Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Gibbs Sampler for a Class of Random Convex Polytopes

Pierre E. Jacob, Ruobin Gong, Paul T. Edlefsen & Arthur P. Dempster

To cite this article: Pierre E. Jacob, Ruobin Gong, Paul T. Edlefsen & Arthur P. Dempster (2021): A Gibbs Sampler for a Class of Random Convex Polytopes, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1881523

To link to this article: https://doi.org/10.1080/01621459.2021.1881523

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2021.1881523

Taylor & Francis
Taylor & Francis Group

A Gibbs Sampler for a Class of Random Convex Polytopes

Pierre E. Jacob[a], Ruobin Gong[b], Paul T. Edlefsen[c], and Arthur P. Dempster[a]

[a]Department of Statistics, Harvard University, Cambridge, MA; [b]Department of Statistics, Rutgers University, New Brunswick, NJ; [c]Fred Hutchinson Cancer Research Center, Seattle, WA

ABSTRACT

We present a Gibbs sampler for the Dempster–Shafer (DS) approach to statistical inference for categorical distributions. The DS framework extends the Bayesian approach, allows in particular the use of partial prior information, and yields three-valued uncertainty assessments representing probabilities "for," "against," and "don't know" about formal assertions of interest. The proposed algorithm targets the distribution of a class of random convex polytopes which encapsulate the DS inference. The sampler relies on an equivalence between the iterative constraints of the vertex configuration and the nonnegativity of cycles in a fully connected directed graph. Illustrations include the testing of independence in 2×2 contingency tables and parameter estimation of the linkage model.

ARTICLE HISTORY
Received June 2020
Accepted January 2021

KEYWORDS
Algorithms; Bayesian methods; Categorical data analysis; Simulation

1. Introduction

Consider observed counts of K possible categories, denoted by $N_1, \ldots, N_K$ and summing to N. We assume that these counts are sums of independent draws from a categorical distribution. The goal is to infer the associated parameters θ in the simplex of dimension K and to forecast future observations. The setting is most familiar to statisticians and if K is small relative to N, and without further information about θ, the story is somewhat simple with the maximum likelihood estimator being both very intuitive and efficient. The plot thickens quickly if N is small or indeterminate, if partial prior information is available, if observations are imperfect, or if additional constraints are imposed, especially when uncertainty quantification is simultaneously sought (Fitzpatrick and Scott 1987; Berger and Bernardo 1992; Sison and Glaz 1995; Liu 2000; Lang 2004; Chafai and Concordet 2009; Dunson and Xing 2009). As any probability distribution on a finite unordered set is necessarily categorical, the setting often arises as part of more elaborate procedures. Besides, the canonical nature of categorical distributions has made them a common test bed for various approaches to inference (Walley 1996; Bernard 1998).

The Dempster–Shafer (DS) theory is a framework for probabilistic reasoning based on observed data and modeling of knowledge. In the DS framework, inferences on user-defined assertions are expressed probabilistically. These assertions can be statements concerning parameters ("the parameter belongs to a certain set") or concerning future observations. Contrary to Bayesian inference, no prior distribution is strictly required, and partial prior specification is allowed (see Section 4.2). Rather than posterior probabilities, DS inference yields three-valued assessments of uncertainty, namely probabilities "for," "against," and "don't know" associated with the assertion of interest, and denoted by $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ (see Section 2.2). In his pioneering work, Dempster (1963, 1966, 1967, 1968) developed the idea of upper and lower probabilities for assertions of interest. Together with the contributions of Shafer (1976, 1979), the approach became known as the DS theory of belief functions. The framework has various connections to other ways of obtaining lower and upper probabilities and to robust Bayesian inference (Wasserman 1990). Over the past decades, the DS theory saw various applications in signal processing, computer vision and machine learning (see, e.g., Bloch 1996; Vasseur et al. 1999; Denoeux 2000; Basir and Yuan 2007; Denoeux 2008; Díaz-Más et al. 2010). As outlined in Dempster (2008, 2014), the DS framework is an ambitious tool for carrying out statistical inferences in scientific practice. During the developments of DS, categorical distributions were front and center due to their generality and relevance to ubiquitous statistical objects such as contingency tables.

The computation required by the DS approach for categorical distributions proved to be demanding. The approach involves distributions of convex polytopes within the simplex, some properties of which were found in Dempster (1966, 1968, 1972). Unfortunately, no closed-form joint distribution of the vertices has been found, hindering both theoretical developments and numerical approximations. The challenge prompted Denœux (2006) to comment that, "Dempster studied the trinomial case […] However, the application of these results to compute the marginal belief function […] has proved, so far and to our knowledge, mathematically intractable." Likewise, Lawrence et al. (2009) commented: "[…] his method for the multinomial model is seldom used, partly because of its computational complexity." Over the past fifty years, the literature saw a handful of alternative methods for categorical inference via generalized fiducial inference (Hannig et al. 2016; Liu and Hannig 2016) the imprecise Dirichlet model (IDM) (Walley 1996),

CONTACT Pierre E. Jacob pjacob@g.harvard.edu Department of Statistics, Harvard University, Cambridge, MA 02138.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

the Dirichlet-DSM method (Lawrence et al. 2009), the vector-valued Poisson model (Edlefsen, Liu, and Dempster 2009), and the Poisson-projection method for multinomial data (Gong 2018, unpublished PhD thesis). The latter three methods were motivated in part to circumvent the computational hurdle put forward by the original DS formulation. The present article aims at filling that gap by proposing an algorithm that carries out the computation proposed in Dempster (1966, 1972). The presentation does not assume previous knowledge on DS inference.

Section 2 introduces the formal setup. Section 3 presents an equivalence between constraints arising in the definition of the problem and the existence of "negative cycles" (defined in Section 3.2) in a certain weighted graph, that leads to a Gibbs sampler. Various illustrations and extensions are laid out in Section 4. Section 5 concerns applications to 2×2 contingency tables and the linkage model. Elements of future research are discussed in Section 6.

2. Inference in Categorical Distributions

We describe inference in categorical distributions as proposed in Dempster (1966), using the following notation. The observations are $\mathbf{x} = (x_n)_{n\in[N]}$, with $x_n \in [K]$ for all $n \in [N]$, where $[m]$ denotes the set $\{1, \ldots, m\}$ for $m \geq 1$. The number of categories is K. The K-dimensional simplex is $\Delta = \{z \in \mathbb{R}_+^K : \sum_{k=1}^K z_k = 1\}$. The set of measurable subsets of Δ is denoted by $\mathcal{B}(\Delta)$. We denote the vertices of Δ by $V_1, \ldots, V_K$. In barycentric coordinates, V_k is a K-vector with kth entry equal to one and other entries equal to zero. A polytope is a set of points $z \in \mathbb{R}^K$ satisfying linear inequalities, of the form $Mz \leq c$ understood component-wise, and where M is a matrix with K columns and c is a vector. For a given $\mathbf{x} \in [K]^N$, $\mathcal{I}_k$ is the set of indices $\{n \in [N] : x_n = k\}$. The counts are $N_k = |\mathcal{I}_k|$ and $\sum_{k\in[K]} N_k = N$. Coordinates of $u_n \in \Delta$ are denoted by $u_{n,k}$ for $k \in [K]$. The volume of a set A is denoted by $\mathrm{Vol}(A)$. The uniform variable Z over $\mathcal{S}$ is written $Z \sim \mathcal{S}$.

2.1. Sampling Mechanism and Feasible Sets

The goal is to infer the parameters $\theta = (\theta_1, \ldots, \theta_K) \in \Delta$ of a categorical distribution using observation $\mathbf{x} = (x_n)_{n\in[N]} \in [K]^N$. Viewing x_n as a random variable, the model states $\mathbb{P}(x_n = k) = \theta_k$ for all $k \in [K]$, $n \in [N]$. Generating draws from a categorical distribution can be done in different ways. In the DS approach, the choice of sampling mechanism of the observable data has an impact on inference of the parameters, a feature that distinguishes DS from likelihood-based approaches, and aligns it with fiducial (Fisher 1935; Hannig et al. 2016), structural (Fraser 1968), and functional (Dawid and Stone 1982) approaches. Appendix A in the supplementary materials illustrates this impact in a simple setting. We follow Dempster (1966) and consider the following sampling mechanism for x_n, which is invariant by permutation of the labels of the categories; it is equivalent to the "Gumbel-max trick" (Maddison, Tarlow, and Minka 2014) as explained in Appendix B in the supplementary materials. Given θ, for each $k \in [K]$, define $\Delta_k(\theta)$ to be a "subsimplex" obtained as the polytope with the same vertices as Δ except that vertex V_k is replaced by θ. The sets $(\Delta_k(\theta))_{k\in[K]}$ form a partition of Δ, shown in Figure 1(a). It can be checked that $\mathrm{Vol}(\Delta_k(\theta)) = \theta_k$. Then, introduce $u_n \sim \Delta$, and define x_n as

$$x_n = \sum_{k\in[K]} k \mathbb{1}(u_n \in \Delta_k(\theta)). \tag{1}$$

In other words, x_n is the unique index $k \in [K]$ such that u_n belongs to $\Delta_k(\theta)$. Since $\mathrm{Vol}(\Delta_k(\theta)) = \theta_k$, x_n indeed follows the categorical distribution with parameter θ. Lemma 2.1 recalls a useful characterization of $\Delta_k(\theta)$.

Lemma 2.1 (Dempster 1966, Lemma 5.2). For $k \in [K]$, $\theta \in \Delta$ and $u_n \in \Delta$, $u_n \in \Delta_k(\theta)$ if and only if $u_{n,\ell}/u_{n,k} \geq \theta_\ell/\theta_k$ for all $\ell \in [K]$.

Given fixed observations $\mathbf{x} = (x_n)_{n\in[N]}$, the sampling mechanism in (1) can be turned into constraints on the values of $\mathbf{u} = (u_n)_{n\in[N]}$ and θ that could have led to the observations. A central piece of the machinery is the following set,

$$\mathcal{R}_\mathbf{x} = \big\{(u_1, \ldots, u_N) \in \Delta^N : \exists \theta \in \Delta \quad \forall n \in [N] \quad u_n \in \Delta_{x_n}(\theta)\big\}. \tag{2}$$

It is the set of all possible realizations of $\mathbf{u}$ which could have produced the data $\mathbf{x}$ for (at least) some θ, via the specified sampling

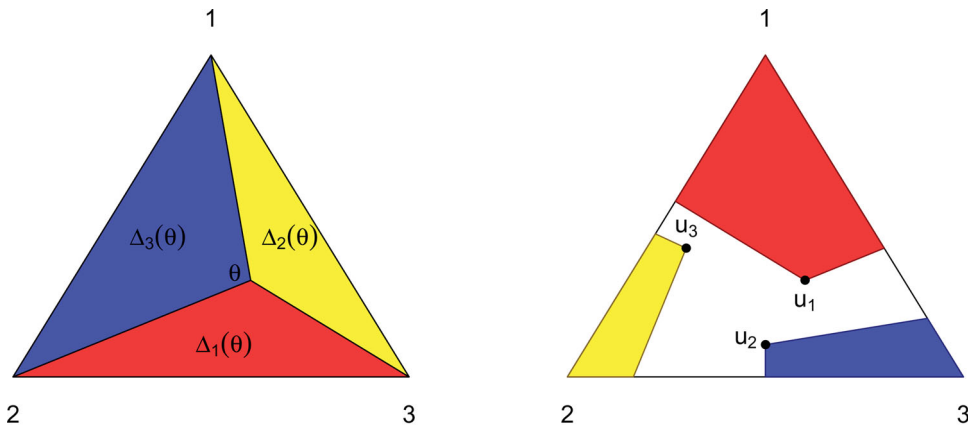

Figure 1. Partition of Δ into $(\Delta_k(\theta))_{k\in[K]}$ in 1a, with $K = 3$. Each point $u_n \in \Delta$ defines, for a fixed $x_n \in [K]$, a set of $\theta \in \Delta$ such that $u_n \in \Delta_{x_n}(\theta)$; three such sets are colored in 1b, for $x_1 = 1, x_2 = 3, x_3 = 2$. Here, no $\theta \in \Delta$ is such that $u_n \in \Delta_{x_n}(\theta)$ for $n = 1, 2, 3$.

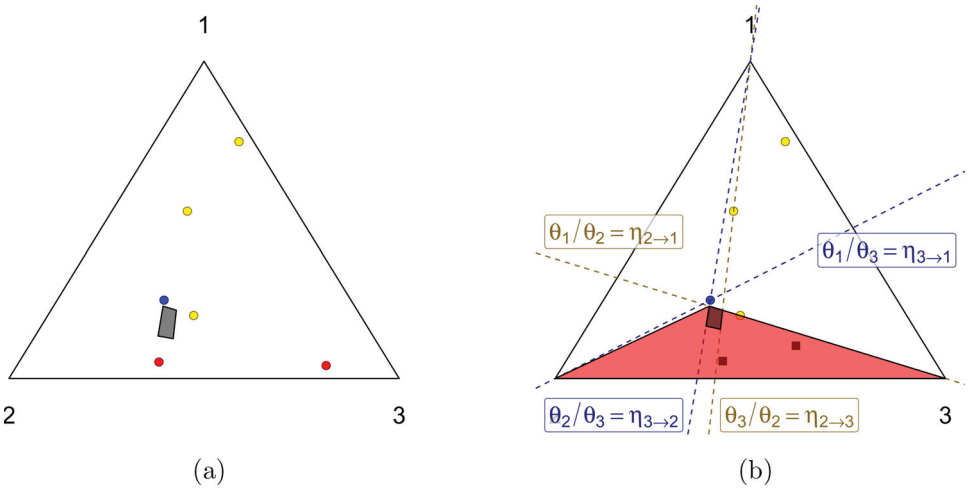

Figure 2. Given $\mathbf{u} \in \mathcal{R}_\mathbf{x}$ (shown in (a)), drop components $\mathbf{u}_{\mathcal{I}_k}$ for some $k \in [K]$ (the red dots in (a)) and draw new components $\mathbf{u}_{\mathcal{I}_k}$ (the red squares in (b)) from their conditional distribution identified in Proposition 3.2 with support represented by the shaded triangle.

mechanism. Given a realization of $\mathbf{u} \in \mathcal{R}_\mathbf{x}$ by definition there is a non-empty "feasible" set $\mathcal{F}(\mathbf{u}) \subset \Delta$ defined as

$$\mathcal{F}(\mathbf{u}) = \left\{\theta \in \Delta\colon\ \forall n \in [N] \quad u_n \in \Delta_{x_n}(\theta)\right\}. \tag{3}$$

On the other hand if $\mathbf{u}$ is an arbitrary point in Δ^N then $\mathcal{F}(\mathbf{u})$ defined above can be empty. For example, $\mathbf{u} = (u_1, u_2, u_3)$ shown in Figure 1(b) leads to an empty $\mathcal{F}(\mathbf{u})$ for the observations $x_1 = 1, x_2 = 3, x_3 = 2$. The goal of the proposed method is to obtain non-empty sets $\mathcal{F}(\mathbf{u})$, as illustrated for another dataset in Figure 2(a). We can rewrite (2) as $\mathcal{R}_\mathbf{x} = \{\mathbf{u}\colon \mathcal{F}(\mathbf{u}) \neq \emptyset\}$.

The ingredients introduced thus far specify the "source" of a belief function (e.g., Wasserman 1990). The central object of interest here is the distribution of the random sets $\mathcal{F}(\mathbf{u})$ conditional on them being nonempty. We consider the uniform distribution on $\mathcal{R}_\mathbf{x}$ denoted by $\nu_\mathbf{x}$, with density

$$\begin{aligned}\forall u_1, \ldots, u_N \in \Delta^N \quad & \nu_\mathbf{x}(u_1, \ldots, u_N) \\ &= \mathrm{Vol}\,(\mathcal{R}_\mathbf{x})^{-1}\, \mathbb{1}\left((u_1, \ldots, u_N) \in \mathcal{R}_\mathbf{x}\right).\end{aligned} \tag{4}$$

Our main contribution is an algorithm to sample $\mathbf{u}$ from $\nu_\mathbf{x}$. The sets $\mathcal{F}(\mathbf{u})$ obtained when $\mathbf{u} \sim \nu_\mathbf{x}$ constitute the class of random convex polytopes studied in Dempster (1972) and referred to in the title of the present article. The distribution $\nu_\mathbf{x}$ is also the result of Dempster's rule of combination (Dempster 1967) applied to the information provided separately by each of the N observations.

2.2. Inference Using Random Sets

We recall briefly how random sets can be processed into "lower" and "upper" probabilities as in Dempster (1966), or into "belief" and "plausibility" as in Shafer (1976, 1990) and Wasserman (1990), or $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ probabilities as in Dempster (2008). The user provides a measurable subset $\Sigma \in \mathcal{B}(\Delta)$ corresponding to an "assertion" of interest about the parameter, for instance, $\Sigma = \{\theta \in \Delta\colon \theta_1 \leq 1/3\}$, or $\Sigma = \{\theta \in \Delta\colon \theta_1/\theta_2 > \theta_3/\theta_4\}$. The belief function assigns a value to each $\Sigma \in \mathcal{B}(\Delta)$ defined as $\mathrm{Bel}(\Sigma) = \nu_\mathbf{x}(\{\mathbf{u}\colon \mathcal{F}(\mathbf{u}) \subset \Sigma\})$. This can be called lower probability and written $\underline{P}(\Sigma)$. The upper probability or "plausibility" $\bar{P}(\Sigma)$ is defined as $1 - \underline{P}(\Sigma^c)$, or equivalently $\nu_\mathbf{x}(\{\mathbf{u}\colon \mathcal{F}(\mathbf{u}) \cap \Sigma \neq \emptyset)\}$. Bayesian inference is recovered exactly when combining the distribution of $\mathcal{F}(\mathbf{u})$ obtained from $\mathbf{x}$ with a prior distribution on θ, see Dempster (1968) and Section 4.2. Following Dempster (2008) DS inference can be summarized via the probability triple $(\mathtt{p}, \mathtt{q}, \mathtt{r})$:

$$\mathtt{p}(\Sigma) = \underline{P}(\Sigma), \quad \mathtt{q}(\Sigma) = 1 - \bar{P}(\Sigma), \quad \mathtt{r}(\Sigma) = \bar{P}(\Sigma) - \underline{P}(\Sigma), \tag{5}$$

with $\mathtt{p} + \mathtt{q} + \mathtt{r} = 1$ for all Σ, quantifying support "for," "against," and "don't know" about the assertion Σ. As argued in Dempster (2008) and Gong (2019), the triple $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ draws a stochastic parallel to the three-valued logic, with $\mathtt{r}$ representing weight of evidence in a third, indeterminate logical state. A $\mathtt{p}$ or $\mathtt{q}$ value close to 1 is interpreted as strong evidence toward Σ or Σ^c, respectively. A large $\mathtt{r}$ suggests that the model and data are structurally deficient in making precise judgment about the assertion Σ or its negation.

Sampling methods enable approximations of these probabilities via standard Monte Carlo arguments. A simple strategy is to draw $(u_n)_{n\in[N]}$ from the uniform on Δ^N until $\mathcal{F}(\mathbf{u})$ is nonempty. However, the rejection rate would be prohibitively high as N increases. Some properties of $\nu_\mathbf{x}$ have been obtained in Dempster (1966, 1972). For example, Equation (2.1) in Dempster (1972) states that, for a fixed $\theta \in \Delta$, $\nu_\mathbf{x}(\{\mathbf{u}\colon \theta \in \mathcal{F}(\mathbf{u})\})$ is equal to the multinomial probability mass function with parameter θ evaluated at $N_1, \ldots, N_K$. Equation (2.5) in Dempster (1972) gives the expected volume of $\mathcal{F}(\mathbf{u})$. Dempster (1972) also obtained the distribution of vertices of $\mathcal{F}(\mathbf{u})$ under $\mathbf{u} \sim \nu_\mathbf{x}$ with smallest and largest coordinate θ_k for any $k \in [K]$, which are Dirichlet distributions. These enable the approximation of $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ for certain assertions, including the sets $\{\theta\colon \theta_k \in [0, c]\}$ for arbitrary $c \in [0, 1]$. However, for general assertions the joint distribution of all vertices of $\mathcal{F}(\mathbf{u})$ under $\mathbf{u} \sim \nu_\mathbf{x}$ is necessary, as in the case of both applications in Section 5.

3. Proposed Gibbs Sampler

3.1. Strategy

The proposed algorithm is a Markov chain Monte Carlo (MCMC) method targeting $\nu_\mathbf{x}$, thus referred to as the target

Algorithm 1 Uniform sampling in $\Delta_k(\theta)$.

Input: $k \in [K]$, $\theta \in \Delta$, and the vertices of Δ denoted by $(V_\ell)_{\ell \in [K]}$.

- Sample $(w_1, \ldots, w_K)$ uniformly on Δ, for example, $\tilde{w}_\ell \sim$ Exponential(1) for all $\ell \in [K]$ and $w_\ell = \tilde{w}_\ell / \sum_{j=1}^K \tilde{w}_j$.
- Define the point $z = w_k \theta + \sum_{\ell \neq k} w_\ell V_\ell$, for example, $z_k = w_k \theta_k$ and $z_\ell = w_k \theta_\ell + w_\ell$ for $\ell \neq k$.
- Return z, a uniformly distributed point in $\Delta_k(\theta)$.

distribution. At the initial step, we set $\theta^{(0)}$ arbitrarily in Δ, for example, a draw from a Dirichlet distribution. Given $\theta^{(0)}$ we can sample, for $k \in [K]$ and $n \in \mathcal{I}_k$, $u_n \sim \Delta_k(\theta^{(0)})$. Then $\mathbf{u} = (u_n)_{n \in [N]}$ is in $\mathcal{R}_\mathbf{x}$ because $\theta^{(0)}$ is in $\mathcal{F}(\mathbf{u})$ by construction. Sampling uniformly over $\Delta_k(\theta)$ can be done following equation (5.7) in Dempster (1966), as recalled in Algorithm 1. In this section, we assume that $N_k = |\mathcal{I}_k| \geq 1$ for all $k \in [K]$, and describe how to handle empty categories in Section 4.1.

We draw components of $\mathbf{u}$ from conditional distributions given the other components under $\nu_\mathbf{x}$, namely we draw $\mathbf{u}_{\mathcal{I}_k} = (u_n)_{n \in \mathcal{I}_k}$ for $k \in [K]$ from $\nu_\mathbf{x}(d\mathbf{u}_{\mathcal{I}_k} | \mathbf{u}_{[N] \setminus \mathcal{I}_k})$. Drawing $\mathbf{u}_{\mathcal{I}_k}$ from this conditional distribution will constitute an iteration of a Gibbs sampler, illustrated in Figure 2 for the data $N_1 = 2, N_2 = 3, N_3 = 1$. Figure 2(a) shows a sample $\mathbf{u} \in \mathcal{R}_\mathbf{x}$, with each u_n colored according to $x_n \in [K]$. Sampling from $\nu_\mathbf{x}(d\mathbf{u}_{\mathcal{I}_k} | \mathbf{u}_{[N] \setminus \mathcal{I}_k})$ can be understood as drawing all the points of the same color conditional on the other points. The overall Gibbs sampler cycles through the K categories to generate a sequence of draws $\mathbf{u}^{(t)}$ that converges to $\nu_\mathbf{x}$ as $t \to \infty$, for example, in distribution. To each $\mathbf{u}^{(t)}$ is associated a feasible set $\mathcal{F}(\mathbf{u}^{(t)})$ that can contribute to the approximation of $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ triples described in Section 2.2. The next question is how to sample from the adequate conditional distributions. Toward this aim, we will draw on a representation of $\mathcal{R}_\mathbf{x}$ connected to the presence of "negative cycles" in a complete graph with K vertices.

3.2. Non-emptiness of Feasible Sets

We can represent $\mathcal{R}_\mathbf{x}$ without mention of the existence of some $\theta \in \Delta$ as in (2), but instead with explicit constraints on the components of $\mathbf{u}$. We first find an equivalent representation of $\theta \in \mathcal{F}(\mathbf{u})$ for a fixed $\mathbf{u}$. By definition $\theta \in \mathcal{F}(\mathbf{u})$ satisfies for all $n \in [N]$ $u_n \in \Delta_{x_n}(\theta)$. For each $k \in [K]$, using Lemma 2.1 we write

$$\forall n \in \mathcal{I}_k \quad \forall \ell \in [K] \setminus \{k\} \quad \frac{u_{n,\ell}}{u_{n,k}} \geq \frac{\theta_\ell}{\theta_k} \quad \Leftrightarrow \quad \forall \ell \in [K] \setminus \{k\} \quad \min_{n \in \mathcal{I}_k} \frac{u_{n,\ell}}{u_{n,k}} \geq \frac{\theta_\ell}{\theta_k}.$$

This prompts the definition

$$\forall k \in [K] \quad \forall \ell \in [K] \quad \eta_{k \to \ell}(\mathbf{u}) = \min_{n \in \mathcal{I}_k} \frac{u_{n,\ell}}{u_{n,k}}. \tag{6}$$

Observe that the values $(\eta_{k \to \ell}(\mathbf{u}))$ depend on the observations through the sets $(\mathcal{I}_k)$. At this point, $\theta \in \mathcal{F}(\mathbf{u})$ is equivalent to $\theta_\ell / \theta_k \leq \eta_{k \to \ell}(\mathbf{u})$ for $\ell, k \in [K]$. Next assume $\theta \in \mathcal{F}(\mathbf{u})$ and consider some implications. First, for all k, ℓ

$$\frac{\theta_\ell}{\theta_k} \leq \eta_{k \to \ell}(\mathbf{u}), \quad \text{and} \quad \frac{\theta_k}{\theta_\ell} \leq \eta_{\ell \to k}(\mathbf{u}), \quad \text{thus} \quad \eta_{k \to \ell}(\mathbf{u}) \eta_{\ell \to k}(\mathbf{u}) \geq 1.$$

If $K \geq 3$ we can write θ_ℓ / θ_k as $(\theta_\ell / \theta_j)(\theta_j / \theta_k)$, and apply a similar reasoning to obtain the inequalities $\eta_{\ell \to k}(\mathbf{u}) \eta_{k \to j}(\mathbf{u}) \eta_{j \to \ell}(\mathbf{u}) \geq 1$ for all k, ℓ, j. Overall we can write, for all $K \geq 2$, with any number L of indices $j_1, \ldots, j_L \in [K]$, the following constraints:

$$\forall L \in [K] \quad \forall j_1, \ldots, j_L \in [K] \quad \eta_{j_1 \to j_2}(\mathbf{u}) \eta_{j_2 \to j_3}(\mathbf{u}) \ldots \eta_{j_L \to j_1}(\mathbf{u}) \geq 1. \tag{7}$$

Hereafter we drop "$(\mathbf{u})$" from the notation for clarity. The case $L = 1$ gives inequalities $\eta_{k \to k} \geq 1$ which are always satisfied since $\eta_{k \to k} = 1$ following (6). Furthermore, it suffices to consider only indices $j_1, \ldots, j_L$ that are unique, otherwise the associated inequality in (7) is implied by inequalities associated with smaller values of L.

At this point, we observe a fruitful connection between (7) and directed graphs. The indices in $[K]$ can be viewed as vertices of a fully connected directed graph. Directed edges are ordered pairs (j_1, j_2). We associate the product $\eta_{j_1 \to j_2} \eta_{j_2 \to j_3} \ldots \eta_{j_L \to j_1}$ with a sequence of edges, (j_1, j_2), (j_2, j_3), up to (j_L, j_1). That sequence forms a path from vertex j_1 back to vertex j_1, of length L, in other words a directed cycle of length L. Define $w_{k \to \ell} = \log \eta_{k \to \ell}$ for all $k, \ell \in [K]$, and treat it as the weight of edge (k, ℓ). Then the inequality (7) is equivalent to $w_{j_1 \to j_2} + w_{j_2 \to j_3} + \cdots + w_{j_L \to j_1} \geq 0$. The sum of weights along a path is called its "value." The inequalities in (7) are then equivalent to all cycles in the graph having nonnegative values. See Figure 3 for an illustration for $K = 3$ of the equivalent conception of constraints in (7) as graph cycle values. Detecting whether graphs contain cycles with negative values, called "negative cycles," can be done with the Bellman–Ford algorithm (Bang-Jensen and Gutin 2008).

At this point we have established that $\theta \in \mathcal{F}(\mathbf{u})$ implies the inequalities of (7), which can be understood as constraints on the weights of a graph. Our next result states that the converse also holds.

Proposition 3.1. There exists $\theta \in \Delta$ satisfying $\theta_\ell / \theta_k \leq \eta_{k \to \ell}$ for all $k, \ell \in [K]$ if and only if the values $(\eta_{k \to \ell})$ satisfy

$$\forall L \in [K] \quad \forall j_1, \ldots, j_L \in [K] \quad \eta_{j_1 \to j_2} \eta_{j_2 \to j_3} \ldots \eta_{j_L \to j_1} \geq 1. \tag{8}$$

Furthermore it suffices to restrict (8) to distinct indices $j_1, \ldots, j_L$.

Proof. The proof of the reverse implication explicitly constructs a feasible θ based on the values $(\eta_{k \to \ell})$, assuming that they satisfy (8). Introduce the fully connected graph with K vertices, with weight $\log \eta_{k \to \ell}$ on edge (k, ℓ). Thanks to $(\eta_{k \to \ell})$ satisfying (8), there are no negative cycles thus one cannot decrease the value of a path by appending a cycle to it. Since there are only finitely many paths without cycles there is a finite minimal value over all paths from k to ℓ, which we denote by $\min(k \to \ell)$. In other words (8) implies that $\min(k \to \ell)$ is finite.

We choose a vertex in $[K]$ arbitrarily, for instance vertex K. We define θ by $\theta_k = \exp(\min(K \to k))$ and then by

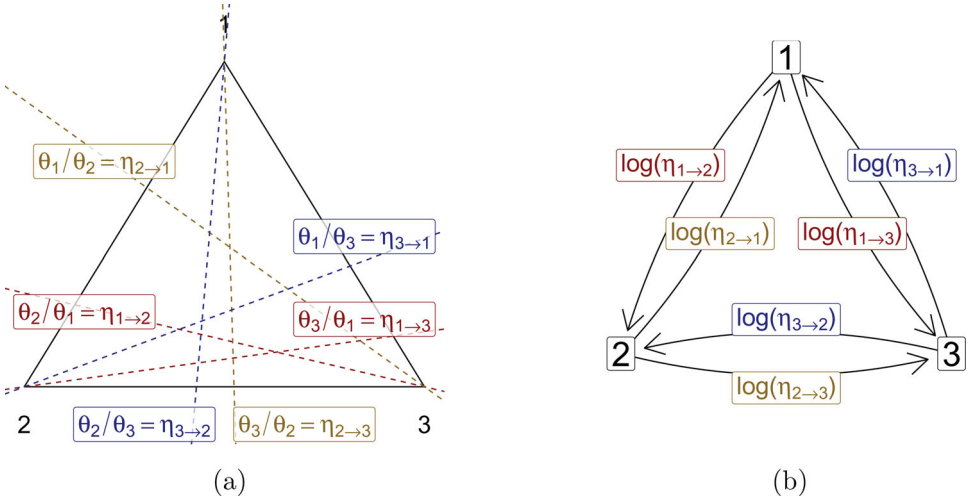

Figure 3. Two views on the constraints in (7). In (a), the values $\eta_{k\to\ell}$ define linear constraints $\theta_\ell/\theta_k = \eta_{k\to\ell}$. In (b), the log values are weights on the edges of a complete directed graph.

normalizing the entries so that $\theta \in \Delta$. We can write $\min(K \to \ell) \le \min(K \to k) + \log \eta_{k\to\ell}$, because the right-hand side is the value of a path from K to ℓ (via k), while the left-hand side is the smallest value over all such paths. Upon taking the exponential, the above inequality is equivalent to $\theta_\ell/\theta_k \le \eta_{k\to\ell}$. □

3.3. Conditional Distributions

Thanks to Proposition 3.1 we can write $\mathcal{R}_\mathbf{x}$ defined in (2) as the set of $\mathbf{u}$ for which the values $\eta_{k\to\ell}$ satisfy (8), with $\eta_{k\to\ell}$ defined in (6). We next provide a representation of the conditional distribution of $\mathbf{u}_{\mathcal{I}_k}$ under $\nu_\mathbf{x}$ that is convenient for sampling purposes.

Proposition 3.2. Let $\mathbf{u} = (u_1, \dots, u_N) \in \mathcal{R}_\mathbf{x}$, and define $\eta_{k\to\ell} = \min_{n\in\mathcal{I}_k} u_{n,\ell}/u_{n,k}$ for all $k, \ell \in [K]$. Let $k \in [K]$. Define for $\ell \in [K]$,

$$\theta_\ell = \frac{\exp(-\min(\ell \to k))}{\sum_{\ell' \in [K]} \exp(-\min(\ell' \to k))}, \tag{9}$$

where $\min(\ell \to k)$ is the minimum value over all paths from ℓ to k, in a fully connected directed graph with weight $\log \eta_{j\to\ell}$ on edge (j, ℓ). Then, $\nu_\mathbf{x}(d\mathbf{u}_{\mathcal{I}_k} | \mathbf{u}_{[N]\setminus\mathcal{I}_k})$ is the uniform distribution on $\Delta_k(\theta)^{N_k}$.

In other words, $\nu_\mathbf{x}(d\mathbf{u}_{\mathcal{I}_k} | \mathbf{u}_{[N]\setminus\mathcal{I}_k})$ is the product measure with each component u_n following the uniform distribution on $\Delta_k(\theta)$, with θ defined in (9). The proposition is key to the implementation of the proposed Gibbs sampler.

Proof. We consider an arbitrary $k \in [K]$, and assume that $\mathbf{u} \in \mathcal{R}_\mathbf{x}$. Listing the inequalities in (8) that involve the index k and separating the terms $\eta_{k\to\ell}$ from the others, we obtain

$$\forall \ell \in [K] \qquad \eta_{k\to\ell} \ge \eta_{\ell\to k}^{-1}, \tag{10}$$

$$\forall \ell \in [K] \quad \forall L \in [K-2] \quad \forall j_1, \dots, j_L \in [K] \setminus \{k, \ell\} \quad \eta_{k\to\ell} \ge \left(\eta_{\ell\to j_1} \dots \eta_{j_L\to k}\right)^{-1}. \tag{11}$$

Thus, for $\mathbf{u}$ to remain in $\mathcal{R}_\mathbf{x}$ after updating its components $\mathbf{u}_{\mathcal{I}_k}$, it is enough to check that the ratios $u_{n,\ell}/u_{n,k}$ for $\ell \in [K]$ and $n \in \mathcal{I}_k$ are lower bounded as above.

The finiteness of $\min(\ell \to k)$ results from the same reasoning as in the proof of Proposition 3.1. Note that $\min(\ell \to k)$ can be constructed without the entries $\mathbf{u}_{\mathcal{I}_k}$ of $\mathbf{u}$, because the shortest path from ℓ to k should pay no attention to any directed edges that stem from k, and the entries $\mathbf{u}_{\mathcal{I}_k}$ inform only the weights of edges stemming from k. Thus, we can define θ as in (9).

We next show that the support of $\nu_\mathbf{x}(d\mathbf{u}_{\mathcal{I}_k} | \mathbf{u}_{[N]\setminus\mathcal{I}_k})$ is exactly $\Delta_k(\theta)^{N_k}$. Let $\mathbf{u}_{\mathcal{I}_k} \in \Delta_k(\theta)^{N_k}$. By Lemma 2.1 and the definition of θ, we have

$$\forall \ell \in [K] \quad \min_{n\in\mathcal{I}_k} \frac{u_{n,\ell}}{u_{n,k}} \ge \exp(-\min(\ell \to k)).$$

But $\exp(-\min(\ell \to k)) = (\exp(\min(\ell \to k)))^{-1}$ is greater than $(\eta_{\ell\to j_1} \dots \eta_{j_L\to k})^{-1}$ for any path $\ell \to j_1 \dots j_L \to k$. Thus, with $\eta_{k\to\ell} = \min_{n\in\mathcal{I}_k} u_{n,\ell}/u_{n,k}$, inequalities (10) and (11) are satisfied. Proposition 3.1 guarantees that $\mathbf{u}$ is in $\mathcal{R}_\mathbf{x}$, thus $\Delta_k(\theta)^{N_k}$ is contained in the support of $\nu_\mathbf{x}(d\mathbf{u}_{\mathcal{I}_k} | \mathbf{u}_{[N]\setminus\mathcal{I}_k})$.

Let us show the reverse inclusion by considering $\mathbf{u}_{\mathcal{I}_k} \notin \Delta_k(\theta)^{N_k}$. There, for some $n \in \mathcal{I}_k$ and some $\ell \in [K]$, we have $u_{n,\ell}/u_{n,k} < \exp(-\min(\ell \to k))$. Denote by $\ell \to j_1 \dots j_L \to k$ the path attaining the value $\min(\ell \to k)$. We obtain $\eta_{k\to\ell} \le u_{n,\ell}/u_{n,k} < (\eta_{\ell\to j_1} \dots \eta_{j_L\to k})^{-1}$, and thus $\eta_{k\to\ell}\eta_{\ell\to j_1} \dots \eta_{j_L\to k} < 1$, in other words some inequalities in (8) are not satisfied and thus, by Proposition 3.1, $\mathbf{u}$ is not in $\mathcal{R}_\mathbf{x}$. □

Proposition 3.2 provides a strategy to sample from the conditional distributions of interest, provided that we can obtain $\theta \in \Delta$ in (9), which involves $\min(\ell \to k)$ for all ℓ. These can be obtained from shortest path algorithms such as Bellman–Ford implemented in `igraph` (Csardi and Nepusz 2006). Alternatively we can view θ in (9) as the solution of the linear program,

$$\max \left\{ \theta_k : \theta \in \Delta \quad \forall j \ne k \quad \forall i \ne j \quad \frac{\theta_i}{\theta_j} \le \eta_{j\to i} \right\}. \tag{12}$$

This has a simple interpretation: θ in (9) is precisely the vertex of $\mathcal{F}(\mathbf{u})$ with the largest kth component. The equivalence

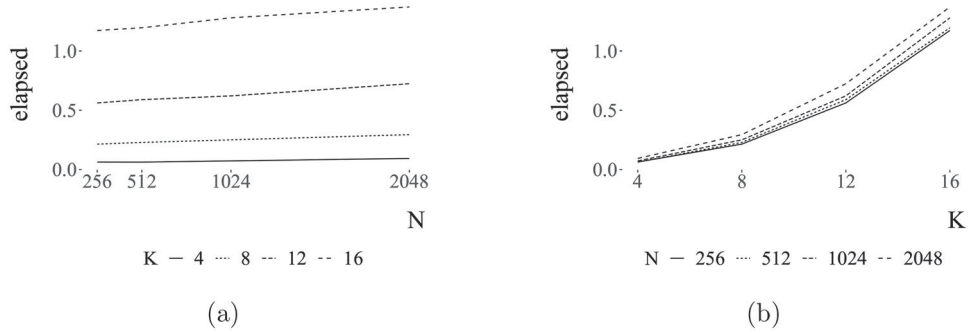

Figure 4. Elapsed time in seconds for 100 iterations of the sampler. In (a), elapsed time as a function of N, for different K. In (b), elapsed time as a function of K, for different N.

Algorithm 2 Gibbs sampler for categorical inference in the Dempster–Shafer framework. Input: observations $\mathbf{x} \in [K]^N$, defining index sets $\mathcal{I}_k = \{n \in [N] : x_n = k\}$ for $k \in [K]$. Output: sequence $(\mathbf{u}^{(t)})_{t\geq 0}$ converging to $\nu_{\mathbf{x}}$, the uniform distribution on $\mathcal{R}_{\mathbf{x}}$.

1. Set $\theta^{(0)}$ in Δ, and for all $k \in [K]$, all $n \in \mathcal{I}_k$, sample $u_n^{(0)} \sim \Delta_k(\theta^{(0)})$ (Algorithm 1).
2. Compute $\eta_{k\to\ell}^{(0)} = \min_{n\in\mathcal{I}_k} u_{n,\ell}^{(0)}/u_{n,k}^{(0)}$ for all $k, \ell \in [K]$.
3. At iteration $t \geq 1$,
   (a) Set $\eta_{k\to\ell}^{(t)} \leftarrow \eta_{k\to\ell}^{(t-1)}$ for all $k, \ell \in [K]$.
   (b) For category $k \in [K]$,
      i. Compute $\theta \in \Delta$ from the values $(\eta_{j\to\ell}^{(t)})$ according to (9),
      either by computing shortest paths or by solving a linear program (12).
      ii. For each $n \in \mathcal{I}_k$, sample $u_n^{(t)} \sim \Delta_k(\theta)$ (Algorithm 1).
      iii. Set $\eta_{k\to\ell}^{(t)} \leftarrow \min_{n\in\mathcal{I}_k} u_{n,\ell}^{(t)}/u_{n,k}^{(t)}$, for all $\ell \neq k$.

between shortest path problems and linear programs is well known. Implementations are provided in `lpsolve` (Berkelaar, Eikland, and Notebaert 2004; Konis 2014).

The Gibbs sampler is described in Algorithm 2. Its outputs include $\mathbf{u}^{(t)}$ converging to $\nu_{\mathbf{x}}$ in distribution as $t \to \infty$, as well as the associated values of $(\eta_{k\to\ell}^{(t)})$ from which we can obtain the sets $\mathcal{F}(\mathbf{u}^{(t)})$ as $\{\theta \in \Delta : \theta_\ell/\theta_k \leq \eta_{k\to\ell}^{(t)} \ \forall k, \ell \in [K]\}$. Such sets can be stored in "half-space representation" or as a list of vertices in Δ, obtained by vertex enumeration (Avis and Fukuda 1992). Convenient functions to store and manipulate polytopes can be found in `rcdd` (Fukuda 1997; Geyer and Meeden 2008). We run 100 iterations of the sampler and record elapsed seconds for different values of N and K. Medians over 50 experiments are reported in Figure 4, for counts set to $\lfloor N/K \rfloor$ in each category.

3.4. Convergence to Stationarity

A common question to all MCMC algorithms is the rate of convergence to stationary (Jerrum 1998; Roberts and Rosenthal 2004), which here might depend on K and the observed counts $N_1, \ldots, N_K$. In the simplest case where $K = 2$, with counts $N_1 \geq 1, N_2 \geq 1$, we obtain an upper bound on the mixing time of the chain in the 1-Wasserstein metric (e.g., Gibbs 2004) as detailed in Appendix C in the supplementary materials. We find that the upper bound increases at most linearly with the total count $N = N_1 + N_2$. The extension of this theoretical result to arbitrary $K \geq 3$ is left as an open question.

For arbitrary K we use the empirical approach of Biswas, Jacob, and Vanetti (2019), that provides estimated upper bounds on the total variation distance (TV) between $\mathbf{u}^{(t)}$ at iteration t and $\nu_{\mathbf{x}}$. These upper bounds are obtained as empirical averages over independent runs of coupled Markov chains; see Appendix D in the supplementary materials for a brief description of the approach. For $K = 4, 8, 12, 16$, we construct synthetic datasets with 10 observations in each category and estimate upper bounds for a range of t shown in Figure 5(a). The number of iterations required for convergence seems to be stable in K. Next, we set $K = 5$ and consider $10, 20, 30, 40$ counts in each category, leading to N varying between 50 and 200. Figure 5(b) shows the associated upper bounds, that increase with N.

4. Adding Categories, Observations, and Priors

4.1. Adding Empty Categories

We describe how to add and remove empty categories based on the output of the Gibbs sampler. Suppose that we have draws $\mathbf{u}$ distributed according to the target $\nu_{\mathbf{x}}$ associated with a dataset $\mathbf{x} \in [K]^N$ with K non-empty categories. We add a category $K+1$ with $\mathcal{I}_{K+1} = \emptyset$, $N_{K+1} = 0$, and consider how to obtain samples $\mathbf{u}'$ from the corresponding target $\nu_{\mathbf{x}'}$. Recall that a variable $(u_1, \ldots, u_K)$ following Dirichlet$(1, \ldots, 1)$ is equal in distribution to the vector with ℓth entry $w_\ell / \sum_{j\in[K]} w_j$ for $\ell \in [K]$, where $(w_\ell)_{\ell\in[K]}$ are independent Exponential(1). Given $(u_1, \ldots, u_K) \sim \Delta$ consider the following procedure. First, draw $s \sim$ Gamma$(K, 1)$, define $w_\ell = s \times u_\ell$ for $\ell \in [K]$, and draw $w_{K+1} \sim$ Exponential(1). Then define $u'_\ell = w_\ell / \sum_{j\in[K+1]} w_j$ for $\ell \in [K+1]$. The resulting vector $\mathbf{u}' = (u'_1, \ldots, u'_{K+1})$ is uniformly distributed on the probability simplex with $K+1$ vertices denoted by Δ'. Since $u'_\ell/u'_k = u_\ell/u_k$ for all $k, \ell \in [K]$, if $(u_1, \ldots, u_K)$ satisfies certain constraints on ratios u_ℓ/u_k, the same constraints are satisfied for $(u'_1, \ldots, u'_{K+1})$. Thus, $\mathbf{u}' \sim \nu_{\mathbf{x}'}$.

We can also remove empty categories. Assume that category $K+1$ is empty and that we have draws $\mathbf{u}' \sim \nu_{\mathbf{x}'}$. For each u'_n, drop the $(K+1)$th component $u'_{n,K+1}$, and define u_n by normalizing the remaining K components. The resulting $\mathbf{u}$ follows $\nu_{\mathbf{x}}$. Importantly, inferences obtained from $\nu_{\mathbf{x}}$ are not necessarily identical to those obtained from $\nu_{\mathbf{x}'}$. This is illustrated with

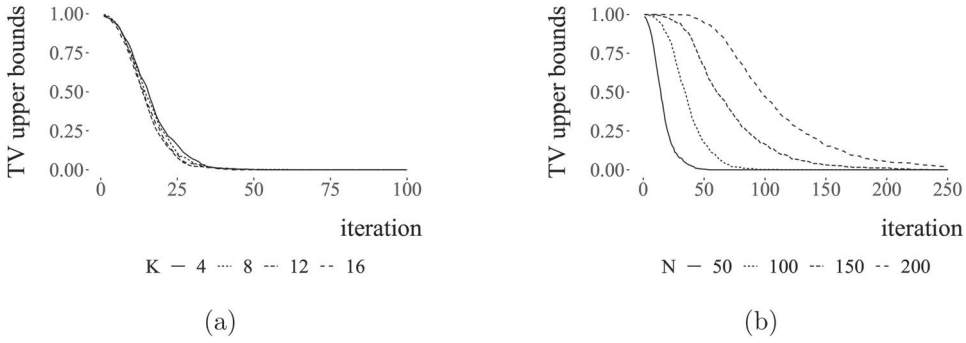

Figure 5. Upper bounds on the TV distance between $\mathbf{u}^{(t)}$ and $\nu_{\mathbf{x}}$ against t. (a) Varying K with 10 counts in each category. (b) Varying N with $K = 5$ and N/K counts in each category.

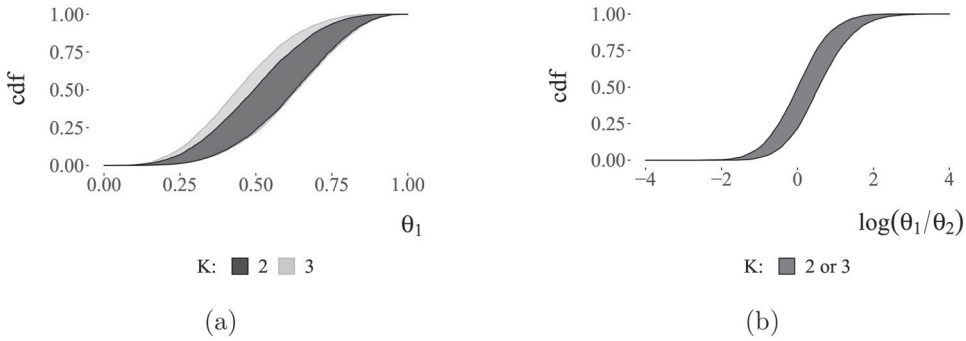

Figure 6. Inference on θ_1 (a) and on $\log(\theta_1/\theta_2)$ (b) using counts $(4, 3)$ ($K = 2$) and $(4, 3, 0)$ ($K = 3$). Including an empty third category modifies the inference on θ_1 but not on θ_1/θ_2.

Figure 6, showing the $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ probabilities associated with the sets $\{\theta\colon \theta_1 \in [0, c)\}$ and $\{\theta\colon \log\theta_1/\theta_2 \in (-\infty, c)\}$, for the counts $(4, 3)$ and $(4, 3, 0)$.

4.2. Adding Partial Prior Information

In the DS framework, multiple sources of information can be merged using Dempster's rule of combination (Dempster 1967, sec. 5; Wasserman 1990, sec. 2). If two sources yield random sets $\mathcal{F}$ and $\mathcal{G}$ the combination is obtained by intersections $\mathcal{F} \cap \mathcal{G}$, under an independent coupling of $\mathcal{F}$ and $\mathcal{G}$ conditional on $\mathcal{F} \cap \mathcal{G} \neq \emptyset$. The rule of combination can be used to incorporate prior knowledge. If the prior is encoded as a probability distribution on $\theta \in \Delta$, we can view each prior draw as a singleton $\mathcal{G}$, thus intersections $\mathcal{F} \cap \mathcal{G}$ are either singletons or empty. It can be checked that the non-empty $\mathcal{F} \cap \mathcal{G}$ are equivalent to draws from the posterior by noting that, for a given $\theta \in \Delta$,

$$\begin{aligned}\nu_{\mathbf{x}}(\{\mathbf{u}\colon \theta \in \mathcal{F}(\mathbf{u})\}) &= \frac{\text{uniform}\left(\{(u_1, \ldots, u_N) \in \Delta^N : \theta \in \mathcal{F}(\mathbf{u})\}\right)}{\text{uniform}\left(\{(u_1, \ldots, u_N) \in \Delta^N : \mathcal{F}(\mathbf{u}) \neq \emptyset\}\right)} \\ &\propto \theta_1^{N_1} \ldots \theta_K^{N_K},\end{aligned}$$

which is proportional to the multinomial likelihood associated with θ and $(N_1, \ldots, N_K)$ (Dempster 1972). This justifies why DS can be seen as a generalization of Bayesian inference. In $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ for an assertion $\Sigma \in \mathcal{B}(\Delta)$ this leads to $\mathtt{p} = \mathbb{P}(\theta \in \Sigma|\mathbf{x})$, the posterior mass of Σ, $\mathtt{q} = 1 - \mathtt{p}$ and $\mathtt{r} = 0$.

The DS framework allows the inclusion of partial prior information. We follow the above reasoning except that the prior is formulated as random sets that are not necessarily singletons. For example, we can specify a prior on some components of θ and extend these into random subsets of Δ by "up-projection" (Dempster 2008) or "minimal extension" (Wasserman 1990, sec. 2.5). Concretely suppose that we observe counts (N_1, N_2) of two categories. We specify a Dirichlet prior on (θ_1, θ_2) and obtain a Dirichlet posterior. Next we are told that there exists in fact a third category, which we could not observe before. This is different than being told that there is a new category with zero counts, $N_3 = 0$, which we could handle as in Section 4.1. Up-projection of each posterior draw (θ_1, θ_2) onto the 3-simplex Δ goes as follows. We compute $\eta_{1\to 2} = \theta_2/\theta_1$ and $\eta_{2\to 1} = \theta_1/\theta_2$, and set $\eta_{3\to k} = \eta_{k\to 3} = +\infty$ for $k = 1, 2$. Denote by $\mathcal{F}$ the resulting feasible sets $\{\theta \in \Delta\colon \theta_\ell/\theta_k \leq \eta_{k\to\ell}\ \forall k, \ell\}$. These sets $\mathcal{F}$ correspond to a "minimal extension" in that inference on θ_1/θ_2 is unchanged, while inference on θ_3 is vacuous. Vacuous means that for any assertion $\Sigma = \{\theta \in \Delta\colon \theta_3 \in A\}$ with $A \subset [0, 1]$, the sets $\mathcal{F}$ result in $\mathtt{p} = 0, \mathtt{q} = 0, \mathtt{r} = 1$. Using the rule of combination we can subsequently intersect such sets $\mathcal{F}$ with independent random sets corresponding to new observations of the three categories. Visuals are provided in Figure 7, with Figure 7(b) showing random sets corresponding to counts of three categories using a partial Dirichlet(2,2) prior on (θ_1, θ_2).

4.3. Adding Observations

We consider the addition of new observations to existing categories. Denote by $\mathbf{x}_{N+1}$ the original data $\mathbf{x}_N$ augmented with an observation x_{N+1}, which we assume equal to $k \in [K]$.

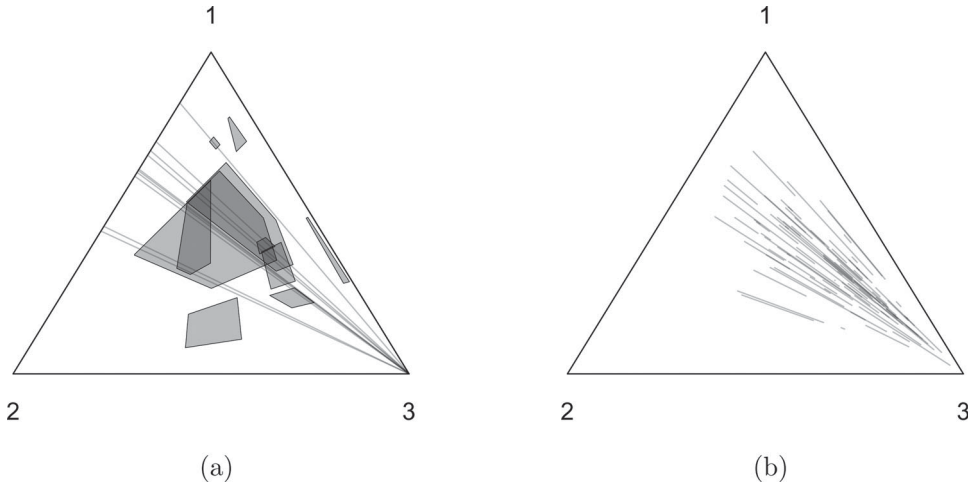

Figure 7. Up-projection of posterior samples (θ_1, θ_2), obtained from $(N_1 = 8, N_2 = 4)$ and a Dirichlet(2,2) prior (segments in (a)), and feasible sets obtained independently for counts (2, 1, 3) (polygons in (a)). The rule of combination retains nonempty intersections of these sets (b).

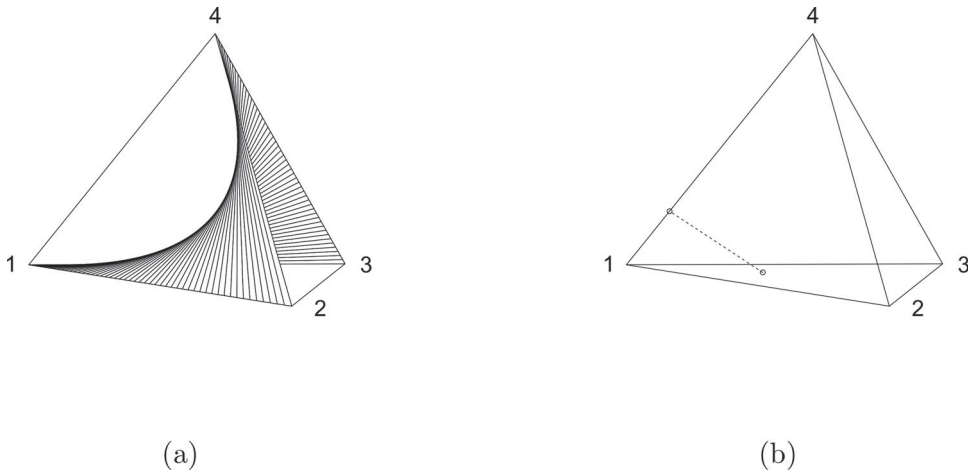

Figure 8. Two surfaces in the 4-simplex. (a) The independence surface $\theta_1\theta_4 = \theta_2\theta_3$ (Fienberg and Gilbert 1970). (b) The linkage constraint of (13) for $\phi \in (0, 1)$ as a dashed segment.

Any $u_{1:N+1} \in \mathcal{R}_{\mathbf{x}_{N+1}}$ is such that $u_{1:N} \in \mathcal{R}_{\mathbf{x}_N}$ and $u_{N+1} \in \Delta_k(\theta)$, with $\theta \in \Delta$ constructed from $u_{1:N}$ as in Proposition 3.2. Indeed if $u_{1:N+1} = (u_1, \ldots, u_{N+1}) \in \mathcal{R}_{\mathbf{x}_{N+1}}$, there exists $\theta' \in \Delta$ such that, for all $n \in [N+1]$, $u_{n,\ell}/u_{n,k} \geq \theta'_\ell/\theta'_k$. Thus, $u_{1:N} \in \mathcal{R}_{\mathbf{x}_N}$. We can check that u_{N+1} is in $\Delta_k(\theta)$. Since $u_{1:N+1} \in \mathcal{R}_{\mathbf{x}_{N+1}}$, then $(u_n)_{n\in\mathcal{I}_k}$ belongs to the support of $\nu_{\mathbf{x}_{N+1}}(d\mathbf{u}_{\mathcal{I}_k}|\mathbf{u}_{[N+1]\setminus\mathcal{I}_k})$, which is $\Delta_k(\theta)^{N_k}$ by Proposition 3.2. Here we have redefined $\mathcal{I}_k = \{n \in [N+1] : x_n = k\}$. Conversely, if $u_{1:N} \in \mathcal{R}_{\mathbf{x}_N}$ and $u_{N+1} \in \Delta_k(\theta)$ then $u_{1:N+1} \in \mathcal{R}_{\mathbf{x}_{N+1}}$; again because $\Delta_k(\theta)$ is precisely the support of $\nu_{\mathbf{x}_{N+1}}(du_{N+1}|\mathbf{u}_{[N+1]\setminus\mathcal{I}_k})$.

This motivates an importance sampling strategy. For $u_{1:N} \sim \nu_{\mathbf{x}_N}$, generate $u_{N+1} \sim \Delta_k(\theta)$, with $\theta \in \Delta$ as above. Denote this distribution by $q_{N+1}(du_{N+1}|u_{1:N})$. The density $u_{N+1} \mapsto q_{N+1}(u_{N+1}|u_{1:N})$ equals $(\theta_k)^{-1}$ for $u_{N+1} \in \Delta_k(\theta)$, since the volume of $\Delta_k(\theta)$ is θ_k. We can correct for the discrepancy between proposal and target by computing weights

$$w_{N+1}(u_{1:N+1}) = \frac{\nu_{\mathbf{x}_{N+1}}(u_{1:N+1})}{\nu_{\mathbf{x}_N}(u_{1:N})q_{N+1}(u_{N+1}|u_{1:N})} = \frac{Z_N}{Z_{N+1}}\mathrm{Vol}(\Delta_k(\theta)),$$

where Z_N is the volume of $\mathcal{R}_{\mathbf{x}_N}$. We can thus implement self-normalized importance sampling. The reasoning can be extended to assimilate observations recursively with a sequential Monte Carlo sampler (Del Moral, Doucet, and Jasra 2006), alternating importance sampling and Gibbs moves. This strategy will be employed in Section 5.1.

5. Applications

We present two applications. In both examples, the $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ probabilities require distributional information about the entire random polytopes, and not only the extreme vertices elicited in Dempster (1972). Both examples involve $K = 4$ categories and curves in the simplex shown in Figure 8. Our main objective is to illustrate the output of the algorithm. We briefly recall from Section 2 that the inferred $\mathtt{p}$ and $\mathtt{q}$ probabilities can be understood as the degree of evidential support "for" or "against" the hypothesis of interest based on available observations and the model specification. The $\mathtt{r}$ probability, which is a distinctive feature of DS compared to standard Bayes, indicates a degree of epistemological indeterminacy, with a larger value encouraging

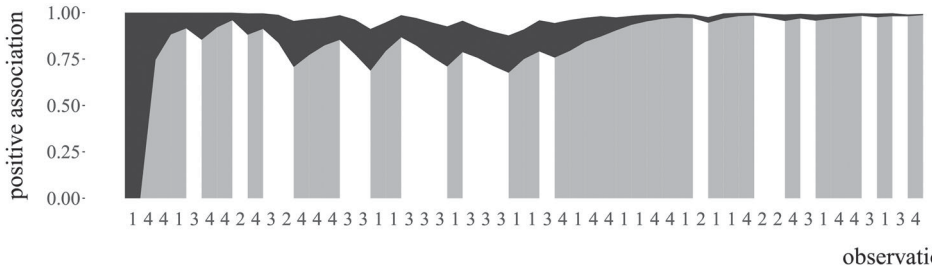

Figure 9. Support for the hypothesis of positive association $H_+ : \theta_1\theta_4 \geq \theta_2\theta_3$ as observations in $\{1, 2, 3, 4\}$ are incorporated one by one. The dark ribbon delineates the probability $\mathfrak{p}$ for H_+, and one minus the support against it, respectively, as its lower and upper rims. The width of the ribbon represents the amount of "don't know" about the hypothesis.

the analyst to suspend judgment about the assertion of interest. The $\mathfrak{r}$ probability can be useful to make decisions or to postpone them, as with other types of imprecise probabilities and robust Bayesian analysis (Berger et al. 1994). The way that decisions can be informed by DS uncertainties has received some attention, for example, see Section 12 of Shafer (1990), and also Yager (1992) and Bauer (1997).

5.1. Testing Independence

In the case of $K = 4$, count data may be arranged in a 2×2 table with proportions $(\theta_1, \theta_2, \theta_3, \theta_4)$ as cell probabilities, row by row. We may be interested in testing independence, $H_0 : \theta_1\theta_4 = \theta_2\theta_3$, see Wasserman (2013, chap. 15). Classic tests include the Pearson's chi-squared test with $\chi^2 = \sum_{i,j} (x_{ij} - e_{ij})^2/e_{ij}$, where e_{ij} is the expected number of counts in cell "ij" under H_0. The Pearson test statistic is asymptotically χ^2_1. The likelihood ratio test with statistic $G^2 = 2\sum_{i,j} x_{ij} \log(x_{ij}/e_{ij})$, is asymptotically equivalent; see Diaconis and Efron (1985) for further interpretations.

Evaluating the posterior probability of H_0 raises the issue that the set $\{\theta \in \Delta : \theta_1\theta_4 = \theta_2\theta_3\}$, a surface in the 4-simplex as depicted in Figure 8(a), might be of zero measure under the posterior. As a remedy one can employ Bayes factors (e.g., Albert and Gupta 1983), or we can consider the evidence toward either positive or negative association, that is, $H_+ : \theta_1\theta_4 \geq \theta_2\theta_3$ or $H_- : \theta_1\theta_4 \leq \theta_2\theta_3$, and interpret such evidence as being against independence.

We consider the dataset presented in Rosenbaum (2002, p. 191) regarding the effect of drainage pits on incident survival in the London underground. Some stations are equipped with drainage pits below the tracks. Passengers who accidentally fall off the platform may seek refuge in the pit to avoid an incoming train. For stations without a pit, only 5 lived out of 21 recorded incidents. In the presence of a pit, 18 out of 32 lived. Ding and Miratrix (2019) reanalyzed the data to assess the difference in mortality rates. Their analysis suggests that the existence of a pit significantly increases the chance of survival. The data can be summarized as counts $(16, 5, 14, 18)$. Pearson's chi-squared test statistic is $\chi^2 = 5.43$ with a p-value of 0.02, while the likelihood ratio test yields a p-value of 0.017. The Bayesian analysis shows strong evidence for positive association, with posterior probabilities $P(H_+ \mid \mathbf{x}) = 0.99$ and $P(H_- \mid \mathbf{x}) = 0.01$.

The DS approach applied sequentially yields the results shown in Figure 9. The horizontal axis shows the observations, in an arbitrary order. The dark ribbon tracks $\mathfrak{p}(H_+)$ and $(1 - \mathfrak{q}(H_+))$ by its lower and upper rims, respectively. The "don't know" probability $\mathfrak{r}(H_+)$, represented by the width of the ribbon, can be seen to progressively shrink, but not systematically. The support for H_+ increases with each observation in $\{1, 4\}$ and decreases with each observation in $\{2, 3\}$ (as highlighted with background shades). Figure 9 is inspired by Figure 4 of Walley, Gurrin, and Burton (1996). In DS inference, the width of the ribbon is part of the inference and could be used, for example, to inform decisions about the collection of additional data.

5.2. Linkage Model

The linkage model from Rao (1973, pp. 368–369) was considered by Lawrence et al. (2009), as an example illustrating inference with an additional constraint. They compare the IDM of Walley (1996) and their method termed Dirichlet DSM (for Dempster–Shafer model). The data consist of $N = 197$ counts over $K = 4$ categories, with probabilities satisfying

$$\theta(\phi) = \left(\frac{1}{2} + \frac{\phi}{4}, \frac{1-\phi}{4}, \frac{1-\phi}{4}, \frac{\phi}{4}\right), \tag{13}$$

for some $\phi \in (0, 1)$. In other words, $\theta(\phi) = A\phi + b$ for appropriately defined 4×1 matrices A and b, as shown in Figure 8(b). The original observations were $(125, 18, 20, 34)$, but Lawrence et al. (2009) considered the counts $(25, 3, 4, 7)$, which results in a more visible amount of "don't know" probability.

We briefly introduce Dirichlet DSM and focus on the comparison between the approaches. They differ by the choice of sampling mechanism: instead of using the mechanism described in Dempster (1966), Lawrence et al. (2009) introduced another mechanism to make inference simpler computationally. For a vector of counts $(N_1, \ldots, N_K)$, the Dirichlet DSM model expresses its posterior inference for the proportion vector θ via the random feasible set $\{\theta \in \Delta : \theta_1 \geq z_1, \ldots, \theta_K \geq z_K\}$, where $\mathbf{z} = (z_0, z_1, \ldots, z_K) \sim \text{Dirichlet}_{K+1}(1, N_1, \ldots, N_K)$. Incorporating the parameter constraint $\theta = A\phi + b$, the feasible set for ϕ is $[\phi_{\min}(\mathbf{z}), \phi_{\max}(\mathbf{z})]$ with

$$\begin{aligned} \phi_{\min}(\mathbf{z}) &\equiv \max(4z_1 - 2, 4z_4) \\ &\leq \phi_{\max}(\mathbf{z}) \equiv \min(1 - 4z_2, 1 - 4z_3). \end{aligned} \tag{14}$$

For the approach of Dempster (1966), termed "simplex-DSM" in Lawrence et al. (2009), we first run the proposed Gibbs sampler without taking into account the linear constraint (13).

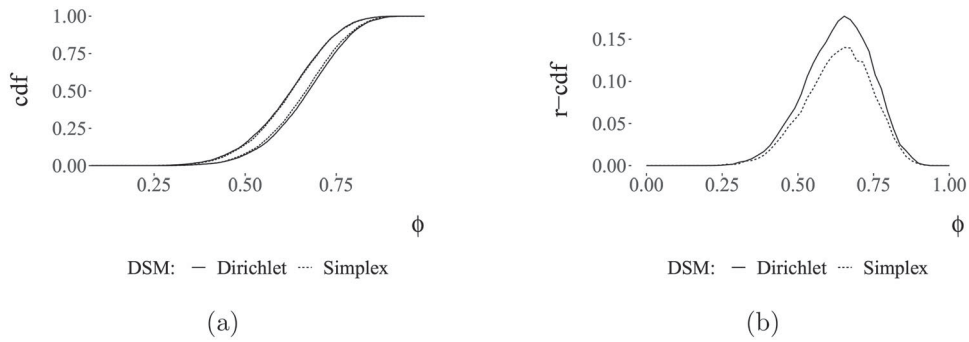

Figure 10. "Dirichlet-DSM" approach of Lawrence et al. (2009) and the original approach of Dempster (1966), for the linkage model with data $(25, 3, 4, 7)$, the latter implemented with the proposed Gibbs sampler. Here (a) shows the lower and upper probabilities for assertions of the form $\{\phi < c\}$ for increasing $c \in [0, 1]$, while (b) depicts the difference between these upper and lower probabilities, or equivalently the $\mathtt{r}$ values.

Among the generated feasible sets, only those that intersect with the linear constraint are retained, and an interval $[\underline{\phi}, \overline{\phi}]$ is obtained for each such set, where

$$\underline{\phi} = \text{argmin}_{\phi} \{\theta(\phi) \in \mathcal{F}(\mathbf{u})\},$$
$$\overline{\phi} = \text{argmax}_{\phi} \{\theta(\phi) \in \mathcal{F}(\mathbf{u})\}.$$

For the data considered here, this retains 5% of the iterations, and is therefore a practical solution. However, the approach would become impractical if the counts were much less "compatible" with the linkage constraint, in which case novel computational methods would be necessary. We estimate $(\mathtt{p}, \mathtt{q}, \mathtt{r})$ for sets $\{\phi \in [0, c)\}$ for $c \in (0, 1)$, that is, lower and upper cumulative distribution functions, under both approaches and represent them in Figure 10(a). The plot shows the overall agreement between the two approaches. Figure 10(b) highlights the difference in $\mathtt{r}$ values, and illustrates that multiple approaches within the DS framework lead to different results.

6. Discussion

The discipline of statistics does not have a single framework for parameter inference. The setting of count data is rich enough to contrast various approaches. Before any other considerations, for a framework to be useful to scientists and decision-makers, the ability to perform the associated computation is essential, and allows for grounded discussions and concrete comparisons. Our work helps with the computation in the DS framework for categorical distributions, which will hopefully motivate further theoretical investigations of its statistical features.

One of the appeals of the DS framework is its flexibility to incorporate types of partial information which are difficult to express in the Bayesian framework. This includes vacuous or partial priors, coarse data which arise from imprecise measurement devices and imperfect surveys. These elements can be represented as random sets (Nguyen 2006; Plass et al. 2015) in the DS framework while circumventing assumptions about the coarsening mechanism (e.g., Heitjan and Rubin 1991).

Whether a perfect sampler could be devised as an alternative to the proposed Gibbs sampler is an open question. Generic algorithms for uniform sampling on polytopes (Vempala 2005; Narayanan 2016; Chen et al. 2018) could also provide competitive results. The proposed Gibbs sampler could itself be accelerated, for instance, by using warm starts in the linear program solvers over subsequent iterations.

The typical challenge of DS computations is the generation of nonempty intersections of random sets. The proposed Gibbs sampler can be seen as a way of avoiding inefficient rejection samplers in the setting of inference in categorical distributions. It remains to see whether similar ideas can be used to deploy the DS framework in other models, for example, to avoid rejection sampling in the linkage model, or in hidden Markov models and models with moment constraints (Chamberlain and Imbens 2003; Bornn, Shephard, and Solgi 2019), which are natural extensions of the categorical distribution.

Supplementary Materials

The supplementary materials describe the choice of sampling mechanism, its effect on statistical inference and its relation with the Gumbel-max trick. They also describe the convergence rate of the algorithm in a simple case and provide reminders on empirical convergence analysis using coupled Markov chains.

Acknowledgments

The authors thank Rahul Mazumder for useful advice on linear programming.

Funding

The authors gratefully acknowledge support from the National Science Foundation (DMS-1712872, DMS-1844695, DMS-1916002), and the National Institute of Allergy and Infectious Disease at the National Institutes of Health [2 R37 AI054165-11 and 75N93019C00070]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Albert, J. H., and Gupta, A. K. (1983), "Bayesian Estimation Methods for 2 × 2 Contingency Tables Using Mixtures of Dirichlet Distributions," *Journal of the American Statistical Association*, 78, 708–717. [9]

Avis, D., and Fukuda, K. (1992), "A Pivoting Algorithm for Convex Hulls and Vertex Enumeration of Arrangements and Polyhedra," *Discrete & Computational Geometry*, 8, 295–313. [6]

Bang-Jensen, J., and Gutin, G. Z. (2008), *Digraphs: Theory, Algorithms and Applications*, London: Springer. [4]

Basir, O., and Yuan, X. (2007), "Engine Fault Diagnosis Based on Multi-Sensor Information Fusion Using Dempster–Shafer Evidence Theory," *Information Fusion*, 8, 379–386. [1]

Bauer, M. (1997), "Approximation Algorithms and Decision Making in the Dempster–Shafer Theory of Evidence—An Empirical Study," *International Journal of Approximate Reasoning*, 17, 217–237. [9]

Berger, J. O., and Bernardo, J. M. (1992), "Ordered Group Reference Priors With Application to the Multinomial Problem," *Biometrika*, 79, 25–37. [1]

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D. and Betrò, B. (1994), "An Overview of Robust Bayesian Analysis," *Test*, 3, 5–124. [9]

Berkelaar, M., Eikland, K., and Notebaert, P. (2004), "lpsolve: Open Source (Mixed-Integer) Linear Programming System," Eindhoven University of Technology, p. 63. [6]

Bernard, J.-M. (1998), "Bayesian Inference for Categorized Data," in Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (eds.), *New Ways in Statistical Methodology*, Bern: Peter Lang, pp. 159–226. [1]

Biswas, N., Jacob, P. E., and Vanetti, P. (2019), "Estimating Convergence of Markov Chains With L-Lag Couplings," in *Advances in Neural Information Processing Systems*, pp. 7389–7399. [6]

Bloch, I. (1996), "Some Aspects of Dempster–Shafer Evidence Theory for Classification of Multi-Modality Medical Images Taking Partial Volume Effect Into Account," *Pattern Recognition Letters*, 17, 905–919. [1]

Bornn, L., Shephard, N., and Solgi, R. (2019), "Moment Conditions and Bayesian Non-Parametrics," *Journal of the Royal Statistical Society*, Series B, 81, 5–43. [10]

Chafai, D., and Concordet, D. (2009), "Confidence Regions for the Multinomial Parameter With Small Sample Size," *Journal of the American Statistical Association*, 104, 1071–1079. [1]

Chamberlain, G., and Imbens, G. W. (2003), "Nonparametric Applications of Bayesian Inference," *Journal of Business & Economic Statistics*, 21, 12–18. [10]

Chen, Y., Dwivedi, R., Wainwright, M. J., and Yu, B. (2018), "Fast MCMC Sampling Algorithms on Polytopes," *The Journal of Machine Learning Research*, 19, 2146–2231. [10]

Csardi, G. and Nepusz, T. (2006), "The igraph Software Package for Complex Network Research," *InterJournal, Complex Systems*, 1695, 1–9, available at http://igraph.org. [5]

Dawid, A. P., and Stone, M. (1982), "The Functional-Model Basis of Fiducial Inference," *The Annals of Statistics*, 10, 1054–1067. [2]

Del Moral, P., Doucet, A., and Jasra, A. (2006), "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society*, Series B, 68, 411–436. [8]

Dempster, A. P. (1963), "On Direct Probabilities," *Journal of the Royal Statistical Society*, Series B, 25, 100–110. [1]

——— (1966), "New Methods for Reasoning Towards Posterior Distributions Based on Sample Data," *The Annals of Mathematical Statistics*, 37, 355–374. [1,2,3,4,9,10]

——— (1967), "Upper and Lower Probabilities Induced by a Multivalued Mapping," *The Annals of Mathematical Statistics*, 38, 325–339. [1,3,7]

——— (1968), "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society*, Series B, 30, 205–247. [1,3]

——— (1972), "A Class of Random Convex Polytopes," *The Annals of Mathematical Statistics*, 43, 260–272. [1,2,3,7,8]

——— (2008), "The Dempster–Shafer Calculus for Statisticians," *International Journal of Approximate Reasoning*, 48, 365–377. [1,3,7]

——— (2014), "Statistical Inference From a Dempster–Shafer Perspective," in *Past, Present, and Future of Statistical Science*, eds. X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J.-L. Wang, Boca Raton, FL: Chapman and Hall/CRC, pp. 275–288. [1]

Denoeux, T. (2000), "A Neural Network Classifier Based on Dempster–Shafer Theory," *IEEE Transactions on Systems, Man, and Cybernetics A: Systems and Humans*, 30, 131–150. [1]

——— (2006), "Constructing Belief Functions From Sample Data Using Multinomial Confidence Regions," *International Journal of Approximate Reasoning*, 42, 228–252. [1]

——— (2008), "A k-Nearest Neighbor Classification Rule Based on Dempster–Shafer Theory," in *Classic Works of the Dempster-Shafer Theory of Belief Functions*, eds. R. R. Yager and L. Liu, Berlin, Heidelberg: Springer, pp. 737–760. [1]

Diaconis, P., and Efron, B. (1985), "Testing for Independence in a Two-Way Table: New Interpretations of the Chi-Square Statistic," *The Annals of Statistics*, 13, 845–874. [9]

Díaz-Más, L., Muñoz-Salinas, R., Madrid-Cuevas, F. J., and Medina-Carnicer, R. (2010), "Shape From Silhouette Using Dempster–Shafer Theory," *Pattern Recognition*, 43, 2119–2131. [1]

Ding, P., and Miratrix, L. W. (2019), "Model-Free Causal Inference of Binary Experimental Data," *Scandinavian Journal of Statistics*, 46, 200–214. [9]

Dunson, D. B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051. [1]

Edlefsen, P. T., Liu, C., and Dempster, A. P. (2009), "Estimating Limits From Poisson Counting Data Using Dempster–Shafer Analysis," *The Annals of Applied Statistics*, 3, 764–790. [2]

Fienberg, S. E., and Gilbert, J. P. (1970), "The Geometry of a Two by Two Contingency Table," *Journal of the American Statistical Association*, 65, 694–701. [8]

Fisher, R. A. (1935), "The Fiducial Argument in Statistical Inference," *Annals of Eugenics*, 6, 391–398. [2]

Fitzpatrick, S., and Scott, A. (1987), "Quick Simultaneous Confidence Intervals for Multinomial Proportions," *Journal of the American Statistical Association*, 82, 875–878. [1]

Fraser, D. A. S. (1968), *The Structure of Inference*, New York: Wiley. [2]

Fukuda, K. (1997), "cdd/cdd+ Reference Manual," Institute for Operations Research, ETH-Zentrum, pp. 91–111. [6]

Geyer, C. J., and Meeden, G. D. (2008), "R Package rcdd (C Double Description for R), Version 1.1." [6]

Gibbs, A. L. (2004), "Convergence in the Wasserstein Metric for Markov Chain Monte Carlo Algorithms With Applications to Image Restoration," *Stochastic Models*, 20, 473–492. [6]

Gong, R. (2018), "Low-Resolution Statistical Modeling With Belief Functions," PhD thesis, Harvard University. [2]

——— (2019), "Simultaneous Inference Under the Vacuous Orientation Assumption," *Proceedings of Machine Learning Research*, 103, 225–234. [3]

Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016), "Generalized Fiducial Inference: A Review and New Results," *Journal of the American Statistical Association*, 111, 1346–1361. [1,2]

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244–2253. [10]

Jerrum, M. (1998), "Mathematical Foundations of the Markov Chain Monte Carlo Method," in *Probabilistic Methods for Algorithmic Discrete Mathematics*, eds. M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, Berlin, Heidelberg: Springer, pp. 116–165. [6]

Konis, K. (2014), "lpSolveAPI: R Interface for lpSolve Version 5.5," R Package Version 5.2-0. [6]

Lang, J. B. (2004), "Multinomial-Poisson Homogeneous Models for Contingency Tables," *The Annals of Statistics*, 32, 340–383. [1]

Lawrence, E. C., Vander Wiel, S., Liu, C., and Zhang, J. (2009), "A New Method for Multinomial Inference Using Dempster-Shafer Theory," Technical Report, Los Alamos National Lab, Los Alamos, New Mexico, United States. [1,2,9,10]

Liu, C. (2000), "Estimation of Discrete Distributions With a Class of Simplex Constraints," *Journal of the American Statistical Association*, 95, 109–120. [1]

Liu, Y., and Hannig, J. (2016), "Generalized Fiducial Inference for Binary Logistic Item Response Models," *Psychometrika*, 81, 290–324. [1]

Maddison, C. J., Tarlow, D., and Minka, T. (2014), "A* Sampling," in *Advances in Neural Information Processing Systems*, pp. 3086–3094. [2]

Narayanan, H. (2016), "Randomized Interior Point Methods for Sampling and Optimization," *The Annals of Applied Probability*, 26, 597–641. [10]

Nguyen, H. T. (2006), *An Introduction to Random Sets*, Boca Raton, FL: CRC Press. [10]

Plass, J., Augustin, T., Cattaneo, M., and Schollmeyer, G. (2015), "Statistical Modelling Under Epistemic Data Imprecision: Some Results on Estimating Multinomial Distributions and Logistic Regression for Coarse Categorical Data," in *ISIPTA* (Vol. 15), pp. 247–256. [10]

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (Vol. 2), New York: Wiley. [9]

Roberts, G. O., and Rosenthal, J. S. (2004), "General State Space Markov Chains and MCMC Algorithms," *Probability Surveys*, 1, 20–71. [6]

Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer-Verlag. [9]

Shafer, G. (1976), *A Mathematical Theory of Evidence* (Vol. 42), Princeton, NJ: Princeton University Press. [1,3]

——— (1979), "Allocations of Probability," *The Annals of Probability*, 7, 827–839. [1]

——— (1990), "Perspectives on the Theory and Practice of Belief Functions," *International Journal of Approximate Reasoning*, 4, 323–362. [3,9]

Sison, C. P., and Glaz, J. (1995), "Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions," *Journal of the American Statistical Association*, 90, 366–369. [1]

Vasseur, P., Pégard, C., Mouaddib, E., and Delahoche, L. (1999), "Perceptual Organization Approach Based on Dempster–Shafer Theory," *Pattern Recognition*, 32, 1449–1462. [1]

Vempala, S. (2005), "Geometric Random Walks: A Survey," *Combinatorial and Computational Geometry*, 52, 2. [10]

Walley, P. (1996), "Inferences From Multinomial Data: Learning About a Bag of Marbles," *Journal of the Royal Statistical Society*, Series B, 58, 3–34. [1,9]

Walley, P., Gurrin, L., and Burton, P. (1996), "Analysis of Clinical Data Using Imprecise Prior Probabilities," *Journal of the Royal Statistical Society*, Series D, 45, 457–485. [9]

Wasserman, L. (1990), "Belief Functions and Statistical Inference," *Canadian Journal of Statistics*, 18, 183–196. [1,3,7]

——— (2013), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer. [9]

Yager, R. R. (1992), "Decision Making Under Dempster–Shafer Uncertainties," *International Journal of General System*, 20, 233–245. [9]