other two methods.

Overall, the bridge sampling routine shows the better trade-off between accuracy and efficiency. The variability of the estimates remain in an acceptable range and bridge sampling outperforms the other methods for the extreme data set $\mathbf{x_4}$. The reliability of bridge sampling, which was already on display in the empirical application, was again confirmed in this simulation study.

At this point, we would like to refer the interested reader to our online appendix for a more extended simulation study. This additional simulation study further describes under which conditions the encompassing prior approach and sometimes even the conditioning method fail to estimate a realistic Bayes factor.

Appendix C

R code to sample from a truncated Dirichlet distribution

Damien and Walker (2001) propose to sample from a one-dimensional truncated gamma distribution using data-augmentation and Gibbs sampling. To facilitate the implementation of the Gibbs sampling algorithm, Damien and Walker (2001) introduce a latent variable ν_k whose joint density with γ_k results in convenient conditional distributions. Let γ_k be a gamma variable with bounded support $[\gamma_j, \gamma_q]$. The resulting conditional distribution for $\nu_k \mid \gamma_k$ is simply a uniform distribution on the interval $(0, \exp(-\gamma_k))$. The conditional distribution for $\gamma_k \mid \nu_k$ is proportional to $\gamma_k^{\alpha_k - 1}\, \mathbb{I}(\gamma_k \in [l_k, u_k])$, where $l_k = \gamma_j$ and $u_k = \min(\gamma_q, -\log(\nu_k))$.

Based on these conditional distributions, the Gibbs sampling algorithm can be implemented as follows. First, we initialize the value for γ_k by sampling from an unconstrained gamma distribution. Second, we draw a sample for $\nu_k \mid \gamma_k$ from a Uniform$(0, \exp(-\gamma_k))$ distribution. Then, we draw $\gamma_k \mid \nu_k$ using the inverse CDF technique. The inverse CDF technique inverts the cumulative distribution function (CDF) of a distribution to map a probability to the corresponding value of the random variable. Therefore, we first draw a sample from a standard uniform distribution on the interval $(0, 1)$ which serves as the value for which the cumulative distribution function (CDF) of γ_k, that is $P(\gamma_k \leq x_k)$, is evaluated at. We can then find the corresponding value γ_k by applying the following equation:

$$\gamma_k = \sqrt[\alpha_k]{P(\gamma_k \leq x_k)(u_k^{\alpha_k} - l_k^{\alpha_k}) + l_k^{\alpha_k}}.$$

Below we show the `R` implementation for this algorithm. Note however, that for our simulation study we used a C++ implementation to increase the efficiency of the algorithm and to make it numerically stable. This C++ code is also available in our OSF folder.

```
##############################################################################
# Alexandra Sarafoglou, Last updated June 2020
# Samples from a truncated Dirichlet distribution
# Code written by Maarten Marsman and Alexandra Sarafoglou
# Reference:
#   Damien, P., & Walker, S. G. (2001). Sampling truncated
#   normal, beta, and gamma densities. Journal of Computational
```

```
#   and Graphical Statistics, 10, 206--215.
#########################################################################

truncatedSampling <- function(A, boundaries, niter = 1e5) {
  # input:  A           :  vector of Dirichlet paramters
  #         boundaries  :  list with indices for upper and lower truncation
  #                        boundaries for each parameter theta_k
  # output: samples     :  from truncated Dirichlet distribution

  # define 5% of samples as burn-in; minimum number of burn-in samples is 10
  nburnin  <- max(c(10, (niter/100) * 5))

  samples <- matrix(ncol=length(A), nrow = (niter + nburnin))

  # starting values of Gibbs Sampler
  K         <- length(A)
  gamma     <- rgamma(K, A, 1)
  iteration <- 0

  for(iter in 1:(niter+nburnin)){

    for(k in 1:K){

      ## 0. CHECK FOR BOUNDS ##
      there_are_no_bounds <- is.null(unlist(boundaries[[k]]))

      if(there_are_no_bounds){

        #   if there are no bounds, sample from unconstrained gamma distribution
        gamma[k] <- rgamma(1, A[k], 1)

      } else {

        #   if there are bounds, sample from truncated gamma distribution

        ## 1. LOWER BOUND ##

        # initialize lower bound
        Lo <- 0
        # check for lower bound
        there_is_a_lower_bound <- !is.null(boundaries[[k]]$lower)
        if(there_is_a_lower_bound){

          smaller_value <- boundaries[[k]]$lower
          Lo            <- max(gamma[smaller_value])

        }

        ## 2. UPPER BOUND ##

        # initialize upper bound
        v  <- runif(1, 0, exp(-gamma[k]))
        Hi <- -log(v)
        # check for upper bound
```

```
      there_is_a_upper_bound <- !is.null(boundaries[[k]]$upper)
      if(there_is_a_upper_bound) {

        larger_value <- boundaries[[k]]$upper
        Hi           <- min(gamma[larger_value], Hi)

      }

      ## 3. SAMPLING ##

      gamma[k] <- (runif(1)*(Hi^A[k] - Lo^A[k]) + Lo^A[k])^{1/A[k]}

    }

  }
  # 4. TRANSFORM GAMMA TO DIRICHLET SAMPLES
  samples[iter,] <- as.numeric(gamma/sum(gamma))

  # show progress
  if (iter %in% (niter/100 * seq(1, 100, by = 10))) {
    iteration <- iteration + 10
    print(paste('sampling completed:', iteration, '%', collapse = '\n'))
  }
 }
 samples <- samples[-(1:nburnin), ]
 return(samples)
}

## Example
#
## Draw 20 samples from a truncated Dirichlet(1,1,1,1) distribution
## with the following order-constraints: theta1 > theta2 > theta3 > theta4
#
# boundaries <- list(
#   list(lower = c(2, 3, 4), upper = NULL      ), # theta1 > (theta2, theta3, theta4)
#   list(lower = c(3, 4)   , upper = c(1)      ), # theta1 > theta2 > (theta3, theta4)
#   list(lower = c(4)      , upper = c(1, 2)   ), # (theta1, theta2) > theta3 > theta4
#   list(lower = NULL      , upper = c(1, 2, 3))  # (theta1, theta2, theta3) > theta4
# )
# A     <- c(1, 1, 1, 1)
# niter <- 20
# truncatedSampling(A, boundaries, niter)
```