Journal of the American Statistical Association

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

Rejoinder—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) Rejoinder—A Gibbs Sampler for a Class of Random Convex Polytopes, Journal of the American Statistical Association, 116:535, 1211-1214, DOI: 10.1080/01621459.2021.1945458

To link to this article: https://doi.org/10.1080/01621459.2021.1945458

Published online: 07 Sep 2021.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 535, 1211–1214: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2021.1945458

Rejoinder—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[d]

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

We are very grateful to all commenters for their stimulating remarks, questions, as well as useful pointers to the literature which span a wide range of statistical methods over decades of research. We have neither the space nor the knowledge to answer many of the questions raised, and we only aim to offer some clarifications. We hope that readers will be as enthusiastic as ourselves about research on the topics discussed by the commenters. In the following, we refer to Diaconis and Wang as DW, Hoffman, Hannig and Zhang as HHZ, Lawrence and Vander Wiel as LV, Ruggeri as R, Shafer as S, and Williams as W.

1. Our Motivation

The Dempster-Shafer (DS) theory of belief functions was conceived in the 1960s in part as a response to the alluring, yet troubled, Fisherian proposal of fiducial inference (Dempster 1964, 1966). The DS theory stands as a logical framework for uncertain reasoning, employing belief functions and random sets as the basic elements of the *extended calculus of probability*. The theory allows for not only partial specifications of probabilistic knowledge, but also extension and combination operations that venture outside the classic likelihood-based paradigms. While an extensive literature review was not in the scope of the present article, interested readers may refer to the suggestions provided by the commenters, as well as Dempster (2008, 2014) and Cuzzolin (2017) for recent accounts on the subject.

As DW and S recalled, about a decade after its inception, the computational hurdle imposed by the DS theory became evident. Uniquely pertaining to statistical applications is the challenge that typical inference problems, whether parametric and nonparametric, have continuous state spaces. To suit these problems, marginal DS models constructed via multivalued mapping, Equation (3) of this article being an example, call for nondenumerable collections of mass-bearing random sets. While DS found successful implementations of the elegant local computation and propagation schemes (see, e.g., Kong 1986; Shenoy and Shafer 1986) in abundant discrete space problems of artificial intelligence, these tools were not directly transferable to DS models for statistical inference. The lack of computational feasibility became a main reason why the DS theory remained an esoteric interest in this community till this day.

W wrote that the field of statistics lacks a unifying foundation for parametric inference. We find it perfectly acceptable that the field would not have such a unifying foundation when contemplating the breadth of the task and the diversity of situations in which it arises. In any case, the need to go beyond the standard Bayesian framework is widely shared and echoed by scholars who explore alternative inferential paradigms, from structural and functional inference, to generalized fiducial inference, confidence distributions and inferential models, and to other forms of inference using imprecise probabilities. The setting of system reliability described by LV provides another excellent motivation for nonstandard approaches.

As R pointed out, the DS theory is often compared to robust Bayesian approaches, as both were motivated by a quest for a flexible alternative to classic Bayes that is capable of expressing partial, or weakened, knowledge about unknown states of the world. Both DS and robust Bayes use upper and lower probabilities as the semantics to convey uncertain inference. The added theoretical dexterity offers freedom to express a structural type of uncertainty, in addition to the familiar stochastic type. As connections between these approaches are sought (see, e.g., Wasserman 1990), we come to understand the similarities as well as the differences. The recent work of Gong and Meng (2021) discussed how the DS theory and robust Bayes appeal to their preferred updating rules to incorporate observed information. Worth noting is that robust Bayes, due to its objective to preserves coherence at all costs, cannot learn from the data if a fully vacuous prior is employed, and suffers from dilation (Seidenfeld and Wasserman 1993) more so than DS. On the other hand, the DS probability interval does not convey the interpretation as a bound for an unknown "true" probability, but rather a logical consequence resulting from the operational combination of personalist marginal evidence. Just like other paradigms, DS theory and robust Bayes each employ assumptions and imply consequences of their own; both should appeal to self-described Bayesians!

The motivation for the present article is to offer computational feasibility to a classic DS model for categorical distributions. The hope is that our endeavors will motivate theoreticians to pry into its inner workings, and that our efforts provide a step toward a distinct tool for practitioners, to tackle settings

CONTACT Pierre E. Jacob pjacob@g.harvard.edu Department of Statistics, Harvard University, 1 Oxford Street, Science Center 7th Floor, Cambridge, MA 02138.
*Department of Information Systems, Decision Sciences and Statistics, ESSEC Business School, Paris, France.

such as that described by LV and other situations where the quantification of multiple sources of uncertainty is paramount.

2. Relevance of This Approach

As clarified by S, the proposed algorithm implements the DS approach for a specific "structure of the second kind" for Categorical distributions called "simplicial." Both LV and HHZ recalled favorable properties satisfied by another DS method termed "Dirichlet DSM," proposed in Lawrence et al. (2009). Meanwhile, S recalled a property satisfied by the simplicial model only. The experiments in HHZ show that the amount of "don't know" obtained with the simplicial method in the setting of testing uniformity can be much larger than with "Dirichlet DSM." On the other hand, Figure 10(b) of this article suggests that the simplicial model leads to smaller amounts of ignorance regarding parameter inference in the linkage model. S recalled that alternative approaches within the DS framework, have been compared and discussed in the 1960s, for example, in the discussion of Dempster (1968). We might either look for another approach, that would resolve all disturbing aspects of current ones, or be content with the variety of imperfect but available approaches. We certainly agree with HHZ on the numerous appeals of the Dirichlet DSM approach of Lawrence et al. (2009).

As pointed out by W, some aspects of the proposed computational method are very specific to the target distribution under consideration. Indeed, we can only speculate that some of the underlying ideas will inspire algorithms that will tackle significantly different situations. For example, our method instantiates large numbers of auxiliary variables in order to make calculations more tractable on an extended state space; it is directly inspired by the pseudo-marginal method (Andrieu and Roberts 2009) and particle MCMC method (Andrieu, Doucet, and Holenstein 2010), which themselves arose from ingenious work in genealogical inference (Beaumont 2003). We do not claim that the present work will be remotely as influential, but we believe in the value of addressing specific problems in original ways, in parallel to the delineation of general principles.

3. Scalability of the Proposed Algorithm

Some comments concern the computational efficiency of the proposed method. In the experiments of the article, we consider counts summing up to a few thousands, and numbers of categories up to 16. In particular, R questioned the choice of number of MCMC iterations, and HHZ reported long overall run times for moderate numbers of categories, such as 36.

First, it is only necessary to perform the MCMC operations using the nonzero counts. In Section 4.1 we describe how categories with zero counts can be added back, in a postprocessing step. Specifically, we can add M empty categories to a K-dimensional draw $\mathbf{u}$ from $\nu_{\mathbf{x}}$ by sampling a Gamma$(K, 1)$ multiplier for the existing K components, M additional Exponential(1) variables and normalizing the resulting $K+M$ components; this costs of the order of $M+K$ operations, and computing the associated "η" variables costs $K \times (M+K)$ operations, since $\eta_{k\to j} = +\infty$ for each empty category k and $j \neq k$.

Second, we describe in Section 3.4 and Appendix D how the coupling method of Biswas, Jacob, and Vanetti (2019) can be used to obtain practical guidance on the number of MCMC iterations. This often provides support for surprisingly short runs, as shown in Figure 5. Furthermore, the comment of DW suggests that our simple convergence analysis in the case $K = 2$ is accurate. The intriguing connection to the "donkey walk," pointed out by DW and new to us, provides a promising path to an analytical study in the general case $K \geq 3$.

Third, we understand that the long run times reported by HHZ originate from the effort required to enumerate the vertices of the generated polytopes. Thankfully, in many situations vertex enumeration is not a necessary step. The "half-space" representation is enough for the implementation of procedures that scale more favorably with the number of categories. Indeed, recall that the polytope $\mathcal{F}(\mathbf{u})$ can be represented as $\{\theta \in \mathbb{R}^K : A\theta \leq b\}$, where the inequalities hold component-wise. The inequalities include $K+2$ constraints to enforce $\theta \in \Delta$, and $K(K-1)$ constraints of the form $\theta_\ell - \eta_{k\to\ell}\theta_k \leq 0$ for $k, \ell \in [K]$. Thus, the matrix $(\eta_{k\to\ell})_{k,\ell}$ fully specifies the polytope of interest. In passing we note that, with the change of variable $t_k = \log\theta_k - K^{-1}\sum_{\ell=1}^K \log\theta_\ell$, we can write linear equalities on variables $t \in \mathbb{R}^K$ satisfying $\sum_{k=1}^K t_k = 0$, and this is a one-to-one transformation (Dempster 1972, p. 264). Then $\mathcal{F}(\mathbf{u})$ can equivalently be seen as a polytope defined by linear constraints $-\sum_{k=1}^K t_k \leq 0$, $\sum_{k=1}^K t_k \leq 0$, and $t_\ell - t_k \leq \log\eta_{k\to\ell}$ for all k, ℓ. Such representation could be useful in the system reliability setting described by LV, since constraints on products of variables would become linear upon applying a logarithm. Similarly in Section 5.1 we can turn the positive association constraint "$\theta_1\theta_4 > \theta_2\theta_3$" into linear constraints.

Using the half-space representation, lower and upper cumulative distribution functions, such as represented in Figure 6, can be obtained by solving linear programs, namely $\min_{\theta\in\mathcal{F}(\mathbf{u})} \pm\theta_k$ for component k. HHZ mentioned a problem of the form $\min_{\theta\in\mathcal{F}(\mathbf{u})} |\theta - \hat{p}|_2^2$ where $\hat{p} \in \Delta$ is a given point. This is a quadratic program, that can be solved for realistic problem sizes. HHZ also described a program of the form $\min_{\theta\in\mathcal{F}(\mathbf{u})} -|\theta - \hat{p}|_2^2$, which is less standard, and seems challenging. The constraints are again linear, and the objective is quadratic, concave and separable; a quick search points to Kalantari and Rosen (1987), Shen et al. (2016), and Telli, Bentobache, and Mokhtari (2020). Furthermore, there are generic methods to optimize differentiable functions over polytopes, sometimes under the name "linearly constrained global optimization," but these might require more tuning, more computing power, and return results with fewer guarantees.

We illustrate some of these considerations with numerical experiments. The timings are obtained on a single processor (Intel Core i9 at 2.4Ghz). We consider a total of $N = 500$ counts, sampled uniformly over 50 categories; the observed counts are between 4 and 16 across these 50 categories. Using the coupling techniques of Biswas, Jacob, and Vanetti (2019), we find that 40 iterations are enough for the estimated total variation upper bound between the chain and the target distribution to be less than 1%. We measure 0.5 sec to perform

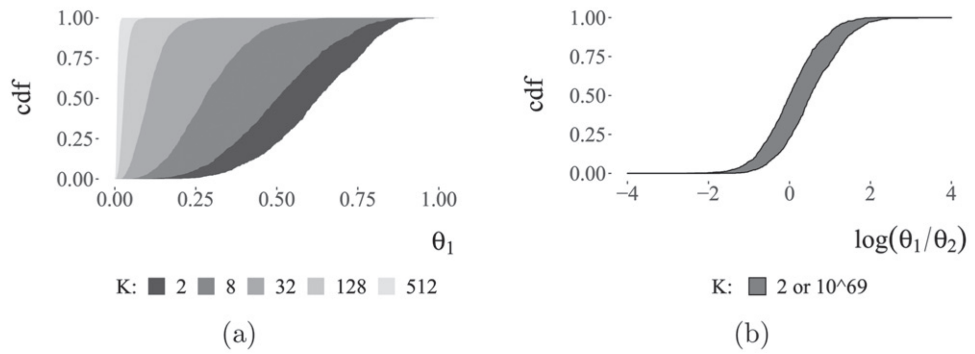

Figure R1. Inference on θ_1 (R1a) and on $\log(\theta_1/\theta_2)$ (R1b) using counts $(4, 3)$ ($K = 2$) and adding various numbers of empty categories to arrive at $K = 512$, or even $K = 10^{69}$. Including empty categories modifies the inference on θ_1 but not on θ_1/θ_2.

each iteration, that is, a full sweep of conditional updates. We next perform 150 iterations and discard the first 50 as burn-in, conservatively. We next add 150 empty categories to each of the 100 polytopes. We record that it takes about 5 sec to incorporate 150 empty categories to all 100 polytopes, resulting in 100 matrices $(\eta_{k\to\ell})$ of size 200 × 200, describing polytopes within the simplex of dimension 200. We measure that the time to solve the linear program $\min_{\theta\in\mathcal{F}(\mathbf{u})} \theta_1$ (once) is about 0.15 sec, and the time to solve the quadratic program $\min_{\theta\in\mathcal{F}(\mathbf{u})} |\theta - \hat{p}|_2^2$ (once), where $\hat{p}$ is the vector of observed frequencies, is about 1 sec.

Therefore, inference using the proposed algorithm can be done in reasonable times, even on a small computer, depending on the assertion of interest. As illustrated by HHZ, it seems important to avoid vertex enumeration if possible, beyond small dimensions. Note that R mentioned works in Robust Bayesian analysis, such as Betrò (2009), where linear semi-infinite programs arise. It seems likely that comparable computational tasks would emerge in any statistical method involving collections of sets of parameters.

4. Dealing With Very Large Numbers of Empty Categories

Some comments suggest the consideration of unknown numbers of categories or very large numbers of them. DW described a setting of card shuffling, where $N = 100$ permutations are observed among $K \approx 10^{69}$ possibilities. As recalled above, empty categories can be added as a postprocessing step, for a cost linear in their numbers, but that remains infeasible if $K \approx 10^{69}$.

It might still be possible to obtain or to anticipate the results of Dempster–Shafer analysis under very large numbers of categories. We revisit the setting of Figure 6 of the article, where the nonzero counts are $N_1 = 4, N_2 = 3$ and where we consider the addition of empty categories. By adding up to a few hundred empty categories, we can see empirically that the upper probabilities on assertions of the form "$\theta_k \in [0, t)$" go to one for all $t \geq 0$, whereas the lower probabilities are unaffected; see Figure R1(a). Presumably, this can be established rigorously; in any case we can guess what would happen with $K = 10^{69}$. Regarding assertions of the form "$\log(\theta_\ell/\theta_k) < t$," both lower and upper probabilities are unaffected by the addition of empty categories; see Figure R1(b).

The R scripts for these additional experiments have been added to *https://github.com/pierrejacob/dempsterpolytope*.

Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank to 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 (grant nos. 2 R37 AI054165-11 and 75N93019C00070).

ORCID

Pierre E. Jacob http://orcid.org/0000-0002-3126-6966
Ruobin Gongb http://orcid.org/0000-0003-2965-9266

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Satistical Society*, Series B, 72, 269–342. [1212]

Andrieu, C., and Roberts, G. O. (2009), "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations," *The Annals of Statistics*, 37, 697–725. [1212]

Beaumont, M. A. (2003), "Estimation of Population Growth or Decline in Genetically Monitored Populations," *Genetics*, 164, 1139–1160. [1212]

Betrò, B. (2009), "Numerical Treatment of Bayesian Robustness Problems," *International Journal of Approximate Reasoning*, 50, 279–288. [1213]

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. [1212]

Cuzzolin, F. (2017), *The Geometry of Uncertainty*, Springer International Publishing. [1211]

Dempster, A. P. (1964), "On the Difficulties Inherent in Fisher's Fiducial Argument," *Journal of the American Statistical Association*, 59, 56–66. [1211]

——— (1966), "New Methods for Reasoning Towards Posterior Distributions Based on Sample Data," *The Annals of Mathematical Statistics*, 37, 355–374. [1211]

——— (1968), "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society*, Series B, 30, 205–247. [1212]

——— (1972), "A Class of Random Convex Polytopes," *The Annals of Mathematical Statistics*, 43, 260–272. [1212]

——— (2008), "The Dempster–Shafer Calculus for Statisticians," *International Journal of Approximate Reasoning*, 48, 365–377. [1211]

——— (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, Chapman and Hall/CRC, pp. 275–288. [1211]

Gong, R., and Meng, X.-L. (2021), "Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox" (with discussion), *Statistical Science*, 36, 169–190. [1211]

Kalantari, B., and Rosen, J. B. (1987), "An Algorithm for Global Minimization of Linearly Constrained Concave Quadratic Functions," *Mathematics of Operations Research*, 12, 544–561. [1212]

Kong, C. T. A. (1986), *Multivariate Belief Functions and Graphical Models*, Ph.D Thesis, Harvard University. [1211]

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, NM: Los Alamos National Lab (LANL). [1212]

Seidenfeld, T., and Wasserman, L. (1993), "Dilation for Sets of Probabilities," *The Annals of Statistics*, 21, 1139–1154. [1211]

Shen, X., Diamond, S., Gu, Y., and Boyd, S. P. (2016), "Disciplined Convex-Concave Programming," 2016 IEEE 55th Conference on Decision and Control (CDC), ARIA Resort & Casino, Las Vegas, NV, USA. pp. 1009–1014. [1212]

Shenoy, P. P., and Shafer, G. (1986), "Propagating Belief Functions With Local Computations," *IEEE Expert*, 1, 43–52. [1211]

Telli, M., Bentobache, M., and Mokhtari, A. (2020), "A Successive Linear Approximation Algorithm for the Global Minimization of a Concave Quadratic Program," *Computational & Applied Mathematics*, 39, 1–28. [1212]

Wasserman, L. (1990), "Belief Functions and Statistical Inference," *Canadian Journal of Statistics*, 18, 183–196. [1211]