Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference

Daniel W. Heck\textsuperscript{a,*,1}, Clintin P. Davis-Stober\textsuperscript{b,2}

\textsuperscript{a} University of Mannheim, Germany
\textsuperscript{b} University of Missouri, United States

HIGHLIGHTS

- Psychological theory often leads to inequality-constrained multinomial models.
- Constraints are defined by inequalities or the convex hull of a set of vertices.
- We develop a Gibbs sampler for Bayesian estimation using either representation.
- We offer improved methods for model testing using the encompassing Bayes factor.
- The R package multinomineq implements the proposed methods.

ARTICLE INFO

Article history:
Received 21 August 2018
Received in revised form 11 February 2019
Available online xxxx

Keywords:
Order constraints
Bayesian model selection
Convex polytope
Gibbs sampling
Bayes factor

ABSTRACT

Many psychological theories can be operationalized as linear inequality constraints on the parameters of multinomial distributions (e.g., discrete choice analysis). These constraints can be described in two equivalent ways: Either as the solution set to a system of linear inequalities or as the convex hull of a set of extremal points (vertices). For both representations, we describe a general Gibbs sampler for drawing posterior samples in order to carry out Bayesian analyses. We also summarize alternative sampling methods for estimating Bayes factors for these model representations using the encompassing Bayes factor method. We introduce the R package multinomineq, which provides an easily-accessible interface to a computationally efficient implementation of these techniques.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

Multinomial random variables form the backbone of discrete and categorical data analysis within psychology and the behavioral sciences. The key to any viable data analysis is the successful translation of an abstract theoretical hypothesis into a concrete, statistical model. As a simple example, consider the hypothesis that overconsumption of drugs (i.e., taking more tablets than prescribed) decreases with the number of daily doses (Paes, Bakker, & Soe-Agnie, 1997). To assess the validity of this prediction, one could test the statistical hypothesis that overconsumption is identical across all dosage regimes. If this hypothesis is rejected, one could carry out subsequent analyses to determine if the rates differ across the dosage conditions in a pairwise fashion. Yet, testing the “straw-man” model of all dosage conditions resulting in identical rates of overconsumption is not necessarily a faithful translation of the original hypothesis, rather, it is a means to an end, serving only as a pretext to carrying out tests on multiple pairs of dosage conditions.

To make this example more concrete, suppose we have three dosage regimes of the drug (i.e., once, twice, and three times daily) in a between-subjects design (Paes et al., 1997). We model the number of participants showing overconsumption in each condition as a binomial random variable and define the parameters $\theta_1$, $\theta_2$, and $\theta_3$ as the corresponding probabilities that an individual takes more tablets than prescribed. While we could test whether the three $\theta_i$ parameters are equal across all conditions (i.e., $\theta_1 = \theta_2 = \theta_3$), this does not directly follow from our original hypothesis which only specified a monotonic relationship.

\url{https://doi.org/10.1016/j.jmp.2019.03.004}
0022-2496/© 2019 Elsevier Inc. All rights reserved.
between overconsumption and dosage regimen. Testing the hypothesis of interest requires specifying an ordering relationship imposed on the overconsumption rates for each of the three dosage conditions:
\[ \theta_1 \geq \theta_2 \geq \theta_3. \]  
(1)

Paired with the binomial likelihood function, these order constraints represent a more faithful statistical analysis of the hypothesis being tested (see also Hoijtink, 2011, for a full discussion). Testing order constraints such as these, and linear inequality constraints more generally, requires a bit more effort than simpler tests of equality, but, as we show, can be carried out efficiently and are more interpretable.

A key difficulty in analyzing inequality-constrained models and theories is that it can quickly become difficult to characterize the resulting restricted parameter space (e.g., Davis-Stober, 2012; Fishburn, 1992). Our drug dosage example is quite simple—indeed, for Eq. (1), there are only two non-redundant pairwise order constraints, namely, \( \theta_1 \geq \theta_2 \) and \( \theta_2 \geq \theta_3 \). When combined with the inequality constraints that the probability of overconsumption must be between zero and one for all conditions (i.e., \( 0 \leq \theta_i \leq 1 \)), this completely characterizes the ordering relationships of interest. However, not all interesting hypotheses are so simple in structure. As we illustrate in Section 5.3, the random preference model of Regenwetter and Davis-Stober (2012) is far more complex with 75,834 non-redundant linear inequalities.

In general, bounded, linearly restricted parameter spaces can be defined in two different, yet equivalent, ways (Bronsted, 2012). First, the restricted parameter space can be defined as the solution space to a system of a finite number of linear inequalities and equalities — similar to our drug dosage example. Alternatively, the same restricted parameter space can be defined as the convex hull of a set of extremal points (vertices). Let \( \theta = (\theta_1, \theta_2, \theta_3) \). For our simple dosage example, the set of all extremal points is the set of all vectors, \( \theta \), where each entry is equal to 0 or 1 and satisfy the above inequalities, which yields the set: \((0, 0, 0), (1, 0, 0), (1, 1, 0), \) and \((1, 1, 1)\). Section 1.1 shows that it is often relatively easy to derive these vertices by enumerating all patterns that are predicted by a psychological theory even though it may be difficult to specify the corresponding system of inequality constraints (Regenwetter & Robinson, 2017).

Irrespective of how inequality constraints are formally specified, their statistical analysis has been a long-standing issue in mathematical psychology (Iverson & Falmagne, 1985) and statistics in general (Barlow, Bartholomew, Bremner, & Brunk, 1972; Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2004). In classical statistics, basic results regarding the asymptotic distribution of the likelihood ratio test are valid when testing equality constraints, but are not when testing inequality constraints (Davis-Stober, 2009; Silvapulle & Sen, 2004). As a remedy, methods for inequality-constrained models have recently been developed in the Bayesian framework (Hoijtink, Klugkist, & Boelen, 2008; Karabatsos, 2005; Klugkist, Laudy, & Hoijtink, 2005; Myung, Karabatsos, & Iverson, 2005; Sedransk, Monahan, & Chiu, 1985) or based on minimum description length (Heck, Wagenmakers, & Morey, 2015; Klauer & Kellen, 2015; Rissanen, 1978). Multinomial models with inequality constraints have also been applied to the Bayesian analysis of contingency tables (e.g., Agresti & Hitchcock, 2005; Klugkist, Laudy, & Hoijtink, 2010; Laudy & Hoijtink, 2007; Lindley, 1964). However, general-purpose software packages for Bayesian statistics such as JAGS (Plummer, 2003) or Stan (Stan Development Team, 2018) are often not suited for the analysis of models with complex inequality constraints. This is due to the fact that the boundary of the constrained parameter space is specified as a, typically complex, function of multiple parameters. As a result, the parameters are highly inter-dependent and often cannot be defined independently (for a counterexample with simple constraints, see Heck & Wagenmakers, 2016).

This article considers computational methods of carrying out Bayesian analyses on multinomial models with linear inequality constraints on the parameters. However, we go further than analyzing simple “toy” models such as the dosage example above and consider models defined by arbitrarily complex linear constraints on multinomial parameters. Analyzing this class of model is known to be computationally challenging, especially for highly complex linear constraints as those defined by random preference models (Smeulders, Davis-Stober, Regenwetter, & Spieksma, 2018) and the axioms of additive conjoint measurement (Karabatsos, 2018). In the following, Section 1.1 highlights the relevance of inequality-constrained multinomial models for testing psychological theories. In Section 2, we introduce the notation, likelihood, and prior for multinomial models and the two types of representations for inequality constraints. Section 3 extends existing computational methods for binomial models with specific order constraints (e.g., Karabatsos, 2005; Myung et al., 2005) to multinomial models with arbitrary sets of linear inequalities. More precisely, we develop a general Gibbs sampler for parameter estimation and offer improved computational methods for estimating the encompassing Bayes factor for carrying out Bayesian model selection. Section 4 develops these methods for models that are specified by a set of predicted patterns using the vertex representation. This is useful, as defining a restricted model may be straightforward for one type of representation but not the other, while switching between representations can be computationally infeasible (Avis, Bremner, & Seidel, 1997). In Section 5, we offer the R package multinomineq (Heck & Davis-Stober, 2019) and show how to apply inequality-constrained multinomial models in practice using concrete examples. Finally, Section 6 discusses the analysis of inequality constraints that arise naturally from theoretic hypotheses that are more complex than simple order restrictions (Hilbig & Moshagen, 2014).

1.1. Where do inequality constraints come from?

Inequality constraints on multinomial parameters can arise in a number of ways. Similar to our drug consumption example, they can arise “organically” by directly instantiating the hypothesis of interest. For this example, the inequalities are implied by the natural hypothesis that the response categories should be ordered by dosage regimen. In this way, inequality constraints can provide a direct evaluation of the hypothesis of interest, in contrast to other, heuristic methods such as testing the equality of all three dosage condition parameters and then carrying out additional, post hoc analyses to determine directional differences. In later sections, we will consider other examples of linear inequality constraints that arise naturally from theoretic hypotheses that are more complex than simple order restrictions (Hilbig & Moshagen, 2014).

While not immediately obvious, linear inequality constraints can also arise when evaluating theories/models/axioms in which multiple predictions are made. Such theories are quite common, especially in the field of judgment and decision making. For example, consider the well-known transitivity of preference axiom (Regenwetter, Dana, & Davis-Stober, 2011). Depending upon an individual’s tastes, there are many ways for a decision maker to have transitive preferences over a set of choice alternatives. Evaluating multiple predictions of a theory simultaneously within a multinomial framework opens up additional ways to operationalize this theory of interest. As an example, we consider methods of stochastic specification for deterministic theories, although we note that the application of such methods (e.g., mixture methods) extends beyond the decision making domain (Davis-Stober, Morey, Gretton, & Heathcote, 2016).
1.1.1. Stochastic specification

Many psychological theories predict deterministic choice patterns across different contexts (e.g., different types of stimuli, items, conditions, measurement occasions, or pre-existing groups). For instance, a theory might provide a specific response pattern such as “participants prefer Option A over B in each of five choice scenarios” (Bröder & Schiffer, 2003). Often, however, theories predict more than one response pattern. As illustrated in Section 5.2 for the description-experience gap in the domain of risky gambles (Hertwig, Barron, Weber, & Erev, 2004), the hypothesis that participants assign more weight to small probabilities results in multiple predicted patterns. The complete set of predicted patterns can be obtained in different ways (Regenwetter & Robinson, 2017), for instance, by (a) translating a verbal theory into predicted patterns, (b) deriving algebraic implications of axioms or formal theories, and (c) brute force enumeration of all of the predictions made by the deterministic theory, typically under a set of theory-specific assumptions (e.g., theory parameter values).3 Irrespective of how the theoretical predictions are derived, observed choice frequencies are inherently noisy and exhibit a certain amount of variance both within and across persons or contexts. Hence, the question arises of how to define a stochastic model for empirical frequencies based on a set of deterministic predicted patterns (Carbone & Hey, 2000; Heck, Hilbig, & Moshagen, 2017; Regenwetter & Davis-Stober, 2012, 2018).

In multinomial models, each predicted choice pattern can be represented by a vector of probabilities of either one (an option is deterministically chosen) or zero (an option is not chosen; Bröder & Schiffer, 2003). Fig. 1 illustrates this for two independent binomial probabilities \( \theta = (\theta_1, \theta_2) \) of preferring Option A over B in a control and an experimental condition, respectively. The three black points in Fig. 1A show three predicted patterns of a hypothetical theory that are represented by the vectors \( \psi^{(1)} = (0, 1), \psi^{(2)} = (1, 1), \) and \( \psi^{(3)} = (1, 0) \). For instance, the pattern \( \psi^{(3)} = (1, 0) \) represents the prediction that Option A is chosen in the control condition (since \( \theta_1 = 1 \)) whereas Option B is chosen in the experimental condition (since \( \theta_2 = 0 \)).

To derive a stochastic model based on a set of predictions \( \psi^{(i)} \), it is important to consider why a psychological theory makes multiple predictions in the first place (Regenwetter & Robinson, 2017). A theory might assume that one of the predicted patterns consistently describes the “true” data-generating mechanism across all measurement occasions. According to this interpretation, theory-inconsistent responses merely emerge from unsystematic errors in responding (e.g., due to inattention) whereas latent preferences are stable. In our example, this assumption results in a stochastic model with two independent error probabilities for the two conditions. These error probabilities serve as free parameters and are usually constrained to be below a predefined, fixed threshold such as 20%. In Fig. 1B, this independent-error model is illustrated geometrically by square boxes around the three predicted patterns.

Alternatively, a theory might assume that latent preference states randomly fluctuate across measurement occasions (e.g., across time, persons, or situations), whereas the response process is error-free (Regenwetter & Robinson, 2017). This means that at each measurement occasion, one of the predicted patterns describes the “true” data-generating mechanism perfectly. However, since we do not know which latent states generated the responses in which trials, this error specification leads to a finite mixture model over the predicted patterns (Regenwetter et al., 2014). Fig. 1C shows the parameter space of this mixture model for our example. Essentially, the model permits only those probability vectors \( \theta \) that are inside the triangle obtained by connecting the three predicted preference patterns by straight lines (i.e., \( \theta_1 \geq 1 - \theta_2 \)). Geometrically, this area is the convex hull of the finite number of predicted patterns \( \psi^{(i)} \) and defines a convex polygon in two dimensions (cf. Eq. (8)). More generally, for \( D = 3 \) choice probabilities, the convex hull results in a convex polyhedron, and for arbitrary number of probabilities \( D \), this geometric object is known as a convex polytope (Koppen, 1995; Suck, 1992).

The present paper is concerned with mixture models as that illustrated in Fig. 1C. Theoretically, these models assume random variation in the latent, data-generating process, which can be represented statistically as a mixture distribution over the finite set of predicted patterns \( \psi^{(i)} \) (Regenwetter & Robinson, 2017). The parameter space of these models can equivalently be described by specifying explicit linear inequality constraints on choice probabilities (e.g., \( \theta_1 \leq \theta_2 \)), or by the convex hull of all response patterns \( \psi^{(i)} \) that are predicted by a theory. These mixture models are quite general and, depending upon the experimental design, can provide a strong test of the theory/axiom of interest. For example, applied to a single individual with choice responses aggregated over multiple time points, a violation of a mixture model over a set of predictions provides evidence that this individual must have violated the theory of interest; as the model allowed for an arbitrary distribution over all possible theory-consistent preferences.

---

3 For the brute force enumeration approach, this is typically handled by simulating the deterministic theory while systematically varying its parameter values. To ensure accurate enumeration, a large number of simulation replications should be used to ensure that all possible predictions are simulated.
2. Multinomial models with linear inequality constraints

In this section, we outline the notation, likelihood function, and prior distribution of multinomial models and introduce the two equivalent formal representations of linear inequality constraints.

2.1. Notation and likelihood

In the following we use the term “item type” to refer to a category system $i$ that is modeled by a multinomial distribution with a fixed total number of observations $n_i$. For instance, an item type might refer to a certain context, an experimental condition, a pre-existing group, or a specific combination of choice alternatives presented to the participants. Each item type $i$ has $j_i$ response options indexed by $j = 1, \ldots, j_i$, and thus the total number of categories is $J = \sum j_i$. This notation is very general, since it includes binary items as a special case (resulting in binomial instead of multinomial distributions), allows us to jointly model items with varying numbers of response options $j_i$ (e.g., binary and ternary items), and even covers paradigms such as ranking tasks. For instance, if item type $i$ asks participants to rank three elements $\{a, b, c\}$, the six possible rankings $(abc, acb, bac, bca, cab, cba)$ define an item type with $j_i = 6$ observable response categories.

Since probabilities must sum to one within each category system, the vector of free parameters $\theta$ is defined by omitting the last probability for each item type,

$$\theta = (\theta_{11}, \ldots, \theta_{1(j_1-1)}, \theta_{21}, \ldots, \theta_{J(j_J-1)}).$$

This parameter vector contains $D = \sum j_i - I$ free parameters for which we will define a likelihood function, specify and test inequality constraints, and derive a Gibbs sampler. Assuming independence and identically distributed responses within and across item types, the likelihood of the unconstrained model is given as the product of $I$ probability mass functions of the multinomial distribution,

$$p(k|\theta) = \prod_{i=1}^{I} \left( \begin{array}{c} n_i \\ k_{i1}, \ldots, k_{ij_i} \end{array} \right) \prod_{j=1}^{j_i} \theta_{ij}^{k_{ij}},$$

where the fixed parameters are defined as $\theta_{ij} = 1 - \sum_{l=j+1}^{j_i} \theta_{il}$ for notational convenience. Note that, in a Bayesian context, it is sufficient to assume only exchangeability instead of independent and identically distributed responses (Karabatsos, 2005). Since the probabilities must sum to one within each multinomial distribution, the unconstrained parameter space $\Omega$ for probability vectors $\theta$ is

$$\Omega = \bigotimes_{i=1}^{I} \left\{ \theta \in [0, 1]^{j_i-1} \mid \sum_{j=1}^{j_i-1} \theta_j \leq 1 \right\}.$$  

2.2. Inequality constraints

We are now in a position to connect the likelihood, responsible for generating the data, with the theory being considered. For multinomial models, theories are often operationalized via constraints on the $\theta$ parameters. If a psychological theory implies inequality constraints on the probability vector $\theta$ (e.g., $\theta_{ij} \leq \theta_{il}$), the parameter space of admissible parameters is constrained to a smaller subset $\Omega_c \subset \Omega$. A model with such a truncated parameter space is more parsimonious than the unconstrained model because it permits a smaller set of probability vectors to account for the data (Myung & Pitt, 1997). In the present paper, we are only concerned with inequality constraints that result in a nested parameter space $\Omega_c$ with the same dimensionality as $\Omega$ (cf. Section 6.2).

Psychological theories often imply a set of linear inequality constraints, for instance, that a set of binary choice probabilities $\theta_{ij}$ is ordered and increases monotonically across conditions, $\theta_{11} \leq \theta_{21} \leq \cdots \leq \theta_{J1}$ (Heck et al., 2017). To summarize a set of $R$ linear inequalities that need to hold jointly, it is convenient to describe the constrained parameter space $\Omega_c$ by a matrix $A \in \mathbb{R}^{R \times D}$ and a vector $b \in \mathbb{R}^R$ as follows:

$$\Omega_c = \left\{ \theta \in \Omega \mid A \theta \leq b \right\},$$

where the vector inequality $A \theta \leq b$ is defined by element-wise inequalities of the components, that is, $A_{r, \theta} \leq b_r$ must hold for all rows $A_r$.

As an example, consider the set of monotonic order constraints for binary choice probabilities $\theta_{11} \leq \theta_{21} \leq \theta_{31} \leq . . . \leq \theta_{J1}$ (Hilbig & Moshagen, 2014), which is described by the $Ab$-representation via

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 0 \\ 0 \\ .50 \end{pmatrix}.$$  

When checking the first row of $A$, we see that $1 \cdot \theta_{11} + (-1) \cdot \theta_{21} + 0 \cdot \theta_{31} \leq 0$ is equivalent to the first order constraint $\theta_{11} \leq \theta_{21}$. The second row of $A$ encodes $\theta_{21} \leq \theta_{31}$ and the last row encodes the final order. Fig. 2 shows the set of choice probabilities that satisfy these three inequalities. In this 3-dimensional parameter space for the parameter vector $\theta = (\theta_{11}, \theta_{21}, \theta_{31})$, the inequality constraints are represented by 2-dimensional planes (i.e., by the sides or facets of the polytope). If the number of inequalities cannot be reduced any further by omitting redundant constraints, the set of linear inequalities is said to be facet-defining (Davis-Stober, 2009).

As an alternative to specifying inequalities on the choice probabilities as in Eq. (5), the restricted parameter space can equivalently be represented by enumerating the predicted patterns of a theory (cf. Section 1.1). Geometrically, this representation lists all vertices $v^\theta$ of the polytope instead of describing its facets. For instance, the model defined by $R = 3$ inequalities in Eq. (6)
can alternatively be specified by \( S = 4 \) vertices. Graphically, these vertices are the coordinates of the four corners of the 3-dimensional polytope shown by black points in Fig. 2. Technically, this means that the truncated parameter space \( \Omega_c \) can be defined by the convex hull of the vertices \( \mathbf{v}^{(i)} \), which contains all possible convex linear combinations of the four vertices (Fukuda, 2004):

\[
\begin{align*}
\theta &= \alpha_1 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0.50 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 0 \\ 0.50 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 0.50 \\ 0 \\ 0 \end{pmatrix} + \alpha_4 \begin{pmatrix} 0.50 \\ 0.50 \\ 0 \\ 0 \end{pmatrix}.
\end{align*}
\]

(7)

where \( \alpha_1, \ldots, \alpha_4 \) are (arbitrary) nonnegative weights that sum to one. Substantively, the mixture weights \( \alpha \) are simply the probabilities that a specific pattern \( \mathbf{v}^{(i)} \) generates the observed responses.\(^{4}\) The restricted parameter space \( \Omega_c \) is thus defined as the convex hull of the predicted patterns \( \mathbf{v}^{(i)} \) (Regenwetter et al., 2014):

\[
\Omega_c = \left\{ \theta = \sum_{s=1}^{S} \alpha_s \mathbf{v}^{(s)} \left| \alpha_s \geq 0 \text{ for all } s = 1, \ldots, S \text{ and } \sum_{s=1}^{S} \alpha_s = 1 \right. \right\}.
\]

(8)

For convenience, we list the vertices in an \( S \times D \) matrix \( \mathbf{V} \), in which each row refers to one vertex \( \mathbf{v}^{(i)} \) for the free parameters \( \theta \):

\[
\mathbf{V} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0.50 & 0 \\ 0 & 0.50 & 0 & 0 \\ 0.50 & 0.50 & 0 & 0 \end{pmatrix}.
\]

(9)

Any convex, inequality-constrained model can be described either by the Ab-representation using the facet-defining inequalities or by the V-representation using a set of vertices to describe the convex hull (Davis-Stober, 2009; Doignon & R榭h清洁, 2016; Fukuda, 2004). Depending on the application, however, one of the two representations is often more convenient for theoretical or technical reasons. For instance, the \( S \) vertices are often easier to derive because they are identical to the predicted patterns \( \mathbf{v}^{(i)} \) that are implied by a psychological theory (cf. Section 1.1). For relatively simple cases (e.g., up to several thousand inequalities), the software PORTA (Christof, L篹, & Stoer, 1997) or polymake (Assaf et al., 2017) provide algorithms for the conversion of the two representations. However, this problem is NP-hard for general convex polytopes, and thus the representation conversion is often infeasible in practice if the dimension \( D \) of the polytope or the number of inequalities or vertices is very large. For an overview of algorithms for the vertex and facet enumeration problems, see Avis et al. (1997).

2.3. Prior and posterior distributions

In the Bayesian framework, a prior distribution is required for the probability parameters \( \theta \). We assume independent Dirichlet distributions for the choice probabilities of the multinomial model, with a truncated support that is defined by the convex, inequality-constrained parameter space \( \Omega_c \) (Karabatsos, 2005). Given the shape parameters \( \beta = (\beta_{11}, \ldots, \beta_{18}) \), the probability density function of the truncated Dirichlet distribution is defined as

\[
p(\theta|k) = \frac{1}{c} \mathbb{1}_{\Omega_c}(\theta) \prod_{i=1}^{j} \prod_{j=1}^{l} \beta_{ij}^{\theta_{ij} - 1}.
\]

(10)

where \( \mathbb{1}_{\Omega_c}(\theta) \) is the indicator function, which equals one if \( \theta \in \Omega_c \) and zero otherwise, and \( c \) is the normalization constant, which ensures that the density integrates to one. The special case \( \beta_{ij} = 1 \) for all shape parameters results in a uniform probability distribution, meaning that all of the admissible parameters in \( \Omega_c \) are equally likely a priori. In Section 6.2, we discuss the substantive motivation for this prior in more detail.

The normalizing constant \( c \) of the prior distribution is obtained by integrating the kernel\(^{5}\) of the prior probability density in Eq. (10) over the restricted parameter space \( \Omega_c \):

\[
c = \int \mathbb{1}_{\Omega_c}(\theta) \prod_{i=1}^{j} \prod_{j=1}^{l} \beta_{ij}^{\theta_{ij} - 1} d\theta.
\]

(11)

Since \( c \) necessarily decreases as the relative volume of \( \Omega_c \) becomes smaller, this constant quantifies the restrictiveness or parsimony of the inequality-constrained model. For some types of order constraints and prior parameters, the constant \( c \) can be derived analytically. This is often the case when assuming a uniform prior distribution on the choice probabilities (i.e., \( \beta_{ij} = 1 \) because then, the integral in Eq. (11) simply equals the volume of the restricted parameter space \( \Omega_c \). For example, consider the inequality constraints \( \theta_{11} \leq \theta_{13} \leq \theta_{21} \leq 0.50 \) illustrated in Fig. 2. In this case, the linear order on the three parameters divides the volume of the 3-dimensional cube with side length 0.50 (and thus, a volume of 0.50\(^3\)) into 3\(^3\) equally-sized parts, which results in

\[
c = 0.50^3 \cdot \frac{1}{3!}.
\]

(12)

However, even when assuming a uniform prior, an analytical solution for \( c \) is often difficult to find because a large number of facets or vertices often results in very complex polytopes for the parameter space \( \Omega_c \). As a remedy, the constant \( c \) often needs to be estimated by Monte Carlo integration (cf. Section 3.3).

Since the Dirichlet distribution is a conjugate prior for the multinomial distribution, the posterior distribution is also a Dirichlet (Lindley, 1964).

\[
p(\theta|k) = \frac{1}{f} \mathbb{1}_{\Omega_c}(\theta) \prod_{i=1}^{j} \prod_{j=1}^{l} \beta_{ij}^{\theta_{ij} + \beta_{ij} - 1}.
\]

(13)

The normalizing constant \( f \) of this truncated distribution is computed via integration similarly as for the prior in Eq. (11) while replacing the exponents in the integrand by \( k_{ij} + \beta_{ij} - 1 \). Note that the prior shape parameters \( \beta_{ij} \) additively combine with the observed frequencies \( k_{ij} \) and can thus be interpreted as the prior sample size assigned to each response category.

3. Bayesian inference using the inequality representation

In this section, we summarize and improve computational methods for the Bayesian analysis of multinomial models given a set of linear inequality constraints.

3.1. Gibbs sampling for the Ab-representation

In the Bayesian framework, parameter estimation focuses on the posterior distribution of the parameters given the data. Usually, computational methods are required to obtain point and uncertainty estimates for the parameters. For this purpose, Markov chain Monte Carlo (MCMC) methods draw samples from the posterior distribution, which can then be summarized by descriptive

\(^{4}\) Since the three binomial conditions provide only three degrees of freedom, the four parameters \( \alpha \) are not identifiable. Nevertheless, the model is testable (Bamher & van Santen, 2000).

\(^{5}\) To clarify, we use the term “kernel” to refer to the form of the probability density function where any factors that are not functions of variables are removed.
statistics such as the mean, standard deviation, or highest-density intervals.

The Gibbs sampler is a specific MCMC algorithm that cycles through the components of a parameter vector by drawing samples of the conditional posterior distributions of one parameter given the remaining parameters. Gibbs samplers are especially useful for models with inequality-constraints, because the constraints merely truncate the range of admissible parameter values, whereas the shape or kernel of the posterior density function remains proportional to that of the unconstrained model (Gelfand, Smith, & Lee, 1992). In previous work, Gibbs samplers have been developed for binomial models with specific, theoretically-derived sets of inequality constraints. For instance, Myung et al. (2005) constructed a sampler tailored to decision axioms such as weak or strong stochastic transitivity. Karabatsos and Sheu (2004) derived constraints based on item response theory, and Prince, Brown, and Heathcote (2012) developed a Gibbs sampler for state-trace analysis. We generalize these approaches by developing a Gibbs sampler for multinomial data given any set of convex, linear inequality constraints defined by the inequality representation $A \theta \leq b$. By relying on an analytical solution of the conditional posterior distribution, the proposed Gibbs sampler is more efficient compared to MCMC samplers that require accept–reject steps in each iteration (e.g., Metropolis within Gibbs; Karabatsos, 2001). Our approach also generalizes the hit-and-run-sampler (Smith, 1984), an efficient MCMC method that draws random samples from a uniform distribution on a convex polytope (Lovász & Vempala, 2006).

In each iteration $t$, the Gibbs sampler cycles through the elements of the parameter vector $\theta^{(t)}$ (either sequentially or at random) and updates the current probability parameter $\theta_{ij}$ given the remaining parameters. This updating step requires the conditional posterior distribution of the probability parameter $\theta_{ij}$ for item type $i$ and response option $j$ (Gelfand et al., 1992). To derive this distribution, we first consider the support of the parameter $\theta_{ij}$ conditional on the remaining parameters $\theta_{-ij} = (\theta_{11}, \ldots, \theta_{ij-1}, \theta_{ij+1}, \ldots, \theta_{ij})$. Geometrically, this problem is identical to computing the intersection of a line with a convex polytope as illustrated in Fig. 3. This problem is also known as “line-clipping” for 3-dimensional polyhedra, and thus the following solution builds on ideas of the Cyrus–Beck algorithm often used in computer graphics (Cyrus & Beck, 1978).

Given a specific parameter $\theta^{(t)}$ inside the polytope (black point), the method searches for the lower an upper bounds in the direction of the parameter dimension of $\theta_{ij}$ while fixing the remaining parameters $\theta_{-ij}$ (shown by the vertical black line). By solving the set of inequalities as derived below, we obtain the lower and upper truncation boundaries (red triangles).

To derive the lower and upper truncation boundaries for the parameter $\theta_{ij}$, we denote the $ij$-th column vector of $A$ by $A_{ij}(\theta_{ij})$ and the remaining matrix without this column by $A_{-ij}(\theta_{ij})$. Using vector notation, we split the sum on the left side of the system of inequalities $A \theta \leq b$ into two parts,

$$\theta_{ij} A_{ij}(\theta_{ij}) + A_{-ij}(\theta_{ij}) \theta_{-ij} \leq b.$$  (14)

Next, we subtract all terms involving the remaining parameters $\theta_{-ij}$,

$$\theta_{ij} A_{ij}(\theta_{ij}) \leq b - A_{-ij}(\theta_{ij}) \theta_{-ij}.$$  (15)

Finally, the lower and upper truncation limits $z_0$ and $z_1$ are derived by dividing both sides by the column entries $A_{ij}(\theta_{ij})$, while reversing those inequalities for which the components $A_{ij}(\theta_{ij})$ in the $r$th row and $ij$-th column of $A$ are negative,

$$\begin{align*}
\{ \theta_{ij} &\geq z_0 \} := \max_{\theta_{ij}} \left( \| b_r - A_{ij}(\theta_{ij}) \theta_{-ij} \| / A_{ij}(\theta_{ij}) \right) \text{ for all } r \text{ with } A_{ij}(\theta_{ij}) < 0, \\
\{ \theta_{ij} \leq z_1 \} &:= \min_{\theta_{ij}} \left( \| b_r - A_{ij}(\theta_{ij}) \theta_{-ij} \| / A_{ij}(\theta_{ij}) \right) \text{ for all } r \text{ with } A_{ij}(\theta_{ij}) > 0,
\end{align*}$$  (16)

where $A_r(\cdot)$ is the $r$th row vector of $A$ without the $ij$-column entry. In Eq. (16), we compute the maximum (minimum) of all lower-bound (upper-bound) inequalities since the parameter $\theta_{ij}$ must satisfy all inequalities defined by the vector notation in Eq. (15). Overall, this derivation shows that the conditional posterior of $\theta_{ij}$ has support on the interval $[z_0, z_1]$. This follows via convexity of the constrained parameter space.

Next, we derive the kernel density of the conditional posterior on the support $[z_0, z_1]$. For this purpose, we define the scaling factor

$$s_{ij} = 1 - \sum_{p=1, p \neq j}^{k_{ij}} \theta_{ip} = \theta_{ij} + \theta_{ij}$$  (17)

which equals the sum of the probabilities for the choice frequencies $k_j$ and $k_{ij}$. Using this notation, it follows that the conditional posterior distribution of $\theta_{ij}$ given fixed values of the remaining probabilities $\theta_{-ij}$ is

$$f(\theta_{ij}|(\theta_{-ij}), k) \propto \theta_{ij}^{s_{ij}} \left( \prod_{j=1}^{k_{ij}} \theta_{ij}^{k_{ij}} \right)^{1 - \sum_{j=1}^{k_{ij}} \theta_{ij}} \prod_{z_0, z_1} \theta_{ij}^{s_{ij}}$$  (18)

$$\propto \theta_{ij}^{s_{ij}} \left( s_{ij} - \theta_{ij} \right)^{k_{ij}} \prod_{z_0, z_1} \theta_{ij}$$  (19)

$$\propto \left( \theta_{ij}/s_{ij} \right)^{s_{ij}} \left( 1 - \theta_{ij}/s_{ij} \right)^{k_{ij}} \prod_{z_0, z_1} \theta_{ij}/s_{ij}.$$  (20)

Importantly, with respect to the scaled parameter $\theta_{ij}/s_{ij}$, Eq. (20) is proportional to the density function of a truncated beta distribution:

$$\left( \theta_{ij}/s_{ij} | (\theta_{-ij}, k) \right) \sim \text{Beta}(k_j + \beta_{ij}, k_{ij} + \beta_{ij}) \text{ truncated to } [z_0/s_{ij}, z_1/s_{ij}].$$  (21)
This analytical solution allows us to sample from the conditional posterior distribution of \( \theta_k \) efficiently. First, one draws a (not yet scaled) random sample \( \eta_{ij}^{(t)} \) from a truncated beta distribution using the inverse transformation method (Devroye, 1986, p. 38),

\[
\eta_{ij}^{(t)} = F^{-1} \left[ F(z_0) + (F(z_1) - F(z_0)) w^{(t)} \right],
\]

where \( w^{(t)} \) is sampled uniformly on the interval [0, 1] and \( F \) is the cumulative distribution function of the non-truncated beta distribution with shape parameters defined in Eq. (21). In the next step, the sampled value \( \eta_{ij}^{(t)} \) is scaled by setting \( \theta_{ij}^{(t)} = \eta_{ij}^{(t)} \), which produces a sample from the conditional target distribution in Eq. (20).

By using this analytical solution for the conditional posterior distribution, we can then update the current values of the parameter vector \( \theta^{(t)} \) either in fixed or random order (systematic vs. random scan; Robert & Casella, 2004, p. 375) within each of the \( T \) iterations of the Gibbs sampler. Since the Gibbs sampler requires a point inside the restricted parameter space \( \Omega_c \), as a starting value, we use convex-constrained optimization to find the maximum a posteriori (MAP) estimate of \( \theta \) (Lange, 2010). If a uniform prior is assumed (i.e., all \( \beta_{0i} = 1 \)), the posterior distribution is guaranteed to be uni-modal. Thus, using the MAP estimate as a starting value improves convergence of the Gibbs sampler and reduces the number of required burn-in samples.

### 3.2. Posterior-predictive p-values

In statistical modeling, it is often important to test the absolute fit of a model to data. In the Bayesian framework, posterior-predictive p-values provide a measure of fit that is both intuitive and easy to compute (Meng, 1994). To quantify the discrepancy between observed and expected frequencies, it is convenient to rely on Pearson’s \( X^2 \)-statistic for multinomial models,

\[
X^2 = \sum_{j=1}^{b} \sum_{i=1}^{k} \frac{(k_{ij} - \theta_{ij} n_{i})^2}{\theta_{ij} n_{i}}.
\]

Essentially, posterior-predictive p-values compare the distribution of the \( X^2 \)-statistic for the observed frequencies \( k_{ij} \) against that for the posterior-predictive frequencies \( k_{ij}^{(t)} \).

From a practical perspective, posterior-predictive p-values are computed by iterating through all \( T \) posterior samples. In each iteration, we apply Eq. (23) to compute (a) the statistic \( X_{qbs}^2(\theta^{(t)}) \) using the observed frequencies \( k_{ij} \) and (b) the statistic \( X_{\text{pred}}^2(\theta^{(t)}) \) using the posterior-predictive frequencies \( k_{ij}^{(t)} \), which are randomly sampled from the product-multinomial distribution in Eq. (3) based on the posterior samples \( \theta^{(t)} \). In both cases, the expected frequencies are \( \theta^{(t)}_{ij} n_{i} \). Finally, the posterior-predictive p-value is computed as the proportion of \( X^2 \)-samples for which the observed test statistic is smaller than the posterior-predictive one,

\[
\hat{p}_B = 1 - \frac{1}{T} \sum_{t=1}^{T} I_{X_{\text{obs}}^2(\theta^{(t)}) > X_{\text{pred}}^2(\theta^{(t)})},
\]

This Bayesian p-value will be very small if the observed frequencies do not match the distribution of frequencies implied by the posterior distribution of the model. However, even though small \( \hat{p}_B \) indicates misfit, their interpretation differs from that of p-values in classical statistics because posterior-predictive p-values are not uniformly distributed under the null hypothesis (Meng, 1994).

Note that different test statistics can be used to obtain posterior-predictive p-values tailored to specific research questions. For instance, one can compute separate \( X^2 \)-statistics for different experimental conditions or stimuli to test whether model fit is moderated by these factors. Another advantage of posterior-predictive p-values is that their computation requires only posterior samples, which can efficiently be obtained using the Gibbs sampler. In contrast, computation of the Bayes factor may often be computationally more costly especially for complex inequality constraints.

### 3.3. The encompassing Bayes factor for the Ab-representation

To test whether the predictions of a theory are valid, we compare the model \( M_t \) with the unconstrained parameter space \( \Omega \) against the model \( M_c \), with the inequality-constrained parameter space \( \Omega_c \). Since the constrained model is a nested model, it can never fit the data better than the unconstrained model. However, due to the restricted parameter space, the constrained model \( M_c \) is more parsimonious, a property that is desirable from a theoretical perspective because it results in improved prediction accuracy for new data (Myung & Pitt, 1997). To compare different models, Bayesian model selection provides a trade-off between model fit and complexity. More specifically, the Bayes factor \( B_{u} \) quantifies the evidence for the inequality-constrained model \( M_c \) versus the unconstrained model \( M_u \) and is defined as the ratio of the marginal likelihoods of the two models:

\[
B_{u} := \frac{p(k|\theta, M_c)}{p(k|\theta, M_u)} = \frac{1}{\int_{\Omega_c} p(\theta|\theta, M_u) d\theta} \frac{\int_{\Omega_c} p(\theta|\theta, M_c) d\theta}{\int_{\Omega_u} p(\theta|\theta, M_u) d\theta}.
\]

Usually, it is difficult to compute the integrals in Eq. (25) directly. As a remedy, we rely on the method of encompassing Bayes factors to compute the Bayes factor (Klugkist & Hoijtink, 2007; Klugkist, Kato, & Hoijtink, 2005). Within this framework, the prior of the nested model \( M_c \) must be proportional to the prior of the encompassing model \( M_u \) within the constrained parameter space \( \Omega_c \). Technically, the priors thus have the same kernel density and only differ by the support and the normalizing constant. The truncated Dirichlet prior in Eq. (10) satisfies this requirement, because the nested and the unconstrained model differ only in the parameter spaces \( \Omega_c \) versus \( \Omega \), respectively. To facilitate the computation of the Bayes factor, Klugkist et al. (2005) used the well-known representation of the marginal probability \( p(k|\theta) = \int p(k|\theta^*|k) p(\theta^*|k) d\theta^* \) for an arbitrary parameter value \( \theta^* \in \Omega_c \). When applied to Eq. (25), we obtain the following identity:

\[
B_{u} := \frac{\int_{\Omega_c} p(k|\theta, M_c) p(\theta|\theta, M_u) d\theta}{\int_{\Omega_u} p(k|\theta, M_u) d\theta} = \frac{c}{c'}.
\]

This derivation uses the fact that for \( \theta^* \in \Omega_c \), the likelihood function \( p(k|\theta^*, M_u) \) is identical for the two models, and the prior function \( p(\theta^*|M_c) \) and the posterior \( p(\theta^*|k, M_u) \) are proportional up to the constants \( c \) and \( f \), respectively. Hence, it follows that the Bayes factor equals the ratio of the normalizing constants of the prior and posterior distribution for the constrained model \( M_c \) (see Eqs. (10) and (13)).

To approximate the constants \( c \) and \( f \) in Eq. (26), we can use Monte Carlo sampling to estimate the proportion of prior and posterior samples from the encompassing model that satisfies the constraints. More precisely, computing the encompassing Bayes factor requires the following steps (for a detailed introduction and proofs, see Hoijtink, 2011; Hoijtink et al., 2008). First, we draw \( T \) random samples from the prior distribution of the
unconstrained model $\mathcal{M}_u$, which can easily be done for the Dirichlet distribution. Second, we count the number of samples $\Theta^{(s)}$ that are within the parameter space of the constrained model $\mathcal{M}_c$. Given the $Ab$-representation of an inequality-constrained multinomial model, we only have to check\(^\text{6}\) whether $A\Theta^{(s)} \leq b$. The observed proportion of prior samples in $\Omega_c$ is denoted by $\hat{c}$ and provides an estimate for the constant $c$. Next, we draw random samples from the posterior distribution of the unconstrained model $\mathcal{M}_u$, which is also straightforward because the posterior is a conjugate Dirichlet distribution. Similar as for the prior samples, we compute the proportion of posterior samples $\hat{f}$ that are within the constrained parameter space $\Omega_c$ as an approximation for the normalizing constant $f$.

Importantly, the encompassing Bayes factor requires prior and posterior samples from the encompassing model only. Thereby, the approach is closely related to the popular Savage–Dickey density ratio for computing the Bayes factor in favor of an equality constraint (Heck, 2019; Wetzels, Grasman, & Wagenmakers, 2010). Recently, the method of encompassing priors has been implemented for binomial models in the software QTest (Regenwetter & Cavagnaro, 2019; Regenwetter et al., 2014).

3.4. Precision of the encompassing Bayes factor approximation

Despite the advantages of being computationally straightforward, the encompassing Bayes-factor approach only provides an approximation of the Bayes factor $B_{cu} = f/\hat{c}$. In substantive applications, it is important to ensure that the Bayes factor approximation is sufficiently precise to draw any substantive conclusions. In the following, we show how to quantify the uncertainty based on recommendations by Hoijtink (2011). Importantly, this approach quantifies the uncertainty of the approximation due to the specific computational implementation (for a similar approach, see Heck, Overstall, Gronau, & Wagenmakers, 2019) and does not provide any information about the empirical sampling variation of the Bayes factor for new data.

The precision of the Bayes factor approximation in Eq. (26) can be quantified by formalizing the sampling process used to approximate the constants $c$ and $f$ (Hoijtink, 2011). Given that $T$ posterior samples were drawn to approximate $f$, the number of samples $T_f$ that satisfy the order constraints can be understood as a binomial random variable with the unknown rate parameter $f$. To estimate $f$, it is sufficient to compute the maximum-likelihood estimate $\hat{f} = T_f/T$ as discussed above. However, by treating the number of samples $T_f$ as a random variable, we can derive a posterior distribution for the unknown constant $f$. By assuming a uniform prior, the posterior distribution of $f$ is the beta distribution

$$f|T_f \sim \text{Beta}(T_f + 1, T - T_f + 1).$$

(27)

This distribution describes the uncertainty associated with the approximation of the constant $f$ given that $T_f$ out of $T$ posterior samples from the encompassing model satisfied the constraints (a similar approach applies to the normalizing constant $c$, which is approximated by the proportion $\hat{c}$ of prior samples satisfying the constraints).

To quantify the uncertainty in the Bayes-factor approximation $B_{cu} = f/\hat{c}$, we use a sampling approach (Hoijtink, 2011). That is, we draw $r = 1, \ldots, R$ samples for the unknown parameters $f$ and $c$ from the beta posterior distributions in Eq. (27) and compute the Bayes factor as the ratio of these posterior samples:

$$B_{cu}^{(r)} = f^{(r)}/\hat{c}^{(r)}.$$  

(28)

Thereby, we obtain a distribution of Bayes factors, which quantifies the uncertainty of the numerical approximation (Heck et al., 2019). Specifically, we can compute the standard deviation of the samples $B_{cu}^{(r)}$ to quantify the precision of $B_{cu} = \hat{f}/\hat{c}$. When testing the same model for multiple participants or data sets, computational time can be saved by approximating the prior constant $c$ once with very high precision (large $T$), whereas the constant $f$ needs to be approximated separately for different vectors of observed frequencies $k$. Moreover, if the model assumes a uniform prior and if the volume of the parameter space $\Omega_c$ is available in closed form (e.g., Heck et al., 2017; Prince et al., 2012), the exact numerical value for $c$ can be used in the sampling approximation in Eq. (28).

3.5. A stepwise algorithm

The encompassing Bayes factor approach has the drawback that very large numbers of samples are required if the constants $c$ and $f$ are very small (Hoijtink, 2011). This will be the case if the polytope defined by the parameter space $\Omega_c$ has a very small volume relative to the encompassing parameter space $\Omega$. Especially for $f$, this issue also arises if the posterior distribution of the encompassing model assigns only very small probability mass to the constrained parameter space $\Omega_c$. In both cases, only a very small proportion of the $T$ samples will be inside $\Omega_c$, and thus, the estimates $\hat{c} = T_c/T$ and $\hat{f} = T_f/T$ will have large sampling error. Unfortunately, this issue becomes even more severe for the Bayes factor approximation, since the total number of samples $T$ cancels out. $B_{cu} = T_c/T$. In the worst case, both $T_c$ and $T_f$ equal zero, in which case the Bayes factor cannot be estimated at all.

As a remedy, Hoijtink (2011) proposed to split large sets of inequality constraints into monotonically increasing subsets (see also Mulder, Hoijtink, & de Leeuw, 2012), an approach that has yet not been applied or implemented for multinomial models yet.

For the specific scenario of the $Ab$-representation, this can easily be achieved by partitioning the rows of the matrix $A$ and the corresponding entries of the vector $b$. For the example in Eq. (6), we can define two nested, inequality-constrained models $\mathcal{M}_1$ and $\mathcal{M}_2$ as follows:

$$\mathcal{M}_1 : \begin{pmatrix} A^{(1)} \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \end{pmatrix} \quad \text{and} \quad b^{(1)} = \begin{pmatrix} 0 \end{pmatrix},$$

(29)

$$\mathcal{M}_2 : \begin{pmatrix} A^{(2)} \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix} \quad \text{and} \quad b^{(2)} = \begin{pmatrix} 0 \end{pmatrix}.$$  

(30)

Here, the model $\mathcal{M}_1 : \mathcal{M}_2$ is obtained by selecting the first one (two) rows of $A$ and the first one (two) entries of $b$. By dropping distinct subsets of the inequality constraints, the parameter space increases monotonically and we obtain an order of nested models, $\mathcal{M}_c \subset \mathcal{M}_2 \subset \mathcal{M}_1 \subset \mathcal{M}_u$. Based on this decomposition into an order of nested models, we can now compute the encompassing Bayes factor using a stepwise algorithm. This approach relies on approximating multiple constants $c_m$ for each pairwise comparison of two nested models. First, we use prior samples from the Dirichlet distribution of the encompassing model to compute the proportion of samples $\hat{c}_1$ that are inside the parameter space of $\mathcal{M}_1$, similar as before. Second, we use the Gibbs sampler from Section 3.1 to draw prior samples from the inequality-constrained model $\mathcal{M}_1$ and count the proportion of samples $\hat{c}_2$ that are in $\mathcal{M}_2$. Third, we again use Gibbs sampling for the model $\mathcal{M}_2$ and count the proportion
of samples \( \hat{c}_t \) in \( \mathcal{M}_c \). Since each of the intermediate constants \( c_m \) represents the relative decrease in volume for two consecutive, nested models,\(^7\) the overall constant \( c \) can be obtained by multiplication (for a detailed proof, see Hoijtink, 2011):

\[
c = c_1 c_2 c_c.
\]

The same strategy applies to the posterior constant \( f \), with the difference that, in each step of the algorithm, samples are drawn from the truncated posterior instead of the prior distribution. Finally, due to transitivity of the Bayes factor, the approximation of the encompassing Bayes factor is:

\[
\hat{B}_{cu} = \hat{B}_{1} \hat{B}_{2} \hat{B}_{2u} \approx \frac{\hat{c}_1 \hat{c}_2 \hat{c}_t}{f_1 f_2 f_t}.
\]

Why is it that the stepwise approach results in a more precise approximation of the Bayes factor? Essentially, this is due to the fact that the parameter spaces of two consecutive, nested models \( \mathcal{M}_m \) and \( \mathcal{M}_{m+1} \) differ much less in volume in comparison to the difference in the parameter spaces of the encompassing model \( \mathcal{M}_l \) and the most strongly constrained model \( \mathcal{M}_c \) (Hoijtink, 2011). Accordingly, each of the constants \( c_m \) will be much larger than the overall constant \( c \), and can thus be approximated with higher precision. This in turn increases the precision of the approximation \( \hat{c} \), and in turn, the precision of the Bayes factor estimate \( \hat{B}_{cu} = \hat{c} f \). To quantify the uncertainty of the stepwise procedure, we can extend the approach from Section 3.4. Similar as before, we use random samples \( c_m^{(r)} \) from beta distributions to approximate the uncertainty of each intermediate constant in isolation. Next, we use these samples to repeatedly compute the overall constant \( c^{(r)} = c_1^{(r)} c_2^{(r)} \cdots c^{(r)}_c \). To summarize the uncertainty of the approximation \( \hat{c} \), we can then compute the standard deviation of the samples \( c^{(r)} \). Moreover, it is possible to use samples \( c^{(r)} \) and \( f^{(r)} \) for both the prior and the posterior constants to judge the uncertainty of the Bayes factor approximation using Eq. (28).

Several details of the stepwise algorithm can be improved to increase the computational efficiency even further. First, when checking whether samples from an inequality-constrained model \( \mathcal{M}_m \) are inside the parameter space of \( \mathcal{M}_{m+1} \), it is sufficient to check only those constraints that are unique to the model \( \mathcal{M}_{m+1} \) (since we sample from \( \mathcal{M}_m \), we do not need to check whether the corresponding inequalities hold). Second, to obtain a starting value for the Gibbs sampler of an inequality-constrained model \( \mathcal{M}_m \), we re-use the last sample from the next larger model \( \mathcal{M}_{m-1} \) that satisfied the additional constraints in \( \mathcal{M}_m \). Thereby, the Gibbs sampler does not require a burn-in phase for each of the consecutive steps. Finally, since independent sampling from the encompassing model is more efficient than Gibbs sampling, we recommend to split the inequalities \( A \theta \leq b \) into larger subsets. For instance, instead of defining consecutive, nested models by dropping one inequality at a time, the first step from the encompassing model \( \mathcal{M}_l \) to the nested model \( \mathcal{M}_1 \) can include multiple inequalities (e.g., dozens) instead of a single inequality. Thereby, more samples can be drawn from the encompassing model to compute \( \hat{c}_1 \) in the first step, whereas fewer samples are sufficient to compute \( \hat{c}_m \) in the remaining Gibbs-sampling steps.

### 3.6. An automatic stepwise algorithm

Despite the increased precision of the stepwise procedure, the question remains how many samples \( T_m \) for each step \( m \) are sufficient. Whereas some steps will require only few samples (e.g., if the prior constant \( c_m \) is close to one), others will require more samples to ensure the same level of precision (e.g., if the posterior constant \( f_m \) is very small because the constraints in one step are violated by the data). Especially for the approximation of the posterior constant \( f \), it is difficult to judge a-priori which of the steps require more samples.

As a remedy, we propose an automatic stepwise procedure. For each step \( m \), this method continuously samples from the model \( \mathcal{M}_m \) until a minimum number of samples \( T_{\text{min}} > 0 \) has been observed within the constrained parameter space of the next smaller model \( \mathcal{M}_{m+1} \). Thereby, more samples are drawn to approximate those constants \( c_m \) close to zero which would otherwise have a larger approximation uncertainty for identical \( T_m \). Moreover, this approach also ensures that each of the intermediate approximations \( c_m \) are strictly positive, which resolves the issue that the Bayes factor cannot be computed if both \( f \) and \( \hat{c} \) are zero.

To ensure that a minimum number of samples is used in each step, we first iterate through all models by drawing \( T_0 \) samples. After the first round, we switch between the different nested models and always update the model with the smallest number of samples \( T_m \) satisfying the corresponding inequality constraints. Moreover, as a starting value for the Gibbs sampling from a model \( \mathcal{M}_m \), we again use an adaptive scheme that selects the most recent parameter vector that satisfies the corresponding constraints. Thereby, we reduce the issue of requiring a burn-in phase in each step.

The uncertainty of the automatic stepwise procedure can again be quantified by drawing samples from beta distributions similar as for the stepwise procedure in the previous section. This is the case because in each iteration \( m \) of the automatic procedure, the sampling process results in a negative-binomial likelihood for the number of samples \( T_m \) that is required to reach the minimum number of “hits” \( T_{\text{min}} \). Since the beta distribution is a conjugate prior for the negative binomial, we can again draw samples from a beta distribution (with shape parameters \( T_{\text{min}} + 1 \) and \( T_m - T_{\text{min}} + 1 \)) to quantify the precision of the proportion \( \hat{c}_m = T_{\text{min}} / T_m \).

### 4. Bayesian inference using the vertex representation

In the following, we develop computational tools for obtaining posterior samples and computing the Bayes factor for inequality-constrained multinomial models that are defined by the \( V \)-representation. Instead of providing a set of inequalities as in the \( Ab \)-representation, the \( V \)-representation uses an \( S \times D \) matrix that contains one vertex \( v^{(s)} \) (e.g., a predicted pattern) per row as illustrated in Eq. (8). For many psychological theories, it is indeed easier to obtain a list of all admissible predicted patterns (Regenwetter & Robinson, 2017). Since transformation between the two types of representations is in general NP-hard and often infeasible (Doignon & Raxesh, 2016; Fukuda, 2004), the following developments facilitate the statistical test of psychological theories in practical applications.

#### 4.1. Gibbs sampling for the \( V \)-representation

In Section 3.1, we developed a Gibbs sampler for inequality-constrained multinomial models by deriving the conditional posterior distribution of a parameter \( \theta_i \) given the remaining parameters \( \theta_{\sim i} \). The same steps are required for the \( V \)-representation. However, the posterior distribution of an inequality-constrained model does not depend on the type of representation that is used to define the restricted parameter space \( \Omega' \). Hence, it follows that both the full posterior distribution and the conditional posterior distributions for the \( V \)-representation...
are identical to those derived for the Ab-representation in Section 3.1. Specifically, the conditional posterior of a parameter \( \theta_j \) is again the scaled, truncated beta distribution in Eq. (20). However, to draw random samples from this distribution, it is necessary to compute the lower and upper truncation boundaries \( z_0 \) and \( z_1 \) conditional on the remaining parameters \( \theta_{ji} \). For the Ab-representation, these boundaries were simply derived by solving the set of inequalities \( A \theta \leq b \). However, for the V-representation, we do not know of such a simple algebraic solution. As a remedy, the following algorithm uses a geometric derivation to compute the conditional truncation boundaries of a parameter \( \theta_i \).

In Gibbs sampling, each step requires the distribution of the parameter \( \theta_i \) conditional on the current state of the remaining parameters \( \theta_{ji} \). Geometrically, this implies that, starting at the point \( \theta^g \), “we walk” through the polytope in the direction of the \( ij \)-th dimension. Since the polytope is convex, a straight line in this direction has two intersections with the convex hull of the vertices in \( V \) (i.e., the lower and upper truncation boundaries \( z_0 \) and \( z_1 \), respectively; Lovász & Simonovits, 1993). It follows that the conditional truncation boundaries of the parameter \( \theta_i \) can be derived by computing these two intersections.

For this purpose, we solve two linear programs, one for the lower and one for the upper truncation boundary. By construction of the Gibbs sampler, the current sample \( \theta^{(t)} \) is known to be inside the polytope. To formalize the intuition of “walking” in the direction of the \( ij \)-th dimension, we define the direction vector \( e^{(ij)} \) as the \( ij \)-th unit vector in \( \mathbb{R}^D \) (with zero entries except for the \( ij \)-th entry, which equals one). The linear program now maximizes the distance \( \lambda_i \) from the starting point \( \theta^{(t)} \) in the \( ij \)-th direction, under the constraint that the solution (i.e., the intersection) can be represented as a convex combination of the vertices \( \psi^{(s)} \):

\[
\begin{align*}
\text{maximize} & \quad \lambda_i (1 + \alpha_i e^{(ij)}) = \sum_{s=1}^{S} \alpha_i \psi^{(s)} \\
\text{subject to} & \quad \sum_{s=1}^{S} \alpha_i = 1, \quad \lambda_i > 0 \text{ and } \alpha_i \geq 0 \text{ for all } s = 1, \ldots, S.
\end{align*}
\]

The upper truncation boundary \( z_1 \) is then given by the \( ij \)-th coordinate of the intersection, \( z_1 = \psi^{(s)}_i + \lambda_i \). The second intersection is computed by a linear program that maximizes the distance \( \lambda_0 \) in the opposite direction by replacing the left-hand side of the first constraint in Eq. (33) by \( \theta^{(t)} - \lambda_0 e^{(ij)} \). Accordingly, the lower truncation boundary is \( z_0 = \psi^{(s)}_i - \lambda_0 \).

In summary, an iteration of the Gibbs sampler requires to solve two linear programs in Eq. (33) and to use the resulting truncation boundaries for drawing a random sample from the scaled, truncated beta distribution in Eq. (21). As a starting value for the sampler, one can use a random, convex combination of the vertices \( \psi^{(s)} \). However, to speed up convergence and reduce the number of burn-in samples (similarly as for the Ab-representation), we use the MAP estimate of the mixture weights \( \hat{\alpha} \) as a starting value. This is a valid strategy, even though the mixture weights \( \alpha \) are in general not identifiable because the corresponding estimate implied for the probability vector, \( \hat{\theta} = \sum_{s=1}^{S} \hat{\alpha}_s \psi^{(s)} \), is still unique (Klauser, Singmann, & Kellen, 2015).

4.2. The encompassing Bayes factor

Similar as for the Ab-representation, Bayes factors for the V-representation can be computed with the encompassing method (cf. Section 3.3). Essentially, this requires us to draw \( T \) independent samples from a product-Dirichlet distribution (i.e., the prior and the posterior) and test for each sample \( \theta^{(t)} \) whether it is inside the convex hull of the vertices in the matrix \( V \). A conceptually straightforward approach is to transform the V-representation to the Ab-representation (e.g., using software such as PORTA; Christof et al., 1997) and then check whether \( A \theta \leq b \) holds. However, in a tutorial on polyhedral computation, Fukuda (2004) refers to this approach as “a method that we do not recommend but many people use. This method computes an inequality [Ab-] representation […]. Once the system \( Ax \leq b \) is computed, it is easy to check whether \( p \) [a specific vector to be tested] satisfies the system or not. In most cases, this method is too expensive, since the convex hull computation is very hard in general and impossible for large data” (meaning scenarios with a large number of vertices in the matrix \( V \)).

As an alternative, Fukuda (2004) recommends to directly test whether a parameter vector \( \theta^{(i)} \) is inside the convex hull of the vertices \( \psi^{(s)} \) without computing the Ab-representation. For this purpose, an algorithm is required to check whether there exists a set of nonnegative weights \( \alpha \) that sum to one and satisfy the constraint \( V \alpha = \theta^{(i)} \). The computational approach is based on the geometric intuition that a vector is inside the polytope if and only if it is redundant for the definition of the convex hull. Based on this idea, Fukuda (2004) shows that it is sufficient to solve the following linear program:

\[
\text{maximize} \quad z^T \theta^{(i)} - z_0 \quad \text{with } z_0 \in \mathbb{R}^2 \quad \text{and } z \in \mathbb{R}^D
\]

subject to

\[
\begin{align*}
& z^T \psi^{(s)} - z_0 \leq 0 \quad \text{for all } s = 1, \ldots, S \\
& z^T \theta^{(i)} - z_0 \leq 1.
\end{align*}
\]

The parameter vector \( \theta^{(i)} \) is non-redundant (i.e., outside the convex hull) if and only if the optimal value of this linear program is strictly positive. In this case, the solution \( (z_0, z) \) of the linear program implies that the vector \( \theta^{(i)} \) has a positive distance to the polytope (i.e., is outside of it) and thus is non-redundant for the convex hull.

Despite the advantage of not having to compute the Ab-representation, this approach has the disadvantage that a linear program has to be solved for each prior or posterior sample \( \theta^{(i)} \). For highly constrained models, this might require millions of samples, which renders the approach computationally costly. Smeulders et al. (2018) addressed this challenge via a novel application of a column-generation algorithm — however, this solution may not be general for arbitrary constraints. Future work may improve the efficiency of the algorithm by considering an extended formulation of the polytope (Davis-Stober, Doignon, Fiorini, Glineur, & Regenwetter, 2018).

5. The R package multinomineq

We implemented the above computational methods for multinomial models with convex, inequality constraints in C++ using the linear-algebra library Armadillo (Sanderson, 2010). This has the advantage that many of the sequential computations can efficiently be performed using precompiled code. To also make the methods available to a broad audience, the functions are embedded in the R package multinomineq, which is freely available on GitHub (https://github.com/danheck/multinomineq; Heck & Davis-Stober, 2019). In the following, we show how to translate substantive theories into both the Ab- and the V-representation in R, how to estimate the inequality-constrained parameters \( \theta \), and how to test the constrained multinomial model by computing the encompassing Bayes factor. R scripts for all analyses are available at the Open Science Framework (https://osf.io/ex9u3/) and the package vignette provides more detailed explanations how to use the functions.

---

8 The package will also be made available on CRAN.

9 https://www.dwheck.de/vignettes/multinomineq_intro.html.
5.1. Introductory example: drug dosage and overconsumption

In the drug dosage example in the introduction, Paes et al. (1997) tested the hypothesis that overconsumption increases when the number of daily doses decreases. Daily drug dosage was manipulated in a between-subjects design on three levels: once-daily ($n_1 = 40$), twice-daily ($n_2 = 36$) and three times daily ($n_3 = 15$). Across these conditions, the frequency of participants taking more tablets than prescribed was $k_1 = 16$, $k_2 = 4$, and $k_3 = 2$, respectively. To test the substantive hypothesis of a monotonic relationship (i.e., $\theta_{11} \geq \theta_{21} \geq \theta_{31}$), one can use the inequality representation by defining a matrix $A$ and a vector $b$ in $R^n$:

$$A \leftarrow \text{matrix}(c(-1, 1, 0, 0, -1, 1), \ nrow = 2, \ byrow = \text{TRUE})$$

$$b \leftarrow c(0, 0)$$

As defined in Eq. (5), the first row of $A$ and the first value of the vector $b$ define the inequality $-1 \cdot \theta_{11} + 1 \cdot \theta_{21} + 0 \cdot \theta_{31} \leq 0$. Since we are working with binomial data, the function $\text{bf_binom}$ is used to compute the Bayes factor in favor of the order-constrained hypothesis:

$$\text{bf_binom}(k = c(16, 4, 2), \ n = c(40, 36, 15), \ A = A, \ b = b, \ M = 100000)$$

The vectors $k$ and $n$ provide the frequency of overconsumption and the number of observations per condition, respectively, whereas the argument $M$ specifies how many prior and posterior samples are drawn from the encompassing model (the computation required 0.1 s on an Intel i7-7700). By default, all functions in multinoimineq assume a uniform prior on the restricted parameter space as specified via the shape parameters $\beta_{ij} = 1$ of the Dirichlet distribution (this can be changed via the argument prior).

The function $\text{bf_binom}$ returns a matrix with the approximation of the Bayes factor in the first column and a summary of the sampling error in the remaining columns (i.e., the standard deviation and the 5%- and 95%-quantiles of the samples $B_{11}^M$ from Section 3.4)

$$\text{bf se ci.5}$$

$$\textbf{bf_0u} \quad 2.11 \quad 0.02 \quad 2.08 \quad 2.14$$

$$\textbf{bf_00} \quad 0.47 \quad 0.00 \quad 0.47 \quad 0.48$$

$$\textbf{bf_001} \quad 2.70 \quad 0.03 \quad 2.65 \quad 2.75$$

Note that $\textbf{bf_0u} = 2.11$ refers to the Bayes factor for the constrained model versus the unconstrained model and $\textbf{bf_00} = 0.47$ to its inverse. In this example, the Bayes factor $B_{0u} = 2.11$ indicates that the data provide only anecdotal evidence in favor of the hypothesis of monotonicity. Moreover, the Bayes factor $\textbf{bf_001} = 2.70$ compares the order-constrained model against an alternative model with a parameter space that is defined as the exact complement (i.e., $\Theta_0 = \Theta \setminus \Theta_0$).

5.2. Testing theories via the V-representation: The description-experience gap

In the second example, we use the vertex representation to test the description-experience (DE) gap, which states that the presentation format of probabilities in risky gambles affects individuals’ preferences (Hertwig et al., 2004). In the description condition, participants are presented with the exact numerical value of the probability of receiving a gain (e.g., “you receive $10 with $p = .20 and $0 otherwise”). In the experience condition, participants are presented with random samples of the two possible monetary outcomes $10 and $0 in sequential order, which occur with probabilities $p = .20 and $p = .80, respectively. The DE gap states that rare events (i.e., small probabilities $p$) are overweighted in the description condition but underweighted in the experience condition.

Hertwig et al. (2004) tested the DE gap by presenting participants with six binary decision problems in each of the two experimental conditions. These stimuli were constructed to ensure that over- and underweighting of small probabilities would result in distinct predicted patterns across the six decision problems. However, under the assumption that preferences are heterogeneous across participants and trials, both over- and underweighting imply multiple choice patterns that are in line with the psychological theory (Regenwetter & Robinson, 2017). Given such heterogeneous predictions, observed choice frequencies can be modelled by a mixture distribution over the predicted patterns, and thus, by a multinomial model with inequality constraints (cf. Section 1.1).

To derive all patterns that are predicted when assuming heterogeneous preferences, Regenwetter and Robinson (2017) repeatedly simulated preference patterns with cumulative prospectivetheory (Kahneman & Tversky, 1979) based on random data-generating parameters. These patterns provide the vertex representation via the matrix $V$ to define the inequality-constrained multinomial model. For the DE gap, each row of the matrix $V$ defines a specific pattern of predicted probabilities of choosing Option H in each of the six decision problems. The first of the 32 rows of the matrix $V$ for the overweighting model are:

$$\begin{bmatrix}
p1 & p2 & p3 & p4 & p5 & p6 \\
[1.] & 0 & 0 & 0 & 0 & 0 \\
[2.] & 0 & 0 & 0 & 0 & 1 \\
[3.] & 0 & 0 & 1 & 0 & 0 \\
[4.] & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Bayesian parameter estimation for the parameters $\theta$ proceeds by drawing posterior samples from the restricted model. The corresponding Gibbs sampler for binomial data and the $V$-representation is called via:

$$\text{mcmc} \leftarrow \text{sampling_binom}($$

$$\text{kk} = \text{c}(9, 16, 16, 7, 12, 16), \ \text{n} = 25, $$

$$V = V, \ M = 5000, \ \text{cpu} = 8)$$

Similar as in the first example, binomial data are defined via a vector $k$ (frequencies of choosing Option H for each decision problem) and a vector $n$ (number of responses). Here, the number of responses $n = 25$ was identical for each lottery, and thus we can provide only a scalar value instead of a vector. The Gibbs sampler for the $V$-representation required 62 s to sample $M = 5000$ iterations on each of cpu = 8 processing units in parallel.

This set of posterior samples is stored in the object $\text{mcmc}$ containing a list of matrices, each with $M = 5000$ posterior samples $\theta^{(t)}$. To assess the efficiency of the Gibbs sampler, we compared the number of MCMC iterations $M_{\text{eff}} = (5000 - 10) \cdot 8$ (note that 10 samples are dropped as burnin) against the effective sample size $M_{\text{eff}}$, which is defined as the number of independent samples that would be required to achieve an equivalent estimation accuracy (Heck et al., 2019). The corresponding ratio $M_{\text{eff}}/M_{\text{est}}$ quantifies the loss in information due to dependent sampling and ranged between 0.80 and 1.00 for the six parameters (with an average of 0.94). This shows that Gibbs sampling was very efficient for this specific example. Next, we can test model fit via posterior-predictive $p$-values (cf. Section 3.2) which are computed as:

$$\text{ppp_binom(prob = mcmc, k = c(9, 16, 16, 7, 12, 16), n = 25)}$$

---

10 We opted for the label se to highlight that SD($\textbf{bf_00}$) is conceptually similar to the standard error of the Bayes factor approximation.
For the data by Hertwig et al. (2004), posterior-predictive p-values indicate that choices in the description condition were better described by over- than by underweighting ($p_B = .564$ vs. $p_u = .005$, respectively) whereas choices in the experience condition were better described by under- than by overweighting ($p_B = .587$ vs. $p_u = .058$, respectively).

Whereas Gibbs sampling is performed conditional on the inequality constraints, the encompassing Bayes factor requires prior and posterior samples from the unconstrained model. Hence, the Bayes factor is implemented in a separate function:

\[
bf <- \text{bf_binom}(k = c(9, 16, 16, 7, 12, 16), n = 25, V = V, M = 5000, cpu = 8)
\]

Using $M = 5000$ samples on each of $cpu = 8$ processing units required $34 s$ to compute the Bayes factor for the $V$-representation with $S = 32$ vertices and $D = 6$ parameters. The data provided evidence for over- and against underweighting in the description condition ($B_{0n} = 3.7$ versus $B_{0o} \approx 1.500$, respectively), while showing evidence for under- and against overweighting in the experience condition ($B_{0n} = 34.8$ versus $B_{0o} \approx 1.100$, respectively; cf. Table 6 in Regenwetter & Robinson, 2017). Even though the substantive conclusions are similar to those of the original analysis in this specific example, there are examples in the literature where the reliance on multinomial models with inequality constraints can make a big difference (for instance, when testing transitivity of preferences; Regenwetter & Davis-Stober, 2012).

The package multinomineq also provides a wrapper function $V_{\rightarrow}Ab$ which transforms the vertex representation (i.e., the matrix $V$) to the inequality representation by calling the R package rPorta (Nunkesser, Straatmann, Wenzel, Christof, & Loebel, 2009). If the transformation succeeds, one can rely on the computationally more efficient methods for the $Ab$-representation. Moreover, the reverse transformation is available via the function $Ab_{\rightarrow}V$.

5.3. Multinomial data and complex inequalities: the strict weak order polytope

In the last example, we test a set of complex inequality constraints using multinomial instead of binomial data. We reanalyzed the data by Regenwetter and Davis-Stober (2012) who tested whether preferences for monetary gambles are transitive. Participants were presented with 10 decision problems each featuring two out of the five lotteries $a, b, c, d$, and $e$ (e.g., $a$ versus $b$; $a$ versus $c$; etc.). A ternary choice format allowed for the opportunity to choose one of the two gambles or to respond “indifferent”. To test whether preferences are transitive under the assumption of heterogeneity, Regenwetter and Davis-Stober (2012) used an inequality-constrained multinomial model. The parameter space of this model was defined by the strict weak order for the five lotteries which defines a polytope with 75,834 facet-defining inequalities for the 20 free parameters (i.e., the probabilities of choosing the first or the second gamble in each of the 10 problems). In contrast to the previous example, these constraints are much more complex, as illustrated by the last three rows of the matrix $A$ given in Box I. For illustration purposes, the R package multinomineq provides the data by Regenwetter and Davis-Stober (2012) in the data frame regenwetter2012 and the inequality constraints of the strict weak order polytope for five alternatives in the list swap5 (which contains the matrix $A$ and the vector $b$).

For multinomial response formats with more than two options per item type (e.g., ternary choice), the observed frequencies need to be provided in a different format than for binary responses. In multinomineq, the vector $k$ provides the observed frequencies of all response options (ordered by item type: $k_1, \ldots, k_{10}$) and the vector options provides the number of choice options per condition or item type (i.e., $J_1, J_2, \ldots, J_l$). In the present example, $l = 10$ different item types with $J_l = 3$ choice options were presented $n_l = 45$ times each. To fit the data of the first participant of Regenwetter and Davis-Stober (2012), posterior samples from the inequality-constrained model are drawn via:

\[
\text{mcmc} <- \text{sampling_multinom}(k = c(21, 24, 0, 243, 0, 45, 0, \ldots), \text{options} = c(3, 3, 3, 3, 3, 3, 3, 3, 3, 3), A = A, b = b, M = 1000, cpu = 8)
\]

The Gibbs sampler for the $Ab$-representation with 75,834 facet-defining inequalities on 20 parameters required approximately two minutes to sample $M = 1000$ iterations on each of $cpu = 8$ processing units in parallel. Fig. 4 shows the posterior samples for two probability parameters, indicating that the Gibbs sampler converged very fast (only 10 samples were discarded as burn-in) and resulted in acceptable autocorrelations of 0.73, 0.36, and 0.19 for lags of 1, 5, and 10, respectively (averaged across parameters). Similarly as in the previous example, the ratio $M_{eff}/M_{res}$ quantifies the loss in information and ranged between 0.04 and 0.31 for the 20 parameters (with an average of 0.14). This shows that the efficiency of the implemented Gibbs sampling is acceptable even for complex inequality constraints.

To compute the encompassing Bayes factor for the $Ab$-representation, the function $bf\_multinom$ can be applied similarly as shown in the previous example. Here, we compute the Bayes factor more efficiently by using separate calls to the function $count\_multinom$ to count the number of samples that are inside the inequality-constrained parameter space $\Omega_2$ for (a) the prior distribution and (b) the posterior distribution. This is advantageous since the proportion $c$ of prior samples satisfying the constraints is required repeatedly for the analysis of all participants. To increase precision, we used a larger number of samples to approximate the normalizing constant $c$ (which required 129 s):
approximate estimation uncertainty (cf. Section 3.4):

\[ \text{count}_\text{to}_\text{bf}(\text{prior}, \text{post}) \]

In their paper, Regenwetter and Davis-Stober (2012) reported whether the observed choice proportions of 30 participants across three gamble sets satisfied the inequality constraints descriptively. Moreover, frequentist \( p \)-values were used to test whether the discrepancy of the observed choice frequencies to the inequality-constrained model was significant (Davis-Stober, 2009). To complement the original analysis, Table 1 provides the corresponding Bayes factors for these data with a uniform prior on the choice probabilities (the complete analysis of the 90 datasets required 54 min). The first two columns of Table 1 were previously reported by Davis-Stober, Brown, and Cavagnaro (2015), who calculated the Bayes factors using the encompassing prior approach with rejection sampling. There are some minor discrepancies between the values reported in our Table 1 and Davis-Stober et al. (2015)'s values, reported in their Tables 4 and 5. These discrepancies were due to an error in the original computer code used by Davis-Stober et al. (2015) that underestimated the volume of the strict weak order polytope. Upon correction, both sets of analyses agree, see Davis-Stober, Brown, and Cavagnaro (2018) for an erratum.

In Table 1, Bayes factors for most participants and gamble sets are clearly larger than one, thus providing evidence that preferences are transitive. Moreover, the Bayes factor \( B_{0u} \) takes into account how well the inequality-constrained model describes the data: If the inequality constraints are only barely met (i.e., if the observed choice proportions are close to the facets of the polytope), the Bayes factor \( B_{0u} \) will be smaller compared to the case that the constraints are clearly satisfied (i.e., if the distance is large between the observed choice proportions and the facets of the polytope). In that latter case, the proportion of posterior samples inside the polytope will be larger compared to the former case. In Table 1, this is reflected by the substantial variance in how close the Bayes factors are to the maximum Bayes factor \( B_{0u}^{\text{max}} = 1/c \approx 2187 \) (which is obtained if all posterior samples satisfy the constraints).

6. Discussion

In mathematical psychology in general and judgment and decision making in particular, many theories can be formulated by a set of linear inequality constraints on multinomial models (Iverson, 2006). This includes representational measurement theory (Karakatsos, 2001; Krantz, Luce, Suppes, & Tversky, 1971), state-trace analysis (Prince et al., 2012), decision axioms such as transitivity (Myung et al., 2005; Regenwetter et al., 2011), random utility models (for a review, see Marley & Regenwetter, 2017), and multiattribute probabilistic inference (Heck et al., 2017). Moreover, inequality constraints are often of interest in the analysis of contingency tables (Klugkist et al., 2010; Lindley, 1964) and in cognitive diagnostic assessment (Hojitink, Béland, & Vermeulen, 2014). Geometrically, linear order constraints on the probabilities \( \theta \) result in a parameter space representing a convex polytope, which can be defined either by a set of inequalities (i.e., the \( \mathbf{A}\theta \leq \mathbf{b} \)) or by the convex hull of a set of predicted patterns \( \mathbf{v}(s) \) for the choice frequencies (i.e., the \( \mathbf{V}\)-representation: \( \theta = \sum \alpha_s \mathbf{v}(s) \) with nonnegative weights \( \alpha_s \)). To facilitate statistical tests of such models using either type
models with large (or unconstrained) parameter spaces are penalized more for complexity than parsimonious models with small parameter spaces. Since the approximation of the Bayes factor is usually computationally expensive (or even infeasible), one may restrict model selection to those models that have an adequate fit in terms of posterior-predictive \( p \)-values (Regenwetter et al., 2018).

### 6.1. Analysis of nested data structures

Many psychological experiments use repeated measures per person, and thus, the natural question arises whether and how inequality-constrained multinomial models can account for nested data structures. As a first strategy, it is possible to fit a single inequality-constrained model to the summed individual frequencies (often referred to as “complete pooling”). Even if there is heterogeneity in the parameter vector \( \theta \) across participants, this is a valid approach because the mean of the individual parameters must satisfy the inequality constraints of the model. For each person on the individual level. This is due to the fact that (between-person) mixtures of (within-person) mixture distributions do not affect the convex, inequality-constrained parameter space (Regenwetter & Davis-Stober, 2018). However, the reverse statement does not hold: it is possible that the inequality constraints are satisfied by the aggregated frequencies even though they are violated by some or all of the participants.

To detect variability across participants, researchers often test inequality-constrained models for each person separately, for instance, using Bayesian model selection on the individual level (e.g., Heck et al., 2017; Regenwetter et al., 2014). This “no pooling”-strategy leads to one Bayes factor per person, which opens the question of how to aggregate the distribution of individual Bayes factors on the population level. If a researcher wants to test whether the restricted model accounts for the data by all participants, one can compute the group Bayes factor defined as the product of the individual Bayes factors (Klaassen, Zedelius, Veling, Aarts, & Hoijtink, 2018). Often, however, researchers are interested in classifying participants as users of different decision strategies which are operationalized by separate models (Heck et al., 2017) and want to know whether a majority of participants are better described by one model than by the competitors.

To test hypotheses about latent classes of participants, Cavagnaro and Davis-Stober (2018) proposed a random-effects model that only requires on one or more distributions of individual Bayes factors (e.g., for a control and treatment condition). The random-effects Bayes factor assumes that participants are clustered into latent classes that differ with respect to the data-generating model. The method then tests whether the group sizes of these latent classes differ within or across conditions (for details, see Cavagnaro & Davis-Stober, 2018). The technical framework and implementation of this approach is closely related to the multinomial models discussed in the present manuscript, with the main difference that the individual Bayes factors and not the raw choice frequencies serve as input for statistical inference. Hence, the methods developed in the present manuscript allow computing the random-effects population Bayes factor to test, for instance, whether a specific treatment results in a larger
proportion of risk-seeking versus risk-avoiding participants (Cavagnaro & Davis-Stober, 2018). Overall, inequality-constrained multinomial models (and their implementation in the R package multinomineq) may thus serve as a general and comprehensive framework for testing theories both on the individual and the population level.

As a third alternative (called “partial pooling”), it is possible to develop hierarchical multinomial models with inequality constraints to account for nested data structures. For example, one could assume that the inequality constraints hold for all participants but that the individual parameter vectors \( \theta \) differ across participants (Haaf & Rouder, 2019). To model this substantive hypothesis, one may define a separate parameter vector \( \theta^p \) for each participant \( p \) and assume a Dirichlet distribution of \( \theta^p \) on the group level. Such a hierarchical approach offers the benefit of shrinkage, meaning that the individual parameter estimates will be informed by each other and pulled towards the group mean (Efron & Morris, 1975). However, future work is required to develop computational methods for such high-dimensional models. There are also unresolved questions regarding the direction of parameter shrinkage and how this could interact with the substantive models being tested.

### 6.2. Limitations and future directions

Several open questions and possible limitations remain concerning computational methods for inequality-constrained multinomial models. First, simulation studies are required to assess the efficiency of the developed methods and to assess the benefit of considering the structure of a given set of constraints. For instance, if the parameter space is highly regular such as that of a linear order polytope (Regenwetter et al., 2011), analytical or computational improvements might be possible to speed up estimation and testing of a model (Davis-Stober et al., 2018). Moreover, the algorithms could be adapted by taking into account which of the inequality constraints are violated for a given dataset. For instance, if 95% of the constraints are satisfied descriptively, it is likely that most of the posterior samples from the encompassing model will also adhere to these constraints (see Smeulders et al. (2018) for applications of similar computational heuristics). To approximate the Bayes factor, it might thus be beneficial to reorder the inequalities in the Ab-representation by their relative strength (i.e., by the rejection probability). Similarly, in the stepwise and automatic procedures, the relative strength of the inequalities could be exploited to establish a more efficient clustering into nested models.

Substantively, future work should investigate the choice of prior distribution for multinomial models with inequality constraints. When relying on Bayesian model selection, priors should reflect the psychological theory underlying a specific model and domain as closely as possible (Lee & Vanpaemel, 2018). The computational methods outlined in the present paper assume independent Dirichlet distributions for the choice probabilities, which includes the uniform distribution as a special case (e.g., Hoijtink et al., 2014). In the absence of further knowledge, a uniform prior is often justified since researchers want to assign equal probability to all parameters that are in line with a substantive theory. Moreover, using simulations, Klugkist et al. (2010) showed that a uniform prior results in a good performance of the encompassing Bayes factor for inequality-constrained multinomial models. The assumption of independence may be more controversial, especially if the same choice alternatives are presented repeatedly across different paired comparisons (Regenwetter & Davis-Stober, 2018). As an alternative to independent Dirichlet priors, McCausland and Marley (2013) derived a family of prior distributions for choice probabilities on non-empty subsets of a finite set of objects that take dependency into account. However, this prior is limited to very specific types of decision-making models and not suited as a default for the general class of inequality-constrained multinomial models presented in the present paper. As a remedy, the independent Dirichlet prior may serve as a useful approximation for testing psychological theories and as a convenient default for inequality-constrained data analysis in general. This view is in line with the existing literature as indicated by the conclusion of Regenwetter and Davis-Stober (2018, p. 6) that “statistical analyses of these models usually depend on auxiliary independence and stationarity assumptions to get simple and tractable test statistics”.

Concerning the type of models that can be analyzed, the methods developed in the present manuscript only apply to convex, linear inequality constraints that result in a parameter space of full dimensionality. For instance, a model for \( D = 3 \) binomial probabilities must have three free parameters, as represented by a 3-dimensional polytope (cf. Fig. 2). However, theories often predict that choice probabilities are identical across different item types, which leads to equality constraints of the form \( h_0 = h_k \) (e.g., Bröder & Schiffer, 2003). Formally, any set of linear equality constraints can be defined via a matrix \( C \) and a vector \( d \) similar to the \( Ab \)-representation of inequalities (i.e., by \( C \theta = d \)), thereby defining in a lower-dimensional parameter space. For instance, in Fig. 5A, the 3-dimensional parameter space \( \Omega \) reduces to a constrained parameter space represented by a 2-dimensional plane. For models posing both equality and inequality constraints, the Gibbs sampler in Section 3.1 needs to be adapted to walk through the lower-dimensional projection of the \( D \)-dimensional space of choice probabilities. Moreover, the encompassing approach of computing Bayes factors will fail for such models because random samples from the \( D \)-dimensional prior or posterior distribution will be outside the lower-dimensional parameter space with probability one. As a remedy, the encompassing Bayes factor method allows to test “about equality constraints” defined as \( |C \theta - d| \leq \delta \) which provide similar results as \( \delta \rightarrow 0 \) (Hoijtink, 2011; Klugkist et al., 2010).

Other generalizations of the presented methods concern different types of inequality-constraints. Models with non-convex parameter spaces may emerge if a theory combines multiple inequality constraints by logical OR statements as shown in the error model in Fig. 1B. Whereas the conjunction of multiple inequalities (e.g., \( \theta_{11} \leq \theta_{21} \text{ AND } \theta_{11} \leq \theta_{31} \text{ etc.} \)) always results in a convex parameter space, a disjunction (e.g., \( \theta_{11} + \theta_{21} + \theta_{31} \leq 1 \text{ OR } \theta_{11} + \theta_{21} + \theta_{31} \leq 2 \)) may result in a non-convex parameter space. The latter case is illustrated in Fig. 5B, showing that the parameter space of a disjunction is a (possibly overlapping) union of convex polytopes. To analyze such models, the encompassing Bayes factor can again be approximated by counting how many prior and posterior samples fall into the non-convex parameter space. However, the Gibbs sampler developed in Section 3.1 requires a convex parameter space to ensure that the support of the conditional posterior distribution is always a connected interval with one lower and one upper truncation boundary, which is not necessarily the case for non-convex parameter spaces. An important type of restrictions concerns nonlinear inequality constraints as those shown in Fig. 5C. Nonlinear inequalities emerge in the analysis of contingency tables (Klugkist et al., 2010) and can be represented neither by a finite number of vertices nor by linear inequalities. Instead, a theoretically derived (and measurable) indicator function \( I_{\Omega_0}(\theta) \) is required to define the

---

12 The uniform prior does not belong to the class of objective priors, which are derived based on asymptotic considerations (Ghosh, 2011).
restricted parameter space \( O_n \subset \Omega \). If \( O_n \) is full-dimensional, the encompassing Bayes factor can be computed by counting the number of prior and posterior samples for which \( I_{O_n}(\theta) = 1 \). It is also possible to adapt the Gibbs sampler in Section 3.1 to sample from the posterior distribution as long as the parameter space is full-dimensional and convex. In this case, only the computation of the lower and upper truncation boundaries of the conditional posterior distribution (i.e., the scaled, truncated beta in Eq. (21)) needs to be generalized for nonlinear indicator functions (e.g., via a bisection algorithm). The function \( \text{sampling\_nonlinear} \) provides a Gibbs sampler for convex, nonlinear inequality constraints defined via a user-defined indicator function \( I_{O_n}(\theta) \).

6.3 Conclusion

To test psychological theories, it is important that statistical models reflect a theory’s core predictions without requiring strong and often arbitrary auxiliary assumptions. Multinomial models with inequality constraints provide an ideal framework for this purpose and allow to test both formal theories assuming deterministic axioms (Iverson, 2006) as well as verbal theories predicting multiple choice patterns (Regenwetter & Robinson, 2017). Given the implementation of Bayesian inference for this model class in the R package multinomineq (Heck & Davis-Stober, 2019), it will thus become easier for researchers to test psychological theories.

References

Agresti, A., & Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. Statistical Methods & Applications, 14, 297–330. http://dx.doi.org/10.1007/s10260-005-0121-y.

Aassf, B., Gavrilow, E., Herr, K., Joswig, M., Lorenz, B., Pfaffenholtz, A., & Rehn, T. (2017). Computing convex hulls and counting integer points with polymake. Mathematical Programming Computation, 9, 1–38. http://dx.doi.org/10.1007/s12532-016-0104-z.

Avis, D., Brenner, D., & Seidel, R. (1997). How good are convex hull algorithms? In 11th ACM symposium on computational geometry Computational Geometry, 7, 265–301. http://dx.doi.org/10.1006/comg.1996.0023-5.

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model’s testability and identifiability. Journal of Mathematical Psychology, 44, 20–40. http://dx.doi.org/10.1016/j.jmp.1999.1275.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). Statistical inference under order restrictions: Theory and application of isotonic regression. London, New York: Wiley.

Broder, A., & Schiffer, S. (2003). Bayesian strategy assessment in multiattribute decision making. Journal of Behavioral Decision Making, 16, 193–213. http://dx.doi.org/10.1002/bdm.442.

Brodteard, A. (2012). An introduction to convex polytopes. New York, NY: Springer.

Carbone, E., & Hey, J. D. (2000). Which error story is best? Journal of Risk and Uncertainty, 20, 161–176. http://dx.doi.org/10.1023/A:1007829024107.

Cavagnaro, D. R., & Davis-Stober, C. P. (2018). A model-based test for treatment effects with probabilistic classifications. Psychological Methods, 23, 672–689. http://dx.doi.org/10.1037/met0001737.

Christof, T., Lo, A., & Stoeber, M. (1997). Porta - polyhedron representation transformation algorithm. Retrieved from http://porta.zib.de/.

Cyrus, M., & Beck, J. (1978). Generalized two- and three-dimensional clipping. Computers & Graphics, 3, 23–28. http://dx.doi.org/10.1016/0097-8493(78)90021-3.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: applications to measurement theory. Journal of Mathematical Psychology, 53, 1–13. http://dx.doi.org/10.1016/j.jmp.2008.08.003.

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. Journal of Mathematical Psychology, 56, 86–94. http://dx.doi.org/10.1016/j.jmp.2012.01.003.

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2015). Individual differences in the algebraic structure of preferences. Journal of Mathematical Psychology, 66, 70–82. http://dx.doi.org/10.1016/j.jmp.2014.12.003.

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2018). Erratum to Davis-Stober et al. (2015): individual differences in the algebraic structure of preference, (in preparation).

Davis-Stober, C. P., Doignon, J.-P., Fiorini, S., Glineur, F., & Regenwetter, M. (2018). Extended formulations for order polytopes through network flows. Journal of Mathematical Psychology, 87, 1–10. http://dx.doi.org/10.1016/j.jmp.2018.08.003.

Davis-Stober, C. P., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. Journal of Mathematical Psychology, 72, 116–129. http://dx.doi.org/10.1016/j.jmp.2015.08.004.

Devroye, L. (1986). Non-uniform random variate generation. New York, NY: Springer.

Doignon, J.-P., & Rychep, S. (2016). Primary facets of order polytopes. Journal of Mathematical Psychology, 75, 231–245. http://dx.doi.org/10.1016/j.jmp.2016.07.004.

Efron, B., & Morris, C. (1975). Data analysis using Stein’s estimator and its generalizations. Journal of the American Statistical Association, 70, 311–319. http://dx.doi.org/10.1080/01621459.1975.10479864.

Fishburn, P. C. (1992). Induced binary probabilities and the linear ordering polytope: a status report. Mathematical Social Sciences, 23, 67–80. http://dx.doi.org/10.1016/0165-4896(92)90038-7.

Fukuda, K. (2004). Frequently asked questions in polyhedral computation. Retrieved from http://www.cs.mcgill.ca/~fukuda/soft/polyfaq/polyfaq.html.

Gelfand, A. E., Smith, A. F., & Lee, T.-M. (1992). Bayesian analysis of constrained parameters for state-trace analysis. Journal of Mathematical Psychology, 75, 231–245. http://dx.doi.org/10.1016/0097-8493(92)90024-3.

Ghosh, M. (2011). Objective priors: an introduction for frequentists. Journal of the American Statistical Association, 106, 233–248. http://dx.doi.org/10.1198/jasa.2011.101111.

Grn, D., & Davis-Stober, C. P. (2019). A caveat on the Savage-Dickey density ratio: the case of computing Bayes factors for regression parameters. British Journal of Mathematical and Statistical Psychology, http://dx.doi.org/10.1111/bmsp.12150, (in press).
Silvapulle, M. J., & Sen, P. K. (2004). Constrained statistical inference: Order, inequality, and shape constraints. Hoboken, NJ: Wiley.

Smeulders, B., Davis-Stober, C., Regenwetter, M., & Spieksma, F. C. (2018). Testing probabilistic models of choice using column generation. Computers & Operations Research, 95, 32–43. http://dx.doi.org/10.1016/j.cor.2018.03.001.

Smith, R. L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. Operations Research, 32, 1296–1308. http://dx.doi.org/10.1287/opre.32.6.1296.

Stan Development Team (2018). Stan modeling language users guide and reference manual. Version 2.18. Retrieved from http://mc-stan.org/.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. NeuroImage, 46, 1004–1017. http://dx.doi.org/10.1016/j.neuroimage.2009.03.025.

Suck, R. (1992). Geometric and combinatorial properties of the polytope of binary choice probabilities. Mathematical Social Sciences, 23, 81–102. http://dx.doi.org/10.1016/0165-4896(92)90039-8.

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. http://dx.doi.org/10.1016/j.csda.2010.03.016.