Bayesian Analysis (2023) **18**, Number 1, pp. 287–366

Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics (with Discussion)*

Ryan Giordano[†,§], Runjing Liu[‡,§], Michael I. Jordan[‡], and Tamara Broderick[†]

Abstract. Bayesian models based on the Dirichlet process and other stick-breaking priors have been proposed as core ingredients for clustering, topic modeling, and other unsupervised learning tasks. However, due to the flexibility of these models, the consequences of prior choices can be opaque. And so prior specification can be relatively difficult. At the same time, prior choice can have a substantial effect on posterior inferences. Thus, considerations of robustness need to go hand in hand with nonparametric modeling. In the current paper, we tackle this challenge by exploiting the fact that variational Bayesian methods, in addition to having computational advantages in fitting complex nonparametric models, also yield sensitivities with respect to parametric and nonparametric aspects of Bayesian models. In particular, we demonstrate how to assess the sensitivity of conclusions to the choice of concentration parameter and stick-breaking distribution for inferences under Dirichlet process mixtures and related mixture models. We provide both theoretical and empirical support for our variational approach to Bayesian sensitivity analysis.

Keywords: Dirichlet process, stick breaking, local robustness, variational Bayes, Fréchet differentiability, fastSTRUCTURE.

1 Introduction

Scientists and engineers working in a wide range of fields are often interested in inferring the number of clusters in a given data set, as well as inferring which data points belong together. Such inferential questions can be posed naturally within a Bayesian nonparametric (BNP) framework, building on tools such as the Dirichlet process (Ferguson, 1973; Sethuraman, 1994). The Dirichlet process has two useful attributes that have made it be suggested as a natural model of clustering phenomena. First, it is a combinatorial stochastic process, exhibiting discrete structure that allows multiple data points to be associated with the same underlying value of a parameter. Second, its nonparametric nature means that the number of unique parameter values generally grows with the size of the data set, accommodating growth in the number of inferred clusters as data accrue. Such growth is appropriate in many real-world settings; for example, we

*Runjing Liu is supported by the National Science Foundation graduate research fellowship program. Ryan Giordano and Tamara Broderick were supported in part by an NSF CAREER Award and an ONR Early Career Grant.

[†]Department of EECS, MIT, 77 Massachusetts Ave., 38-401, Cambridge, MA 02139
[‡]Department of Statistics, 367 Evans Hall, UC Berkeley, Berkeley, CA 94720
[§]Equal contribution.

 https://doi.org/10.1214/22-BA1309

might expect to keep discovering new species as we examine more individual organisms, and we might expect to discover more topics as we read more articles in a scientific literature. Finally, the overall Bayesian framework in which the Dirichlet process is embedded allows clustering to be treated as one aspect of a larger inferential problem. In particular, the Dirichlet process can be flexibly incorporated into more complex models that exhibit other forms of structure, including hierarchical, spatio-temporal, and topological structure.

Although the BNP framework offers flexibility, it is important to recognize that it is not a black-box method. As with any Bayesian methodology, the deployment of a BNP model involves choices of hyperparameters. Often, these choices are made for reasons of mathematical or computational convenience. Indeed, the nonparametric nature of BNP models can make it particularly difficult to express prior belief subjectively. For example, the latent frequencies of clusters provided by the Dirichlet process are obtained by recursively removing beta-distributed fractions of probability mass from the unit interval. The use of the beta distribution is motivated by its mathematical tractability under recursion and by the fact that it yields a form of conditional conjugacy that can be exploited by Gibbs sampling. These are appealing properties, but it is difficult to imagine justifying this specific choice subjectively, particularly given that observable consequences of the choice are indirect. Even having accepted the beta distribution as a choice of convenience, there remains the problem of choosing the parameter α associated with this distribution. The implications of this choice are again difficult to assess subjectively. In practice the choice is often made based on previous applications or by simply employing a heuristic (Teh et al., 2006; Gelman et al., 2013, Chapter 23).

In summary, it is important to recognize that there will exist many possible values of α, and many possible forms of stick-breaking prior, that might correspond to one's prior beliefs, but which the Dirichlet process framework and other complex BNP models bundle in a way that makes it difficult to understand and to specify *a priori*. Choices of convenience are therefore made, and, unfortunately, these choices can change the results of a data analysis. For instance, α has a direct, proportional relationship to the number of clusters obtained asymptotically in draws from the Dirichlet process. Thus the number of clusters inferred at any particular data size may depend strongly on α. If our scientific conclusions varied substantially because of such dependence, we might worry that these conclusions were driven not by the data and meaningful prior beliefs but instead by our arbitrary or default choices. It behooves us, then, to check how sensitive our conclusions are to these choices.

The outputs of Bayesian inference arise not just from a model and collection of data but also via the use of some posterior approximation. Accordingly, when we assess sensitivity, we should assess the sensitivity of this full procedure to our model choices. In the current paper we focus on Dirichlet process mixture (DPM) models and Variational Bayesian (VB) posterior approximations based on reverse Kullback-Leibler (KL) divergence. VB methods have several favorable properties that motivate their use in the DPM setting. First, they exhibit fast computational scaling due to their use of gradient-based optimization. Second, they avoid the label-switching problem exhibited by Markov Chain Monte Carlo (MCMC) in the mixture-model setting (Jasra et al., 2005). Third, their implementation has become increasingly straightforward due to au-

tomatic differentiation tools (Ranganath et al., 2014; Kucukelbir et al., 2017). Finally, and of particular interest in the current paper, the variational formulation makes it possible to compute closed-form derivative-based expansions of posterior distributions as a function of model hyperparameters (Giordano et al., 2018). Thus VB provides a natural pathway to quantifying the robustness of Bayesian inference.

Concretely, with a fully specified model and inference procedure in hand in the setting of DPM models, we can ask how sensitive some quantity of interest is to the choices of α and the stick-breaking distributions. One option is to propose a number of potential α values, compute the variational approximation at each α value, and report our quantity of interest for each α value. We might similarly assess sensitivity to the stick-breaking distribution over a range of distributional choices. There are at least two major issues with this proposal: (1) while VB is a relatively fast form of approximate Bayesian inference in general, it may still be prohibitively expensive to have to re-run it many times, and (2) it is unclear how best to choose a collection of α and (especially) the stick-breaking distribution values—and how many to choose.

In this work, we address these challenges by making full use of the variational nature of VB methodology. We show how to approximate the nonlinear dependence of the VB optimum on prior choices using a first-order Taylor series expansion. We build on the local robustness tools developed by Giordano et al. (2018) for VB and Gustafson (1996b) for the exact posterior and MCMC approximations. To enable their application to DPM models, we solve a number of open problems: (1) we establish that the optimal VB parameters are a continuously differentiable function of α and a particular parameterization of the stick-breaking form; (2) we show that the sensitivity of the VB approximation to functional prior perturbations takes the form of an integral against a computationally tractable *influence function*—and illustrate how the influence function can provide an interpretable summary of the effect of arbitrary changes to the prior density; (3) to justify using linear approximations over a ball describing different stick-breaking densities, we show that our method is a *uniformly* good approximation by establishing Fréchet differentiability; (4) we show how to compute our approximation efficiently in high-dimensional problems; and (5) we establish the accuracy, practicality, and computational efficiency of our approximation for a variety of models that use stick-breaking, and for various quantities of interest in both clustering and topic modeling.

Our ambition is not to draw conclusions concerning the robustness of BNP procedures in general, nor even of the DPM model in particular. Rather, we offer an easy-to-use computational tool to quickly and automatically assess the sensitivity to prior specification of VB approximations in a particular problem at hand. Though we focus on demonstrating the effectiveness of our methods on the canonical DPM model, our methods apply immediately to any discrete BNP model that admits a truncated stick-breaking approximation. Our ambition, then, is to encourage and empower researchers to explore the robustness of a wide array of datasets and models, including the DPM, but also other stick-breaking variants.

Even further, despite the present paper's focus on BNP, we develop theory that applies directly to all VB approximations based on reverse KL divergence. Indeed, the formation and analysis of our approximation depends only on the implicit function theorem, and so could be readily extended to VB approximations based on other divergence

measures. Thus, though our discussion and experiments will focus on BNP applications, we hope that the present work can serve as a template for the development and analysis of similar local robustness tools in other popular applications of VB.

The remainder of the paper is organized as follows. We briefly review related work in Section 2. In Section 3, we review the stick-breaking construction of the Dirichlet process and our chosen variational approximation. In Section 4, we derive the form of local prior robustness measures for VB approximations. We consider functional perturbations to the stick-breaking density in Section 5, and define the influence function from which we can construct influential and worst-case perturbations. In Section 6, we address scalability and other computational considerations for computing local sensitivity on real applications. In Section 7, we apply our tools to assess the sensitivity of BNP models in several data analysis problems.

2 Related Work

Evaluating sensitivity to prior choices is typically a desirable step of applied Bayesian data analysis (Gelman et al., 2013, Chapter 6), and a central aim of Bayesian robustness is to provide methods and metrics to measure sensitivity of posterior quantities to variations in the model (Insua and Ruggeri, 2000). Our approach to robustness quantification falls in the category of "local robustness" techniques, which are based on differential approximations to model sensitivity (Gustafson, 2000). The contrasting set of "global robustness" techniques avoid differential approximation, but are computationally expensive or infeasible in all but special cases (Sivaganesan, 2000).

In the present work, we study the robustness of a user's problem-specific posterior quantities of interest, such as the expected number of distinct clusters, or the membership of a particular cluster (as in, e.g., Gustafson (1996b)). In contrast, other work attempts to measure the sensitivity of the entire posterior using, for example, the Wasserstein distance or the largest change within an expressive class of posterior expectations (e.g., Roos et al. (2015); Ghaderinezhad and Ley (2019)). In other words, we study the robustness of particular posterior conclusions rather than attempting to measure the robustness, in some sense, of the entire posterior.

Our focus on VB contrasts with much of the previous Bayesian local robustness literature. For posteriors that are approximated via MCMC, the derivatives of local robustness must be approximated with potentially noisy sample covariances (e.g., Gustafson (1996a)). In contrast, the VB optima that we study admit closed-form derivatives via the implicit function theorem. As an optimization procedure, the evaluation of the sensitivity of VB estimates inherits a long tradition of robustness methods in frequentist statistics (e.g. (Jaeckel, 1972; Cook, 1986; Hampel et al., 2011)), a connection which is explored in Giordano et al. (2018). Our work extends Giordano et al. (2018) by providing more easily verifiable sufficient conditions for Theorem 2 of Giordano et al. (2018) and proving results for nonparametric perturbations to the functional form of the prior, including continuous Fréchet differentiability (and non-differentiability). Our theoretical improvements on Giordano et al. (2018) apply to any VB approximation based on reverse KL divergence, not only BNP models. Ultimately, our theoretical work amounts to

an application of the implicit function theorem (Krantz and Parks, 2012), and a similar approach to ours could yield comparable results for VB approximations based on other divergences (e.g., Li and Turner (2016); Liu and Wang (2016); Ambrogioni et al. (2018)).

Many authors have considered the potential sensitivity of discrete BNP posterior quantities to prior specification. Typically, such work relies either on the existence of closed-form solutions or on running multiple MCMC chains with different prior choices (e.g., Nieto-Barajas and Prünster (2009); Saha and Kurtek (2019)). This work has shown that alternatives to the DPM may exhibit improved robustness properties (Barrios et al., 2013; Lijoi et al., 2007; Canale et al., 2017). In the present work, we take the DPM as our starting point only in order to demonstrate our robustness methodology on a well-known and canonical choice of BNP prior. We hope that our methods could act as a supplement to the computationally or analytically intensive techniques employed by the aforementioned papers to quantify robustness. Indeed, our techniques should apply directly to VB approximations of any discrete BNP prior that admits a truncated approximation (Doshi et al., 2009; Roychowdhury and Kulis, 2015; Campbell et al., 2019).

A final distinction between our work and much of the prior Bayesian local robustness literature is underscored by comparison with Basu (2000), a work that also employs local robustness (applied to MCMC) to measure sensitivity to the concentration parameter of a DPM prior specification. Unlike Basu (2000), who considers the norm of the derivative to be a measure of robustness *per se* (following, for example, Basu et al. (1996); Gustafson (1996b)), we focus on the ability of our linear approximation to *extrapolate* to alternative priors. In this spirit, we hope that our work provides tools for quickly and interactively exploring the space of subjectively reasonable prior alternatives, without committing researchers to a single robustness measure chosen more for mathematical convenience than intuitive validity.

3 The Model and Variational Approximation

3.1 A Stick-Breaking Model for Clustering

Consider a standard Bayesian nonparametric generative model for clustering, with observed data $x = (x_n)_{n=1}^N$. We assume a countable infinity of latent components, with frequencies $\pi = (\pi_1, \pi_2, \ldots)$, such that $\pi_k \in [0,1]$ for all $k \in \{1, 2, \ldots\}$, and $\sum_k \pi_k = 1$. For the nth data point, the vector $z_n = (z_{n1}, z_{n2}, \ldots)$ is an indicator vector; $z_{nk} = 1$ represents the assignment of the nth data point to the kth component, with all other vector elements set equal to zero. We generate $z_{nk} = 1$ with probability π_k, i.i.d. across n. To generate the x_n, we assume the kth component is characterized by a component-specific parameter, $\beta_k \in \Omega_\beta \subseteq \mathbb{R}^{D_\beta}$, and that a data point arising from component k is generated as $\mathcal{P}(x_n|\beta_k)$. Then $\mathcal{P}(x_n|z_n, \beta) = \prod_{k=1}^\infty \mathcal{P}(x_n|\beta_k)^{z_{nk}}$. The β_k in turn are generated i.i.d. from a prior $\mathcal{P}_{\text{base}}(\beta_k)$. For instance, in a Gaussian mixture model, β_k could be a vector representing the mean and covariance of a Gaussian distribution.

It remains to place a prior on the component frequencies π. We will focus on stick-breaking priors for π, so we first replace π with a stick-breaking representation. Let

$\nu = (\nu_1, \nu_2, \ldots)$ represent proportions: $\nu_k \in [0, 1]$. Take

$$\pi_k := \nu_k \prod_{k'<k} (1 - \nu_{k'}). \tag{1}$$

We then define a stick-breaking prior by placing a prior on the ν_k. Fix a density, $\mathcal{P}_{\text{stick}}(\cdot)$, with respect to the Lebesgue measure on $[0, 1]$ and let $\nu_k \overset{iid}{\sim} \mathcal{P}_{\text{stick}}(\nu_k)$ for $k \in \{1, 2, \ldots\}$. A common choice of $\mathcal{P}_{\text{stick}}$ is Beta$(1, \alpha)$, with *concentration parameter* $\alpha > 0$. With this choice, the π are distributed according to the size-biased weights associated with the atoms of a draw from a Dirichlet process. This particular beta stick-breaking prior is often favored due to its convenient mathematical properties and ease of use in inference.

Posterior quantities of interest. In theory, with our generative model and observed data in hand, we can find the Bayesian posterior $\mathcal{P}(\beta, z, \nu|x)$ and report any posterior summaries of interest. For instance, the posterior $\mathcal{P}(\beta, z, \nu|x)$ induces a posterior distribution on the number of clusters $G_{\text{cl}}(z)$, where *clusters* are components to which at least one data point has been assigned:

$$G_{\text{cl}}(z) := \sum_{k=1}^{\infty} \mathbb{I}\left(\left(\sum_{n=1}^{N} z_{nk}\right) > 0\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function taking value 1 when the argument is true and 0 otherwise.

In practice, though, neither the posterior nor the posterior summary is readily accessed. An approximation must be used instead.

3.2 Variational Approximation

To assess the sensitivity of a procedure in practice, we need to consider the approximate Bayesian inference algorithm used as well. Here we focus on a variational Bayes approximation due to Blei and Jordan (2006).

Variational Bayes (VB) posits a class of tractable distributions over the model parameters and chooses the element of this class that minimizes the reverse Kullback-Leibler (KL) divergence to the exact posterior. One approach to apply VB to Dirichlet process stick-breaking models assumes $\nu_{K_{\max}} = 1$ for all distributions in the variational class and some truncation level $K_{\max}$. Let ζ collect the first $K_{\max} - 1$ elements of ν, the first $K_{\max}$ elements of β, and the first $K_{\max}$ elements of z_n across n. In what follows, then, we effectively consider the reverse KL divergence to the posterior marginal $\mathcal{P}(\zeta|x)$. By setting $K_{\max}$ sufficiently large, one can make this truncation as accurate as desired.

Mean-field VB is a particularly popular VB variant where the tractable approximating distributions $\mathcal{Q}$ factorize over the parameters. In our case, then, we consider approximations of the form

$$\mathcal{Q}(\zeta|\eta) = \left(\prod_{k=1}^{K_{\max}-1} \mathcal{Q}(\nu_k|\eta)\right)\left(\prod_{k=1}^{K_{\max}} \mathcal{Q}(\beta_k|\eta)\right)\left(\prod_{n=1}^{N} \mathcal{Q}(z_n|\eta)\right), \tag{2}$$

where $\eta \in \Omega_\eta \subseteq \mathbb{R}^{D_\eta}$ represents *variational parameters* that determine the factors of the $\mathcal{Q}$ distribution. When the observation likelihood $\mathcal{P}(x_n|\beta_k)$ is conditionally conjugate with the component-parameter prior $\mathcal{P}_{\text{base}}(\beta_k)$, no further assumptions are needed on the form of $\mathcal{Q}(\beta_k|\eta)$; one can show that it will take the form of the conjugate exponential family after the KL optimization (Blei et al., 2017). Similarly, when $\mathcal{P}_{\text{stick}}$ is a beta distribution, no further assumptions are needed on $\mathcal{Q}(\nu_k|\eta)$; it will take a beta form. However, since we will consider non-beta forms of $\mathcal{P}_{\text{stick}}$, we must specify a more generic approximation—one that will work even when conditional conjugacy does not hold. To that end, we first transform the ν_k to a value that is unbounded and then use a Gaussian approximation. Define the logit-transformed stick-breaking proportions $\tilde{\nu}_k$:

$$\tilde{\nu}_k := \log(\nu_k) - \log(1-\nu_k) \quad \Leftrightarrow \quad \nu_k = \frac{\exp(\tilde{\nu}_k)}{1+\exp(\tilde{\nu}_k)}.$$

We take $\mathcal{Q}(\tilde{\nu}_k|\eta)$ to be a normal distribution, which induces a logit-normal distribution on ν_k. We approximate all resulting integrals over $\mathcal{Q}(\tilde{\nu}_k|\eta)$, as in the KL objective for VB or in our later sensitivity calculations, with Gauss-Hermite (GH) quadrature; see Supplement D.4 (Giordano et al., 2022).

GH quadrature yields an approximation, which we call $\text{KL}(\eta)$, to the full KL, $\text{KL}(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x))$. We minimize that approximation to perform approximate posterior inference:

$$\text{KL}(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x)) = \mathop{\mathbb{E}}_{\mathcal{Q}(\zeta|\eta)}[\log \mathcal{Q}(\zeta|\eta) - \log \mathcal{P}(x,\zeta)] + \log \mathcal{P}(x) \tag{3}$$

$$\hat{\eta} := \mathop{\text{argmin}}_{\eta\in\Omega_\eta} \text{KL}(\eta) \quad \text{where} \quad \text{KL}(\eta) \approx \text{KL}(\mathcal{Q}(\zeta|\eta)||\mathcal{P}(\zeta|x)). \tag{4}$$

Our final approximation to the marginal posterior $\mathcal{P}(\zeta|x)$ is $\mathcal{Q}(\zeta|\hat{\eta})$.

Posterior quantities of interest. To approximate any functional of the exact posterior, we apply the equivalent functional to $\mathcal{Q}(\zeta|\hat{\eta})$. For instance, the approximation to the posterior expected number of clusters among the N observed data points is

$$\mathop{\mathbb{E}}_{\mathcal{Q}(\zeta|\hat{\eta})}[G_{\text{cl}}(z)] = \mathop{\mathbb{E}}_{\mathcal{Q}(z|\hat{\eta})}[G_{\text{cl}}(z)] = \sum_{k=1}^{K_{\max}-1}\left(1 - \prod_{n=1}^{N}(1 - \mathop{\mathbb{E}}_{\mathcal{Q}(z_n|\hat{\eta}_z)}[z_{nk}])\right). \tag{5}$$

We will see examples in Section 7 where our quantity of interest is (a) the expected posterior number of clusters in the observed data, (b) the expected posterior number of clusters in a new set of (as yet unobserved) data, (c) some aspect of a co-clustering matrix, or (d) the topic assignments of certain data points. In all of these cases, as in (5), we are able to express our (approximate) posterior quantity of interest as a smooth function g of the optimized variational parameters $\hat{\eta}$: $g(\hat{\eta})$. Indeed, as we will discuss in Section 4, our methods and results apply to any quantity of interest that can be written as a smooth function of $\hat{\eta}$.

Once we have an (approximate) posterior quantity of interest, we can ask how this quantity would change—and whether our substantive scientific conclusions would change—if we had made reasonably different prior choices.

4 A Local Approximation for Sensitivity

We would like to understand how our quantity of interest $g(\hat{\eta})$ changes when the concentration parameter or, more generally, the stick-breaking density $\mathcal{P}_{\text{stick}}$ changes. To efficiently compute these changes, we use a first-order Taylor series approximation in the optimal VB parameters. In this section, we first present the Taylor series and then show how to compute its terms.

Sensitivity to the concentration parameter. First, we show how to approximate the sensitivity of $g(\hat{\eta})$ to the choice of concentration parameter α. Let $\hat{\eta}(\alpha)$ represent the value of $\hat{\eta}$ for a particular choice of α. For our approximation, we choose some initial value α_0 of the concentration parameter and solve the optimization problem to compute $\hat{\eta}(\alpha_0)$. We then approximate $\hat{\eta}(\alpha)$ with the linear approximation $\hat{\eta}^{\text{lin}}(\alpha)$, and in turn approximate $g(\hat{\eta}(\alpha))$ with $g(\hat{\eta}^{\text{lin}}(\alpha))$:

$$\hat{\eta}^{\text{lin}}(\alpha) := \hat{\eta}(\alpha_0) + \left.\frac{d\hat{\eta}(\alpha)}{d\alpha}\right|_{\alpha_0} (\alpha - \alpha_0) \quad \text{and} \quad g(\hat{\eta}(\alpha)) \approx g(\hat{\eta}^{\text{lin}}(\alpha)). \tag{6}$$

If $\alpha \mapsto \hat{\eta}(\alpha)$ is continuously differentiable, and g is sufficiently smooth, then we expect $g(\hat{\eta}(\alpha)) \approx g(\hat{\eta}^{\text{lin}}(\alpha))$ when $|\alpha - \alpha_0|$ is small. We will show in Theorem 1 below that the map $\alpha \mapsto \hat{\eta}(\alpha)$ is continuously differentiable for our chosen VB approximation.

Sensitivity to the stick-breaking density. Next, we show how to approximate the sensitivity of $g(\hat{\eta})$ to the choice of concentration stick distribution $\mathcal{P}_{\text{stick}}$. Technically, perturbations of α are perturbations of $\mathcal{P}_{\text{stick}}$. But here we consider more general perturbations of the form of $\mathcal{P}_{\text{stick}}$, potentially outside the beta class. To define our perturbations, let $\tilde{\mathcal{P}}$ represent a potentially unnormalized (but normalizable) density with respect to Lebesgue measure; the same notation without the tilde will give the normalized density. Now start from an initial setting of $\mathcal{P}_{\text{stick}}$ at $\mathcal{P}_0$; we will typically start from Dirichlet-process stick-breaking; i.e., $\mathcal{P}_0 = \text{Beta}\,(1, \alpha_0)$ for some α_0. Then take any Lebesgue-measurable function $\phi(\cdot)$ on $[0, 1]$. We consider a range of alternative (potentially unnormalized) stick-breaking forms $\tilde{\mathcal{P}}(\cdot|t)$ defined on $[0, 1]$ by

$$\log \tilde{\mathcal{P}}(\cdot|t) = \log \mathcal{P}_0(\cdot) + t\phi(\cdot). \tag{7}$$

Note that the perturbation applies equally to every stick break ν_k. This style of multiplicative functional perturbation was proposed by Gustafson (1996b); we deviate from Gustafson (1996b) by considering VB (rather than MCMC) approximations and by allowing ϕ to take on negative values.

If we now let $\hat{\eta}(t)$ represent the value of $\hat{\eta}$ for a particular choice of $\tilde{\mathcal{P}}(\cdot|t)$, we can form an approximation analogous to (6):

$$\hat{\eta}^{\text{lin}}(t) := \hat{\eta}(0) + \left.\frac{d\hat{\eta}(t)}{dt}\right|_{t=0} (t - 0) \quad \text{and} \quad g(\hat{\eta}(t)) \approx g(\hat{\eta}^{\text{lin}}(t)). \tag{8}$$

As in the case of expansions with respect to α, (8) is useful only if the map $t \mapsto \hat{\eta}(t)$ is continuously differentiable for the chosen ϕ. As we will show in Theorem 1 below, a

sufficient condition for differentiability is given in terms of the following norm on the perturbation ϕ.

$$\text{Define } \|\phi\|_\infty := \operatorname*{esssup}_{\nu_0 \sim \mathcal{P}_0} |\phi(\nu_0)| \text{ and } \quad \mathcal{B}_\phi(\delta) := \{\phi : \|\phi\|_\infty < \delta\}. \tag{9}$$

The set of priors that arise by considering functional perturbations $\phi \in \mathcal{B}_\phi(\delta)$ live in a multiplicative band around the original prior, $\mathcal{P}_0$, as shown in Figure 1. Theorem 1 below states that $t \mapsto \hat{\eta}(t)$ is continuously differentiable whenever $\|\phi\|_\infty < \infty$. So, for sufficiently smooth g, we expect the approximation (8) to be good for small t, given a particular choice of ϕ with $\|\phi\|_\infty < \infty$.

The functional perturbation given in (7) is useful because, if we consider any other distribution $\mathcal{P}_1$ for $\mathcal{P}_{\text{stick}}$, we can continuously warp $\mathcal{P}_0$ to $\mathcal{P}_1$ by setting $\phi(\cdot) = \log(\mathcal{P}_1(\cdot)/\mathcal{P}_0(\cdot))$ so long as $\mathcal{P}_1 \ll \mathcal{P}_0$; i.e., $\mathcal{P}_1$ is absolutely continuous with respect to $\mathcal{P}_0$. We will see in Section 5 that we can compute an *influence function* to provide an interpretable summary of the effect of arbitrary changes ϕ. Using the influence function and the $\|\cdot\|_\infty$ norm, we are able to find a worst-case choice of ϕ in $\mathcal{B}_\phi(\delta)$.

However, we note that restricting to $\|\phi\|_\infty < \infty$ limits the kinds of alternative priors $\mathcal{P}_1$ that can be formed using (7). Although we show in Lemma 1 of Supplement A.3 that functional perturbations with $\|\phi\|_\infty < \infty$ yield valid priors, the converse is not true: there exist valid priors $\mathcal{P}_1$ such that the corresponding $\|\phi\|_\infty = \infty$. For instance, perturbing the beta stick-breaking form by changing α provides a counterexample since the log of the beta density is unbounded below; see Example 3 of Supplement A.3 for more details. The limited expressiveness of $\mathcal{B}_\phi(\delta)$ may at first seem like a shortcoming of the perturbation given by (7). However, we show in Section 5 that, among a class of potential functional perturbations such as those proposed by Gustafson (1996b), only the one we defined in (7) is *Fréchet differentiable*—and thus can be used to safely reason about worst-case ϕ.

Computing the terms in the Taylor series. It remains to show that $\alpha \mapsto \hat{\eta}(\alpha)$ and $t \mapsto \hat{\eta}(t)$ are continuously differentiable, and to provide a computable formula for the derivative. Differentiability naturally requires some regularity conditions on the VB parameterization and on the optimum. We state sufficient conditions in the following Assumption 1, which is satisfied for any local optimum of a smooth, unconstrained parameterization of the variational approximation.

Assumption 1. *Assume that: (1) the map $\eta \mapsto \mathrm{KL}(\eta)$ is twice continuously differentiable at $\hat{\eta}$; (2) the Hessian matrix $\frac{\partial^2 \mathrm{KL}(\eta)}{\partial \eta \partial \eta^T}\Big|_{\hat{\eta}}$ is non-singular; and (3) there exists an open ball $\mathcal{B}_\eta \subseteq \mathbb{R}^{D_\eta}$ such that $\hat{\eta} \in \mathcal{B}_\eta \subseteq \Omega_\eta$.*

Our next result establishes the differentiability of $\hat{\eta}$ and provides a computable formula for the derivative.

Theorem 1. *Let Assumption 1 hold for the VB approximation given in Section 3.2. Either take $\varepsilon = t$ under the perturbation given by $\log \tilde{\mathcal{P}}(\nu_k|t) = \log \mathcal{P}_0(\nu_k) + t\phi(\nu_k)$ with $\|\phi\|_\infty < \infty$, or take $\varepsilon = \alpha - \alpha_0$ in a perturbation to the concentration parameter α of*

the unnormalized beta distribution $\log \tilde{\mathcal{P}}(\nu_k|\alpha) = \alpha \log(1-\nu_k)$. *Then the map* $\varepsilon \mapsto \hat{\eta}(\varepsilon)$ *is continuously differentiable at* $\varepsilon = 0$ *with derivative*

$$\left.\frac{d\hat{\eta}(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = -\hat{H}^{-1}\hat{J}, \quad \textit{where} \quad \rho_k(\nu_k) := \left.\frac{\partial \log \tilde{\mathcal{P}}(\nu_k|\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0}, \tag{10}$$

$$\hat{H} := \left.\frac{\partial^2 \mathrm{KL}\,(\eta)}{\partial \eta \partial \eta^T}\right|_{\eta=\hat{\eta}}, \quad \mathcal{S}(\zeta|\eta) := \left.\frac{\partial \log \mathcal{Q}(\zeta|\eta)}{\partial \eta}\right|_{\eta}, \textit{ and} \tag{11}$$

$$\hat{J} := \frac{\partial}{\partial \eta} \mathop{\mathbb{E}}_{\mathcal{Q}(\zeta|\eta)} \left[\sum_{k=1}^{K_{\max}-1} \rho_k(\nu_k)\right]\Bigg|_{\eta=\hat{\eta}} = \mathop{\mathbb{E}}_{\mathcal{Q}(\zeta|\hat{\eta})} \left[\mathcal{S}(\zeta|\hat{\eta}) \sum_{k=1}^{K_{\max}-1} \rho_k(\nu_k)\right]. \tag{12}$$

Proof. The result follows from Theorem 4 of Supplement A.1, which states general conditions for the differentiability of VB optima. We show in Supplement A.2 and A.3 that the conditions of Theorem 4 are satisfied in the case of our present BNP problem. The equivalence of the expressions for $\hat{J}$ follows by differentiating through the expectation; see Lemma 3 of Supplement B for more details. □

(10) requires computation of two terms: $\hat{H}^{-1}$ and $\hat{J}$. Typically, $\hat{J}$, which is a derivative of a variational expectation, is straightforward to evaluate: the requisite expectation is evaluated either in closed form or approximated numerically; then, in either case, an application of automatic differentiation provides the gradient (Baydin et al., 2018). Forming and inverting or factorizing $\hat{H}$ can present a challenge due to its high dimensionality—it has dimensions $D_\eta \times D_\eta$, where D_η is the dimension of η. However, in many cases—including the BNP problem that is our focus—we can take advantage of model sparsity to efficiently compute (10) (see Section 6), and our experiments confirm that we can compute $\left.\frac{d\hat{\eta}(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}$ much more efficiently than re-optimizing the VB objective directly (Section 7.4). Moreover, the savings increase dramatically when we are interested in a range of ε values because $\left.\frac{d\hat{\eta}(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}$ can be re-used to for any chosen value of ε.

5 The Influence Function and Worst-Case Functional Perturbations

We next show how to find influential and worst-case functional perturbations to the stick-breaking density. We start by showing how to compute an influence function to summarize the effect of different choices of ϕ. Using the influence function, we are able to design stick-breaking densities that produce a large change in a quantity of interest, including computing the worst-case perturbation in $\mathcal{B}_\phi(\delta)$. To justify such uses of the influence function, we prove that, for multiplicative perturbations and the ∞-norm, the VB objective is Fréchet differentiable—i.e., that it admits a uniformly good linear approximation in a neighborhood of the null perturbation. Finally, we show that our Fréchet differentiability result is unique among a broad class of alternative choices of functional perturbation.

The influence function and worst-case perturbations. We begin by defining the influence function Ψ and discussing its usefulness for understanding the effect of functional perturbations ϕ. Suppose we have a one-dimensional, differentiable quantity of interest, $g(\cdot) : \Omega_\eta \mapsto \mathbb{R}$, and are considering various alternative priors as given by ϕ in (7). Under the approximation in (8), the dependence of $g(\hat{\eta}^{\text{lin}}(t))$ on ϕ is not simple if $g(\cdot)$ is non-linear. However, for a particular choice of ϕ, by applying the chain rule with Theorem 1, we can derive a fully linear approximation $g(\hat{\eta}(t)) \approx g(\hat{\eta}) + \frac{dg(\hat{\eta}(t))}{dt}\Big|_{t=0}(t-0)$. The advantage of linearizing g in this way is that the map $\phi \mapsto \frac{dg(\hat{\eta}(t))}{dt}\Big|_{t=0}$ has a particularly simple form, as given by the following result.

Corollary 1. *Under the conditions of Theorem 1, using (7) with $\|\phi\|_\infty < \infty$ and $\varepsilon = t$, let $g(\cdot) : \Omega_\eta \mapsto \mathbb{R}$ denote a continuously differentiable, real-valued function of interest. Define the* influence function *$\Psi : [0,1] \mapsto \mathbb{R}$:*

$$\Psi(\cdot) := -\sum_{k=1}^{K_{\max}-1} \frac{dg(\eta)}{d\eta^T}\Big|_{\hat{\eta}} \hat{H}^{-1}\mathcal{S}_k(\cdot|\hat{\eta})\mathcal{Q}_k(\cdot|\hat{\eta}), \tag{13}$$

where $\mathcal{S}_k(\cdot|\hat{\eta})$ and $\mathcal{Q}_k(\cdot|\hat{\eta})$ replace $\mathcal{Q}(\zeta|\eta)$ with just the factor of $\mathcal{Q}$ for ν_k. Then the derivative in (10) can be written as

$$\frac{dg(\hat{\eta}(t))}{dt}\Big|_0 = \int_0^1 \Psi(\nu_0)\phi(\nu_0)d\nu_0. \tag{14}$$

Proof. The form of the influence function is given by the chain rule, gathering terms in (10), and re-writing the variational expectation as an integral over $[0,1]$. We establish an analogous general result for general VB approximations in Corollary 3 of Supplement A.3, specializing to the BNP case in Example 4 of Supplement A.3. □

By choosing perturbations ϕ that align with the influence function, we can form priors that we expect to be influential for the function of interest, $g(\cdot)$. For example, in our experiments of Section 7, we show that by choosing ϕ to be a Gaussian bump aligned with particularly high-magnitude positive or negative values of the influence function, one can ensure a large positive or negative gradient, and hence a large predicted change.

Further, with Corollary 1 in hand, we can find a closed-form expression for the worst-case choice of $\phi \in \mathcal{B}_\phi(\delta)$, which is essentially a VB analogue to Gustafson (1996b, Result 11).

Corollary 2. *Under the conditions of Corollary 1,*

$$\sup_{\phi \in \mathcal{B}_\phi(\delta)} \frac{dg(\hat{\eta}(t))}{dt}\Big|_0 = \delta \int |\Psi(\nu_0)|\, \mu(d\nu_0),$$

and the supremum is achieved at the perturbation $\phi^*(\cdot) = \delta \operatorname{sign}(\Psi(\cdot))$.

Proof. The result follows immediately from applying Hölder's inequality to (14). We establish a similar but much more general result for VB approximations with general choices of model and parameters in Corollary 4 of Supplement A.4. The present result is a special case using Example 4 of Supplement A.4. □

In our experiments of Section 7, we use Corollaries 1 and 2 to choose influential perturbations, and then use the partially linearized (8) to make predictions about the effect of the perturbations.

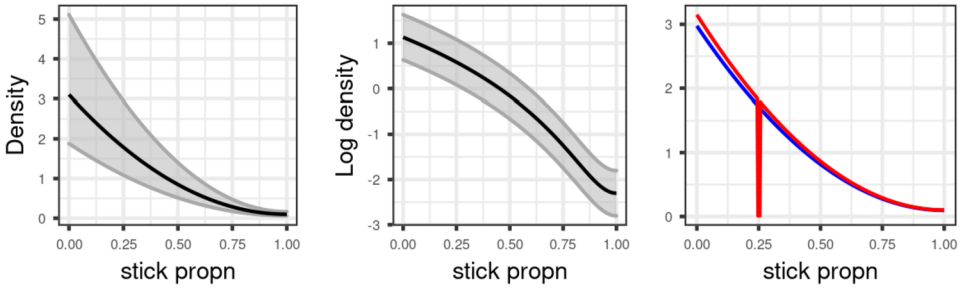

Figure 1: Left two: A multiplicative ball $\mathcal{B}_\phi(\delta)$. Right: Two densities that are distant according to reverse KL divergence and $\|\cdot\|_\infty$ but close according to $\|\cdot\|_p$ for $p \in [1, \infty)$.

Multiplicative perturbations are continuously Fréchet differentiable. The influence function provides a succinct summary of the effect of all perturbations $\phi \in \mathcal{B}_\phi(\delta)$, which we might hope to be accurate for sufficiently small δ. However, the accuracy of our approximation within $\mathcal{B}_\phi(\delta)$ is not guaranteed by Theorem 1 alone. Specifically, Theorem 1 states only that, for a *particular* direction ϕ, $t \mapsto \hat{\eta}(t)$ is continuously differentiable—i.e. that, for a fixed ϕ, one can make t sufficiently small so that the error $\left|\hat{\eta}(t) - \hat{\eta}^{\text{lin}}(t)\right|$ goes to zero faster than t. But, if we write $\hat{\eta}(t\phi)$ and $\hat{\eta}^{\text{lin}}(t\phi)$ to make the dependence on ϕ explicit, then Theorem 1 does not imply that for a fixed δ (no matter how small), the worst-case error $\sup_{\phi \in \mathcal{B}_\phi(\delta)} \left|\hat{\eta}(\phi) - \hat{\eta}^{\text{lin}}(\phi)\right|$ is bounded, much less that it goes to zero faster than δ.

Thus, to be assured that the influence function is a meaningful summary of the effect of all $\phi \in \mathcal{B}_\phi(\delta)$, we wish to establish that the linear approximation given by (8) is uniformly accurate over all ϕ of interest within a sufficiently small neighborhood of the zero function. Specifically, observing that ϕ is a point in the Banach space L_∞ (Dudley, 2018, Theorem 5.2.1), we wish to establish that the map $\phi \mapsto \hat{\eta}(\phi)$ from L_∞ to $\mathbb{R}^{D_\eta}$ is *Fréchet differentiable*, as we now formally define.[1]

[1]Fréchet differentiability is sometimes referred to as "bounded" differentiability. In addition to Fréchet, two other notions of differentiability are common in statistics: Hadamard (i.e., compact) differentiability, and Gateaux (i.e., weak, or directional) differentiability. In each case, the derivative is given by the same linear operator, but comes with different accuracy guarantees, of which Fréchet is the strongest. Consequently, our Theorem 2 below implies both Hadamard and Gateaux differentiability as a consequence of Fréchet differentiability. See Averbukh and Smolyanov (1967) for a general mathematical treatment of differentiability in Banach spaces, or Reeds (1976) for a treatment intended for statisticians.

Definition 1. (Fréchet differentiability, (Zeidler, 1986, Definition 4.5)) Let B_1 and B_2 denote Banach spaces, and let $\mathcal{B}_1 \subseteq B_1$ define an open neighborhood of $\phi_0 \in B_1$. A function $f : \mathcal{B}_1 \mapsto B_2$ is *Fréchet differentiable* at ϕ_0 if there exists a bounded linear operator, $f^{\text{lin}} : B_1 \mapsto B_2$, such that, for $\phi \in B_1$,

$$f(\phi) - f(\phi_0) - f^{\text{lin}}_{\phi_0}(\phi - \phi_0) = o(\|\phi - \phi_0\|) \quad \text{as} \quad \|\phi - \phi_0\| \to 0.$$

The function f is *continuously Fréchet differentiable* if the map $\phi_0 \mapsto f^{\text{lin}}_{\phi_0}(\cdot)$ is continuous as a map from $\mathcal{B}_1$ to the space of all continuous linear operators from B_1 to B_2 equipped with the operator norm. ⊡

By Zeidler (1986, Proposition 4.8), if a function is Fréchet differentiable, then the linear operator f^{lin} is given precisely by the directional derivative $df(\phi_0 + t(\phi - \phi_0))/dt$. Thus, if $\phi \mapsto \hat{\eta}(\phi)$ is Fréchet differentiable, its derivative is given by Corollary 1. Fréchet differentiability guarantees that, for sufficiently small δ, the error of the linear approximation given by Corollary 1 does not blow up in the ball $\mathcal{B}_\phi(\delta)$.

We emphasize that Fréchet differentiability is neither sufficient nor necessary for a derivative to be useful. For example, it is possible in principle for a function to be Fréchet differentiable but still have a very large finite second derivative, and so fail to extrapolate meaningfully to any alternatives one cares about. Conversely, if a function fails to be Fréchet differentiable, the derivative may still perform well in particular directions, including that chosen by Corollary 2. Nevertheless, Fréchet differentiability is a strong local result, and provides some assurance that one can use results such as Corollary 2 without uncovering pathological behavior.

Finally, then, we prove that our perturbation is continuously Fréchet differentiable.

Theorem 2. *Under the conditions of Theorem 1, the map $\phi \mapsto \hat{\eta}(\phi)$ is well-defined and continuously Fréchet differentiable in a neighborhood of the zero function as a map from L_∞ to $\mathbb{R}^{D_\eta}$, with the derivative given in Corollary 1.*

Proof. Our result here is a special case of our general result for VB approximations given in Theorem 5 of Supplement A.4. □

Many other functional perturbations and norms are not Fréchet differentiable. So far we have focused on the multiplicative functional perturbations in (7) combined with the infinity norm in (9). We now ask whether we could perform a similar analysis for other functional perturbations. We show that, of the perturbations proposed by Gustafson (1996b), only multiplicative perturbations yield Fréchet differentiable VB optima.

Specifically, Gustafson (1996b) examines general perturbations, from initial prior $\mathcal{P}_0$ to alternative $\mathcal{P}_1$, that take the following form—with θ a parameter $\theta \in \Omega_\theta \subseteq \mathbb{R}^{D_\theta}$ and $p \in [1, \infty)$:

$$\tilde{\mathcal{P}}(\theta|t_p) := \left((1 - t_p)\mathcal{P}_0(\theta)^{1/p} + t_p \frac{1}{p}\mathcal{P}_1(\theta)^{1/p}\right)^p. \tag{15}$$

Again, let ϕ represent the perturbation, now with:

$$\phi(\theta|\mathcal{P}_1, p) := \mathcal{P}_1(\theta)^{1/p} - \mathcal{P}_0(\theta)^{1/p} \quad \text{and} \quad \|\phi\|_p := \left(\int_0^1 |\phi(\theta)|^p \, d\theta\right)^{1/p}. \tag{16}$$

The limit $p \to \infty$ recovers our multiplicative perturbation in (7) with infinity norm in (9). The choice $p = 1$ recovers a purely additive perturbation. Gustafson (1996b, Result 2) states that $\|\phi\|_p < \infty$ ensures that the corresponding $\tilde{\mathcal{P}}(\theta|t_p)$ can be normalized, strongly motivating using the $\|\cdot\|_p$ norm with the perturbation given by (15).

Our next theorem shows that the reverse KL divergence is discontinuous in $\|\cdot\|_p$ for $p < \infty$. Since Fréchet differentiability implies continuity (Zeidler, 1986, Proposition 4.8 (d)), Theorem 3 implies that it is impossible to derive an analogue of Theorem 2 for perturbations of the form in (15) with the norms in (16).[2]

Theorem 3. *Let μ denote a measure on the Borel sets of some domain Ω_θ, with μ absolutely continuous with respect to the Lebesgue measure, and let $\mathcal{Q}(\theta)$ and $\mathcal{P}_0(\theta)$ denote densities with respect to μ. Without loss of generality, assume that $\mathcal{Q}(\theta) > 0$ on Ω_θ. Assume that $\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}_0(\theta)\right)$ is well-defined and finite.*

Then, for any $\epsilon > 0$ and any $M > 0$, we can find a density $\mathcal{P}_1(\theta)$ such that $\|\phi(\theta|\mathcal{P}_1, p)\|_p < \epsilon$ but $|\mathrm{KL}\left(q(\theta)||\mathcal{P}_1(\theta)\right) - \mathrm{KL}\left(q(\theta)||\mathcal{P}_0(\theta)\right)| > M$.

Proof. See Supplement A.5 for a constructive proof, the key to which is the fact that in any $\|\cdot\|_p$ neighborhood of zero there exist prior densities taking values arbitrarily close to zero on sets of nonzero measure, for which the reverse KL divergence blows up. □

Recall from Section 4 (and particularly Example 3 of Supplement A.3) that there exist priors that cannot be formed from (7) using ϕ with $\|\phi\|_\infty < \infty$. In light of the proof of Theorem 3, the limited expressiveness of multiplicative perturbations with the $\|\cdot\|_\infty$ norm looks like a feature rather than a bug. Consider the rightmost panel of Figure 1, which illustrates the tradeoffs between the various norms. The two blue and red densities are far from one another according to reverse KL divergence since the red density takes values that are nearly zero where the blue density has nonzero mass. The two densities are also distant in $\|\cdot\|_\infty$ since it takes a large multiplicative change to turn the nonzero blue density into the nearly zero red density. However, the two densities are close in $\|\cdot\|_p$ since the region where the red density is nearly zero has a small measure. In order for VB approximations to be continuous (a necessary condition for Fréchet differentiability), one must consider a topology on priors that is no coarser than the topology induced by reverse KL divergence. But since valid priors can take values close

[2]Hadamard differentiability also implies continuity, so our Theorem 3 also implies Hadamard non-differentiability (Averbukh and Smolyanov, 1967, Section 3). In general, a functional may be Gateaux differentiable but discontinuous, though in such cases there are necessarily directions in which the derivative provides an arbitrarily poor approximation to the behavior of the functional (see Averbukh and Smolyanov (1967, Example 1.19)). In the present case, as we discuss in Supplement A.5, there exist pointwise negative priors in every $\|\cdot\|_p$ neighborhood of $\mathcal{P}_0$ for $p < \infty$, so even establishing Gateaux differentiability (i.e., the mere existence of a directional derivative in every direction) requires a somewhat artificial extension of the KL divergence to accommodate pointwise negative prior densities.

to zero, a sacrifice in expressiveness of the neighborhood of zero must be made in order to induce a topology that is compatible with reverse KL divergence. Multiplicative changes and the $\|\cdot\|_\infty$ norm implement such a tradeoff in a natural, easy-to-understand way.

In this sense, VB approximations based on reverse KL divergence are inherently non-robust to priors that ablate mass nearly to zero. No parameterization of the space of priors will relieve this non-robustness. Only by basing variational approximations on divergences other than reverse KL might this non-robustness be alleviated.

6 Fast Computation of the Sensitivity

A principal challenge of computing the sensitivity efficiently is the high-dimensional nature of the parameter ζ and hence the variational parameters η. In particular, we have seen that, in our BNP stick-breaking model, ζ and η both grow linearly with the number of data points N. This growth leads to two major computational challenges: (1) we must solve a high-dimensional optimization problem to extremize the VB objective, and (2) we must solve a linear system given by the Hessian $\hat{H}$. Here we show how we can use special structure in the model to reduce to low-dimensional problems and thereby enjoy efficient computation.

Global and local parameters. In both cases, the key to reducing to a lower-dimensional problem is separating *global* and *local* parameters. Global variables are common to all data points. Local variables are unique to each data point. For instance, in a Gaussian (or other typical) mixture model, the stick-breaking proportions ν and component parameters β are global, whereas the cluster assignment parameters z are local.

Let γ denote the collection of global parameters. When we use a standard mean-field VB parameterization, the VB distributions on γ have their own variational parameters, which we denote η_γ. Similarly, let ℓ denote the local parameters and let η_ℓ be the corresponding local variational parameters.

Reducing to optimization over the global variational parameters. We next show how to reduce the potentially high-dimensional optimization problem over all of η to optimizing over just the global variational parameters η_γ.

In all models we will consider, the conditional posterior $\mathcal{P}(z|\gamma, x)$ has a tractable closed form. Since we choose a conjugate mean-field approximating family for $\mathcal{Q}(z|\eta)$, the optimal local variational parameters $\hat{\eta}_\ell$ can be written as a closed-form function of the global variational parameters η_γ. For some prior parameter ε (as in Theorem 1), let $\hat{\eta}_\ell(\eta_\gamma; \varepsilon)$ denote this mapping, so that

$$\hat{\eta}_\ell(\eta_\gamma; \varepsilon) := \operatorname*{argmin}_{\eta_\ell} \mathrm{KL}\left((\eta_\gamma, \eta_\ell), \varepsilon\right). \tag{17}$$

In Example 6 (Supplement D.1), we illustrate this technique for a Gaussian mixture model. Using (17), we can rewrite our objective as a function of the global parameters. Define

$$\mathrm{KL}_{\mathrm{glob}}(\eta_\gamma, \varepsilon) := \mathrm{KL}\Big((\eta_\gamma, \hat{\eta}_\ell(\eta_\gamma; \varepsilon)), \varepsilon\Big).$$

The $\hat{\eta}_\gamma(\varepsilon)$ that minimizes $\mathrm{KL}_{\mathrm{glob}}(\eta_\gamma, \varepsilon)$ is the same as the corresponding sub-vector of the $\hat{\eta}(\varepsilon)$ that minimizes $\mathrm{KL}(\eta, \varepsilon)$.

Rather than optimizing the $\mathrm{KL}(\eta)$ over all variational parameters, we numerically optimize $\mathrm{KL}_{\mathrm{glob}}$, which is a function only of the relatively low-dimensional global parameters. To minimize $\mathrm{KL}_{\mathrm{glob}}(\eta_\gamma)$ in practice, we run the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm with a loose convergence tolerance followed by the trust-region Newton conjugate gradient method to find a high-quality optimum (the `trust-ncg` method of `scipy.optimize.minimize`, Virtanen et al. (2020); see also Nocedal and Wright (2006, Chapter 7)). After the optimization terminates at an optimal $\hat{\eta}_\gamma$, the optimal local parameters $\hat{\eta}_\ell$ can be set in closed form to produce the entire vector of optimal variational parameters, $\hat{\eta} = (\hat{\eta}_\gamma, \hat{\eta}_\ell)$.

6.1 Computing and Inverting the Hessian

Since the dimension D_η of η scales with N, we can quickly reach cases where inverting or even instantiating a dense matrix of size $D_\eta \times D_\eta$ in memory would be prohibitive. The key to efficient computation is that $\hat{H}$ is not dense; we will again exploit structure inherent in the global/local decomposition.

For generic variables a and b, let H_{ab} denote the sub-matrix $\partial^2 \mathrm{KL}(\eta) / \partial \eta_a \eta_b^T \big|_{\hat{\eta}}$, the Hessian with respect to the variational parameters governing a and b. We decompose the Hessian matrix $\hat{H}$ into four blocks according to the global/local decomposition:

$$\hat{H} = \frac{\partial^2 \mathrm{KL}(\eta)}{\partial \eta \partial \eta^T}\bigg|_{\hat{\eta}} = \begin{pmatrix} H_{\gamma\gamma} & H_{\gamma\ell} \\ H_{\ell\gamma} & H_{\ell\ell} \end{pmatrix}.$$

Similarly, let $\hat{J}_\gamma$ be the components of $\hat{J}$ corresponding to the variational parameters η_γ. The local components, $\hat{J}_\ell$, are zero since no local variables enter the expectation in (12) when we are perturbing the stick-breaking distribution.

In this notation,

$$\frac{d\hat{\eta}(\varepsilon)}{d\varepsilon}\bigg|_{\varepsilon=0} = -\begin{pmatrix} H_{\gamma\gamma} & H_{\gamma\ell} \\ H_{\ell\gamma} & H_{\ell\ell} \end{pmatrix}^{-1} \begin{pmatrix} \hat{J}_\gamma \\ 0 \end{pmatrix}. \tag{18}$$

Applying the Schur complement and focusing on the global parameters (see Supplement D.2 for more details), we find

$$\frac{d\hat{\eta}_\gamma(\varepsilon)}{d\varepsilon}\bigg|_{\varepsilon=0} = -\hat{H}_\gamma^{-1}\hat{J}_\gamma \quad \text{where} \quad \hat{H}_\gamma := \left(H_{\gamma\gamma} - H_{\gamma\ell}H_{\ell\ell}^{-1}H_{\ell\gamma}\right), \tag{19}$$

In the models we consider, $H_{\ell\ell}$ is block diagonal, and the size of $H_{\gamma\gamma}$ is relatively small. Thus each term of (19) can be tractably computed, even on very large datasets. While the Schur complement calculation is illustrative, (19) is equivalent to applying automatic differentiation to the global-only objective $\mathrm{KL}_{\mathrm{glob}}(\eta_\gamma, \varepsilon)$; see Supplement D.2 for details.

In our BNP applications, it is not cost-effective to form and invert or factorize $\hat{H}$ in memory. Instead, we numerically solve linear systems of the form $\hat{H}^{-1}v$ using the conjugate gradient (CG) algorithm (Nocedal and Wright, 2006, Chapter 5), which requires only Hessian-vector products that are readily available through automatic differentiation.

A linear approximation only in the global variational parameters. With the tools above, we can separate out the linear approximation in the global parameters and then directly compute the local parameters. In particular, we compute

$$\hat{\eta}_\gamma^{\text{lin}}(\varepsilon) := \hat{\eta}_\gamma + \left.\frac{d\hat{\eta}_\gamma(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} \varepsilon, \tag{20}$$

and then use $\hat{\eta}_\ell(\eta_\gamma)$ e.g. in computing our quantity of interest. By doing so, our approximation is able to retain non-linearities in the map $\eta_\gamma \mapsto \hat{\eta}_\ell(\eta_\gamma)$. We give an example for the expected number of clusters in Supplement D.3. In all our experiments, we use (20) in this way.

7 Experimental Results

We next evaluate our sensitivity approximations on three real data sets, each with a different model using stick-breaking.[3] We find that our approximations largely agree with ground truth obtained by re-running the VB optimization, but with the evaluation of our derivative an order of magnitude faster than re-optimizing for a given perturbation.

7.1 Gaussian Mixture Modeling on Iris Data

We perform a clustering analysis of Fisher's iris data set (Fisher, 1936; Anderson, 1936). Here each data point (with $N = 150$ total points) represents $d = 4$ measurements of a particular flower, from one of three iris species. We use a standard Gaussian mixture model with a conjugate Gaussian-Wishart prior for the component parameters (detailed in Supplement E.2) and a mean-field VB approximation with truncation parameter $K_{\max} = 15$. We consider two quantities of interest: (1) g_{cl}, the posterior expected number of clusters among the N observed data points, and (2) $g_{\text{pred,cl}}$, the posterior predictive expected number of clusters in N new (i.e. as-yet-unseen) data points. We set the base stick-breaking prior $\mathcal{P}_0(\nu_k)$ to be the standard Beta $(\nu_k|1, \alpha)$ distribution with $\alpha = \alpha_0 = 2$. Under the base stick-breaking prior with α_0, the posterior expected number of clusters matches the three iris species; see also Figure 13 in Supplement E.2 for an illustration.

Sensitivity to the concentration parameter. We approximate the changes in the quantities of interest as α varies over $\alpha \in [0.1, 4.0]$, which corresponds to an *a priori* expected number of clusters among N data points in $[1.5, 15]$ (Supplement E.1). Over this range, the shape of a Beta $(1, \alpha)$ density varies considerably, as shown in Figure 12 in Supplement E.1.

[3]Code and instructions for reproducing our experiments can be found online at https://github.com/Runjing-Liu120/BNPStickBreakingSensitivity.

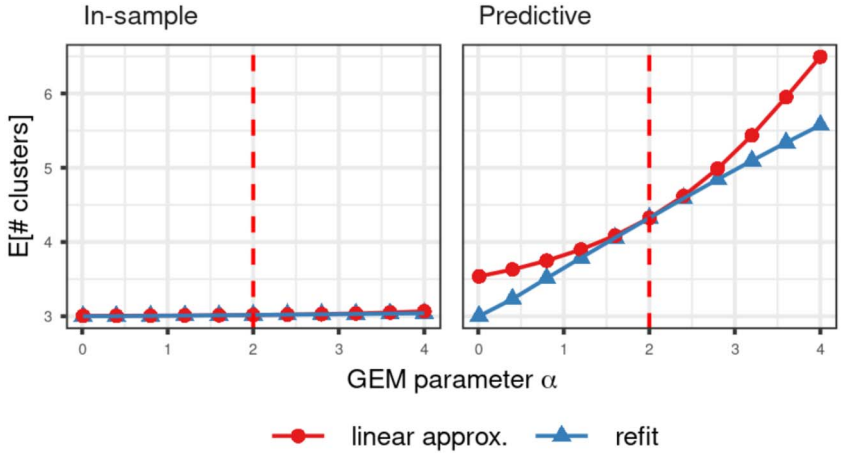

Figure 2: The expected number of clusters in the original data set (g_{cl}, left) and in a new data set of size N ($g_{\text{pred,cl}}$, right) as α varies in the fit of the iris data. We formed the linear approximation at $\alpha_0 = 2$.

Figure 2 compares our linear approximation to ground truth on the two quantities of interest as α varies. Over this range of α, the posterior expected number of clusters in the observed data is quite robust; it remains nearly constant at three. The posterior predictive expected number of clusters in N new data points is less robust; it ranges roughly from 3.0 to 5.6 expected species. Our approximation captures this qualitative behavior. As expected, the approximation is least accurate furthest from the α_0, where the Taylor series is centered.

Sensitivity to functional perturbations. Insensitivity of the expected number of clusters g_{cl} to α does not rule out sensitivity to other prior perturbations. We now check how our approximation fares for the multiplicative perturbations in (7). We consider perturbations ϕ that are Gaussian bumps in logit stick space, with each perturbation centered at a different location on the real line. Each row of Figure 3 corresponds to a different ϕ. Each ϕ is shown in gray in the leftmost plot of its row. The middle column of Figure 3 shows the stick-breaking prior $\mathcal{P}(\nu_k|\phi)$ induced by the corresponding ϕ. The rightmost column of Figure 3 shows the changes produced by the ϕ perturbation for that row. We see that our approximation captures the qualitative behavior of the exact changes.

We also see in this example that we can use the influence function to predict the effect of functional changes to the stick-breaking prior. In the leftmost column, we plot in purple the influence function in the logit space.[4] According to Corollary 1, the sign and magnitude of the effect of a perturbation should be determined by its integral against

[4]Corollary 1 expresses the influence function in the stick domain $[0, 1]$, but, for visualization, it is preferable to express the influence function in the logit stick domain $\mathbb{R}$. The more general Corollary 3 in Supplement A.3 accommodates such transformations.

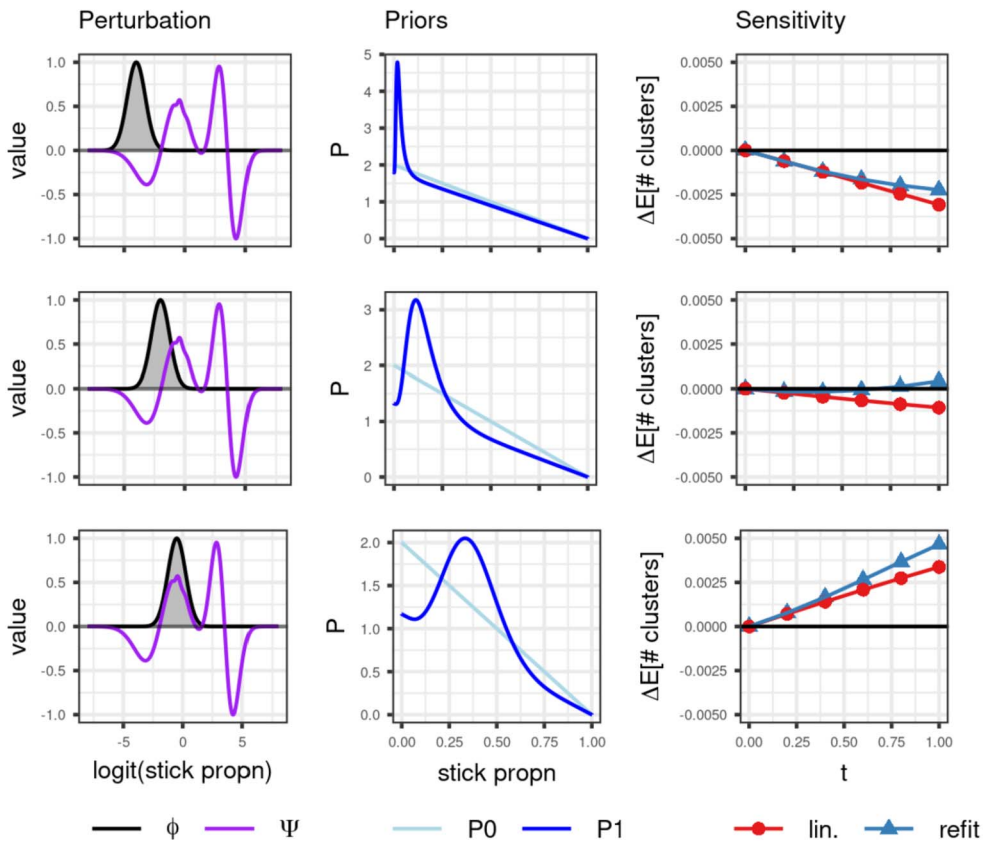

Figure 3: Sensitivity of the expected number of in-sample clusters in the iris data set to three multiplicative perturbations each with $\|\phi\|_\infty = 1$. (Left) The multiplicative perturbation ϕ is in grey. The influence function Ψ, scaled so $\|\Psi\|_\infty = 1$, is in purple. (Middle) The initial $\mathcal{P}_0(\nu_k)$ (light blue) and alternative $\mathcal{P}_1(\nu_k)$ (dark blue) priors. (Right) The effect of the perturbation on the change in expected number of in-sample clusters for $t \in [0, 1]$.

the influence function. Thus, when ϕ lines up with a negative part of Ψ, as in the first row, we expect the change to be negative. Similarly, we expect the perturbation of the bottom row to produce a positive change, and the middle row, in which ϕ overlaps with both negative and positive parts of the influence function, to produce a relatively small change. We see this intuition borne out in the rightmost column.

Worst-case functional perturbation. Finally, Figure 4 shows the worst-case multiplicative perturbation with $\|\phi\|_\infty = 1$, as given by Corollary 2, along with its effect on the prior and g_{cl}. As expected, this worst-case perturbation has a much larger effect on g_{cl} compared to the other unit-norm perturbations in Figure 3. However, even with the worst-case perturbation—which results in an unreasonably shaped prior density—the

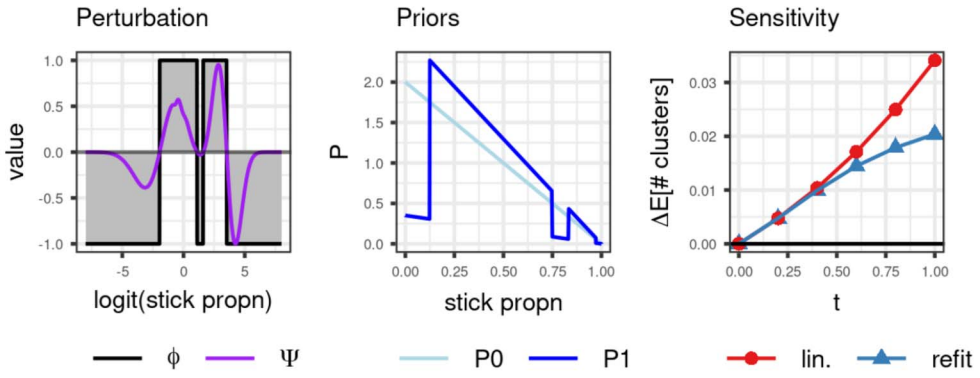

Figure 4: Sensitivity of the expected number of in-sample clusters in the iris data set to the worst-case multiplicative perturbation with $\|\phi\|_\infty = 1$.

change in g_{cl} is still small. We conclude that g_{cl} appears to be a robust quantity for this model and dataset.

7.2 Regression Mixture Modeling

We next check our approximation on a more complex clustering task: clustering time series, with a co-clustering matrix (and summaries thereof) as the quantity of interest.

Data and model. We use a publicly available data set of mice gene expression (Shoemaker et al., 2015). Mice were infected with influenza virus, and expression levels of a set of genes were assessed at 14 time points after infection. Three measurements were taken at each time point (called biological replicates), for a total of $M = 42$ measurements per gene.

The goal of the analysis is to cluster the time-course gene expression data under the assumption that genes with similar time-course behavior may have similar function. Clustering gene expressions is often used for exploratory analysis and is a first step before further downstream investigation. It is important, therefore, to ascertain the stability of the discovered clusters.

The left plot of Figure 14 in Supplement E.3 shows the measurements of a single gene over time. We model each gene as belonging to a latent component, where each component defines a smooth expression curve over time. Then, observations are drawn by adding i.i.d. noise to the smoothed curve along with a gene-specific offset. Following Luan and Li (2003), we construct the smoothers using cubic B-splines.

Let $x_n \in \mathbb{R}^M$ be measurements of gene n at M time points. Let A be the $M \times d$ B-spline regressor matrix, so that the ij-th entry of A is the j-th B-spline basis vector evaluated at the i-th time point. The right plot of Figure 14 in Supplement E.3 shows

the B-spline basis. The distribution of the data arising from component k is

$$\mathcal{P}(x_n|\beta_k, b_n) = \mathcal{N}\left(x_n|A\mu_k + b_n, \tau_k^{-1} I_{M\times M}\right), \tag{21}$$

where b_n is a gene-specific additive offset and I is the identity matrix. We include the additive offset because we are interested in clustering gene expressions based on their patterns over time, not their absolute level. In this model, the component-specific parameters are $\beta_k = (\mu_k, \tau_k)$, the regression coefficients and the inverse noise variance. The component frequencies are determined by stick-breaking according to ν, and cluster assignments z are drawn as in Section 3.1.

Our variational approximation factorizes similarly to (2) except with an additional factor for the additive shift. In our variational approximation, we also make a simplification by letting $\mathcal{Q}(\beta_k|\eta) = \delta(\beta_k|\eta)$, where $\delta(\cdot|\eta)$ denotes a point mass at a parameterized location. See Supplement E.3 for further details concerning the model and variational approximation.

Quantity of interest: the co-clustering matrix and summaries. In this application, we are particularly interested in which genes cluster together, so we focus on the posterior co-clustering matrix. Let $g_{\text{cc}}(\eta) \in \mathbb{R}^{N\times N}$ denote the matrix whose (i,j)-th entry is the posterior probability that gene i belongs to the same cluster as gene j, given by

$$[g_{\text{cc}}(\eta)]_{ij} = \mathop{\mathbb{E}}_{\mathcal{Q}(z|\eta)}\left[\mathbb{I}\left(z_i = z_j\right)\right] = \begin{cases} \sum_{k=1}^{K_{\max}} \left(\mathop{\mathbb{E}}_{\mathcal{Q}(z_i|\eta)}\left[z_{ik}\right] \mathop{\mathbb{E}}_{\mathcal{Q}(z_j|\eta)}\left[z_{jk}\right] \right) & \text{for } i \neq j \\ 1 & \text{for } i = j. \end{cases}$$

Figure 5 shows the inferred co-clustering matrix at α_0.

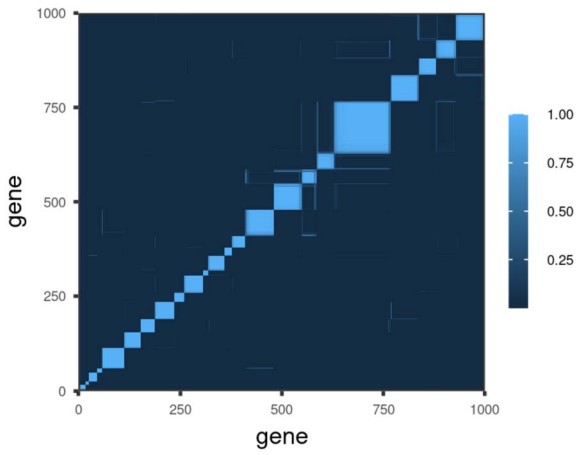

Figure 5: The inferred co-clustering matrix of gene expressions at $\alpha_0 = 6$.

Below, we will use the influence function (Corollary 2) to try and find a perturbation that produces large changes in the co-clustering matrix. To compute the worst-case

perturbation, we must choose a univariate summary of the N^2-dimensional co-clustering matrix whose derivative we wish to extremize. We use the sum of the eigenvalues of the symmetrically normalized graph Laplacian, as given by

$$g_{\text{ev}}(\eta) = \text{Tr}\left(I - D(\eta)^{-1/2} g_{\text{cc}}(\eta) D(\eta)^{-1/2}\right),$$

where $D(\eta)^{-1/2}$ is the diagonal matrix with entries $d_i = \sum_{j=1}^{N} [g_{\text{cc}}(\eta)]_{ij}$. The quantity g_{ev} is differentiable, and has close connection with the number of distinct components in a graph (von Luxburg, 2007). We expect that prior perturbations that produce large changes in g_{ev} will also produce large changes in the full co-clustering matrix.

Sensitivity to the concentration parameter. We first evaluate the sensitivity of the co-clustering matrix g_{cc} to the choice of α in the stick-breaking prior.

We start at $\alpha = \alpha_0 = 6$. We use the linear approximation to extrapolate the co-clustering matrix under prior parameters $\alpha = 0.1$ and $\alpha = 12$. The *a priori* expected number of clusters in the original data at these values is 2 and 50, respectively. Despite this wide prior range, the change in the posterior co-clustering matrix for each α is minuscule (Figure 6). The largest absolute changes in the co-clustering matrix are of order 10^{-2}. Refitting the approximate posterior at $\alpha = 0.1$ and $\alpha = 12$ confirms the insensitivity predicted by the linearized variational global parameters. Beyond capturing insensitivity, the linearized parameters were also able to capture the sign and size of the changes in the individual entries of the co-clustering matrix, even though these changes are small.

Sensitivity to functional perturbations. We now investigate sensitivity of the co-clustering matrix to deviations from the beta prior. In Figure 7, we use the influence function for g_{ev} to construct a nonparametric prior perturbation that we expect to have a large, positive effect. The resulting prior does indeed produce changes an order of magnitude larger than those produced by the perturbations to α shown in Figure 6, and our approximation is again able to capture the qualitative changes. The influence function is also able to explain why α perturbations were unable to produce large changes in this case: Figure 8 shows that changing α (as in Example 3) induces large changes in the prior only where the influence function is small.

However, even with the (unreasonable-looking) selected functional perturbation, the size of the differences in the co-clustering matrix remains modest. It is unlikely that any scientific conclusions derived from the co-clustering matrix would have changed after the functional perturbation. The co-clustering matrix appears robust to perturbations in the stick-breaking distribution.

7.3 Genetic Admixture Modeling with fastSTRUCTURE

Our final analysis illustrates the use of our approximation for stick-breaking priors beyond clustering; namely, in topic modeling.

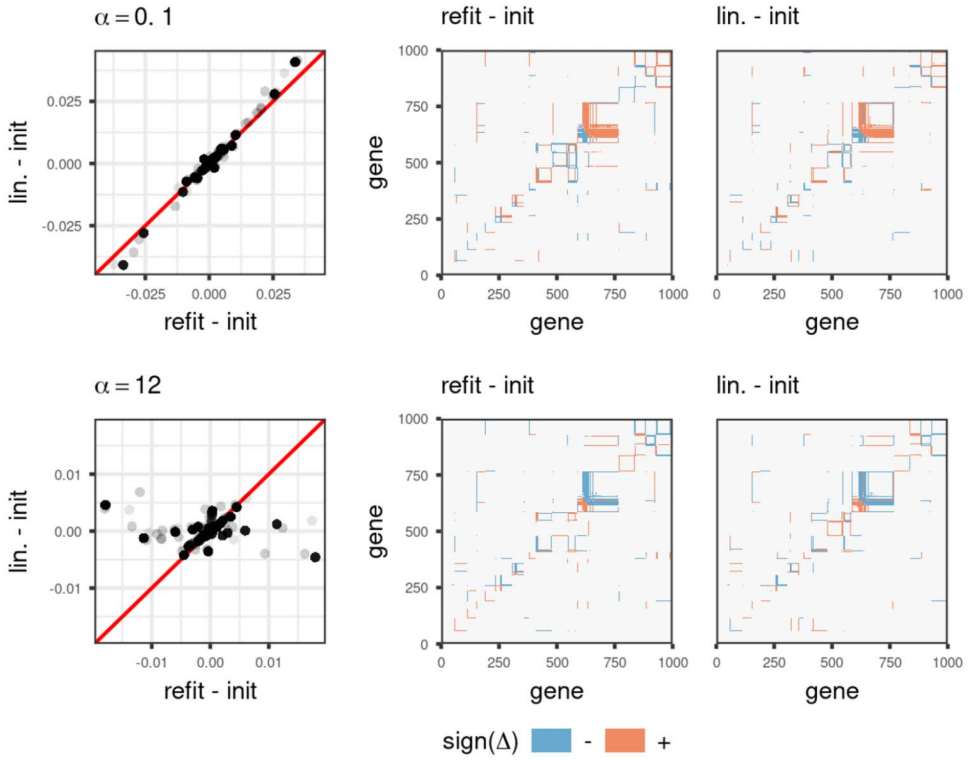

Figure 6: Differences in the co-clustering matrix at $\alpha = 0.1$ (top row) and $\alpha = 12$ (bottom row), relative to the co-clustering matrix at $\alpha_0 = 6$. (Left) A scatter plot of differences under the linear approximation against differences after refitting. Each point represents an entry of the co-clustering matrix. Note the scales of the axes: the largest change in an entry of the co-clustering matrix is ≈ 0.03. (Middle) Sign changes in the co-clustering matrix observed after refitting, ignoring the magnitude of the change. (Right) Sign changes under the linearly approximated variational parameters. For visualization, changes with absolute value $< 10^{-5}$ are not colored.

Data and model. We use a publicly available dataset that contains genotypes from $N = 155$ individuals of an endangered bird species, the Taita thrush (Galbusera et al., 2000). Individuals were collected from four regions in southeast Kenya (Chawia, Mbololo, Ngangao, Yale), and each individual was genotyped at $L = 7$ micro-satellite loci. The four regions were once part of a cohesive cloud forest that has been fragmented by human development. For this endangered bird species, understanding the degree to which populations have grown genetically distinct is important for conservation efforts: well-separated populations with little genetic diversity are particularly at risk of extinction. The goal of the analysis is to infer the population of origin for specific loci and estimate the degree to which populations are admixed in each individual.

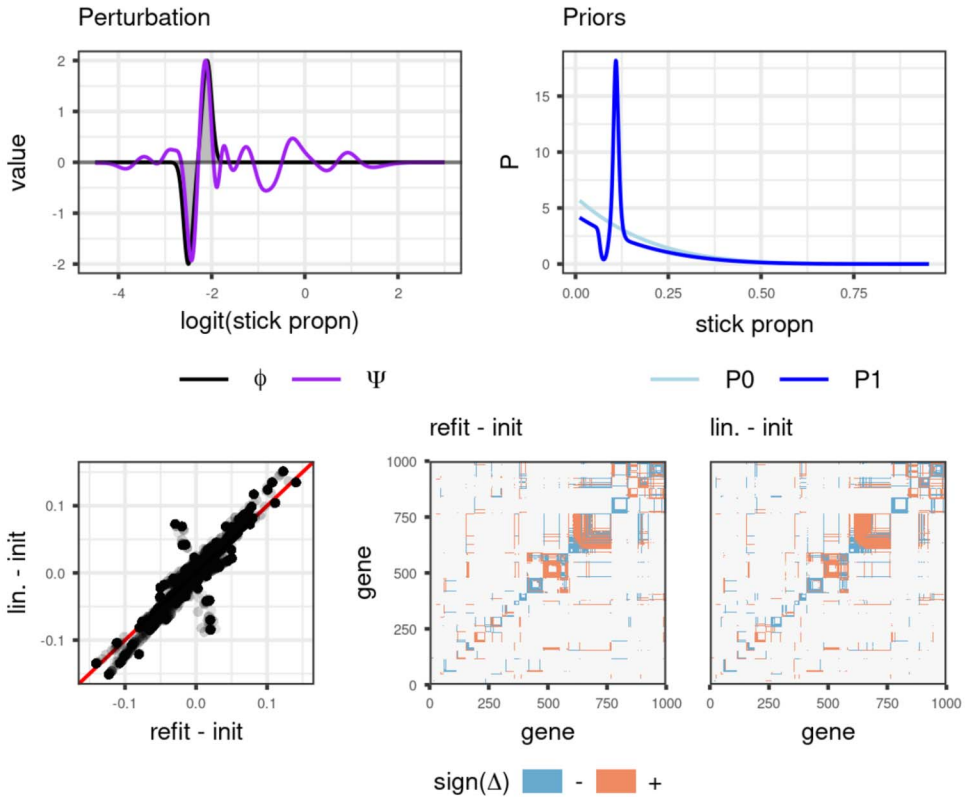

Figure 7: Effect on the co-clustering matrix of a multiplicative functional perturbation. (Top left) The perturbation ϕ is in grey, and the influence function is in purple. (Top right) The effect of this perturbation on the prior density. (Bottom) The effect of this perturbation on the co-clustering matrix. Note the scale of the scatter plot axes compared with the scatter plots in Figure 6.

Let $x_{nli} \in \{1, \ldots, J_l\}$ be the observed genotype for individual n at locus l and chromosome i. J_l is the number of possible genotypes at locus l. For example, if the measurements are all single nucleotides (A, T, C or G) then $J_l = 4$ for all l.

A latent population is characterized by the collection $\beta_k = (\beta_{k1}, \ldots, \beta_{kL})$, where $\beta_{kl} \in \Delta^{J_l - 1}$ are the latent frequencies for the J_l possible genotypes at locus l. Let z_{nli} be the assignment of observation x_{nli} to a latent population. Notice that for a given individual n, different loci (or even different chromosomes at a given locus) may have different population assignments. The distribution of $x_{nli} \in \{1, \ldots, J_l\}$ arising from population k is $\mathcal{P}(x_{nli}|\beta_k) = \text{Categorical}\,(x_{nli}|\beta_{kl})$.

Unlike the previous models, we now have a stick-breaking process for each individual.

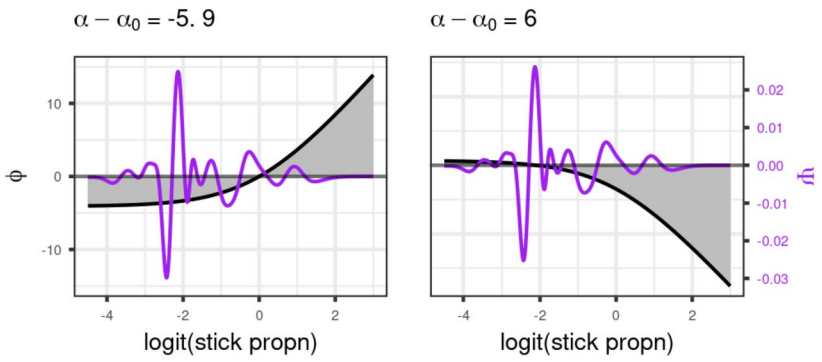

Figure 8: The multiplicative perturbations $\phi_\alpha(\cdot)$ that corresponds to decreasing (left) or increasing (right) the α parameter.

Draw sticks

$$\nu_{nk} \overset{indep}{\sim} \mathcal{P}_{\text{stick}}(\nu_{nk}), \quad n = 1, \ldots, N; k = 1, 2, \ldots.$$

The prior assignment probability vector $\pi_n = (\pi_{n1}, \pi_{n2}, \ldots)$, now unique to each individual, is formed by the same stick-breaking construction as before,

$$\pi_{nk} = \nu_{nk} \prod_{k'<k} (1 - \nu_{nk'}).$$

The population assignment z_{nli} is drawn from a multinomial distribution

$$p(z_{nli}|\pi_n) = \prod_{k=1}^{\infty} \pi_{nk}^{z_{nlik}}.$$

In this genetics application, we call π_n the *admixture* of individual n.

Initially we take $\mathcal{P}_{\text{stick}}$ to be Beta $(1, \alpha)$ with parameter $\alpha = \alpha_0 = 3$. The choice of $\alpha_0 = 3$ corresponds to roughly four distinct populations *a priori*, in agreement with the observation that the individuals come from four geographic regions. Below, we will evaluate sensitivity to this prior choice.

This model is identical to fastSTRUCTURE, a model proposed in Pritchard et al. (2000) and Raj et al. (2014), except that we replace the Dirichlet prior in fastSTRUCTURE with an infinite stick-breaking process. The result is a model similar to a hierarchical Dirichlet process for topic modeling (Teh et al., 2006), but without the top-level Dirichlet process. In addition, genotypes at genetic markers take the place of words in a document; in lieu of inferring "topics," we infer latent populations.

We use a mean-field variational approximation, and all distributions are conditionally conjugate except for the stick-breaking proportions, which remain logit-normal. See Supplement E.4 for further details.

Quantity of interest. The posterior quantities of interest in this application are the admixtures π_n. Figure 16 plots the inferred admixtures $\mathop{\mathbb{E}}_{\mathcal{Q}(\pi_n|\hat{\eta})}[\pi_n]$ for all individuals n.

In the approximate posterior with α_0, there appear to be three dominant latent populations, which we arbitrarily label as populations 1, 2, and 3 (top panel of Figure 9). The inferred admixture proportions generally correspond with geographic regions: Mbololo individuals are primarily population 1, Ngangao individuals are primarily population 2, and Chawia individuals are a mixture of populations 1, 2, and 3 (Figure 16 in Supplement E.4).

Notably, outlying admixtures among individuals from the same geographic region provide clues into the historical migration patterns of this species. For example, while most Mbololo individuals are dominantly population 1, several Mbololo individuals have abnormally large admixture proportions of population 2. Conversely, while most Ngangao individuals are dominantly population 2, several Ngangao individuals have abnormally large admixture proportions of population 1. These patterns suggest that some migration has occurred between the Mbololo and Ngangao regions.

We evaluate the sensitivity of this conclusion to possible prior perturbations. Define the posterior quantity

$$g_{\text{admix}}(\eta; \mathcal{N}, k) = \mathop{\mathbb{E}}_{\mathcal{Q}(\pi|\eta)}\left[\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}}\pi_{nk}\right],$$

the average admixture proportion of population k in a set of individuals $\mathcal{N}$.

Below, we consider g_{admix} with three different sets of individuals: $\mathcal{N} = \{26, \ldots, 31\}$, corresponding to the outlying Mbololo individuals, labeled "A" in Figure 9; $\mathcal{N} = \{125, \ldots, 128\}$, corresponding to the four outlying Ngangao individuals, labeled "B"; and $\mathcal{N} = \{139, \ldots, 155\}$ corresponding to all Chawia individuals, labeled "C". For individuals A, we let $k = 2$ in g_{admix} and examine the robustness of the presence of population 2; for individuals B, we use $k = 1$; and for individuals C, we use $k = 3$. The first two posterior quantities relate to the inferred migration between the Mbololo and Ngangao regions. In the last example, we study the robustness of having a third latent population present, a population that primarily appears in Chawia individuals.

Functional sensitivity. We construct worst-case negative perturbations for each of the three variants of g_{admix}, in order to see whether the biologically interesting patterns can be made to disappear with different prior choices. Figure 9 shows the result of the worst-case perturbations on the prior density and g_{admix}. After the worst-case perturbation, the admixture proportion of population 2 in individuals A was nearly halved. On the other hand, the admixture of population 1 in individuals B is more robust. We conclude that the inferred migration from Ngangao to Mbololo is relatively robust to the stick-breaking prior, while conclusions about migration from Mbololo to Ngangao may be dependent on prior choices.

In this data set and model, the conclusions from the linear approximation did not always agree with the conclusions from refitting the variational approximation. For

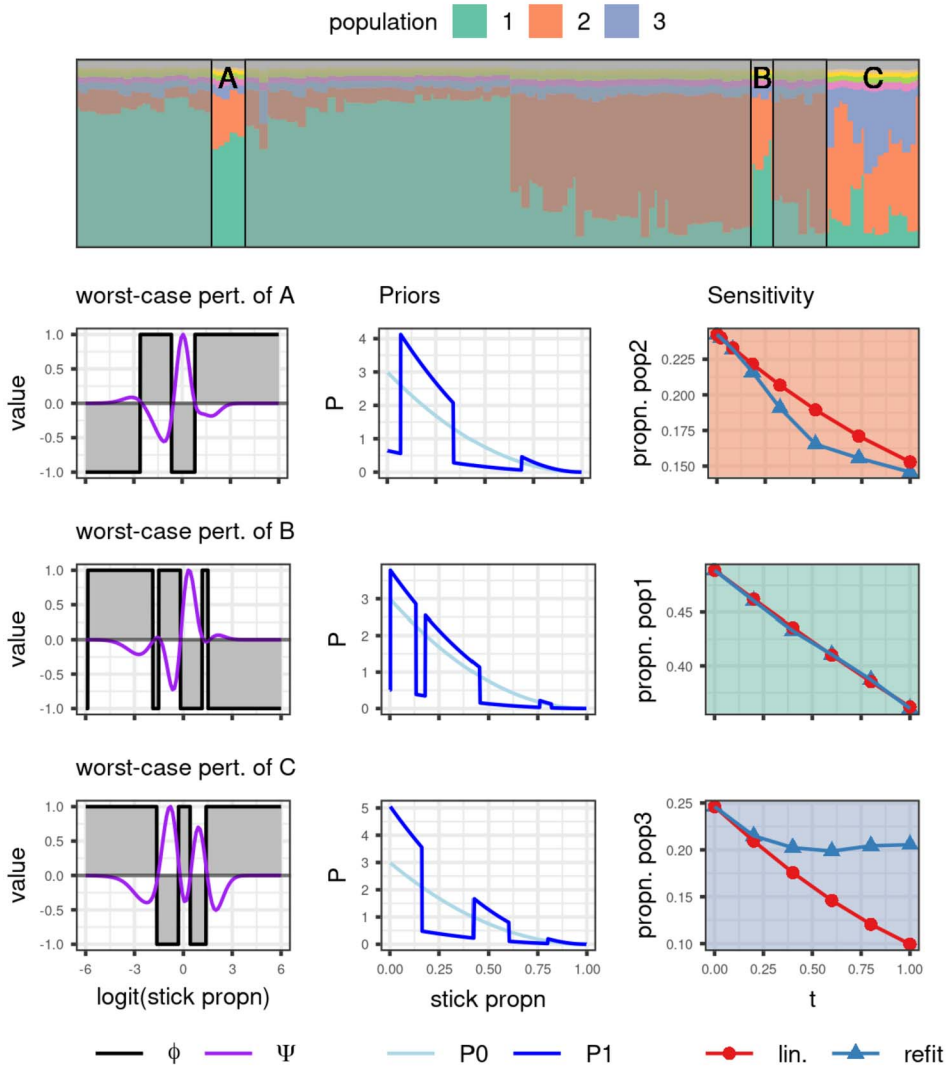

Figure 9: Sensitivity of inferred admixtures for several outlying individuals. For individuals A, we examine the sensitivity of the admixture proportion of population 2. For individuals B, we examine the population 1 admixture. For the individuals C, we examine the population 3 admixture. (Left column) The worst-case negative perturbation with $\|\phi\|_\infty = 1$ in grey, plotted against the influence function in purple (scaled such that $\|\psi\|_\infty = 1$). (Middle column) The effect of the perturbation on the prior density. (Right column) Effects on the inferred admixture.

example, the admixture proportion of population 3 in individuals C were predicted to more sensitive by our linear approximation than were actually observed after refitting

(Figure 9, bottom row).

Moreover, even though the linear approximation agreed with the refits for individuals A in overall admixture proportion (Figure 9, second row), the approximation does not perform uniformly well over all individuals. Figure 10 plots the inferred admixtures computed using the linearized variational parameters and the refitted variational parameters. The admixture proportion of population 2 in individual $n = 25$ dramatically increased after refitting with the perturbed prior; the linearized parameters failed to reproduce this change.

Even though linear approximation works less well in this example, the influence function is still able to guide our choice of functional perturbations at which to refit. While the worst-case perturbations we used may be an adversarial choice, the influence function suggests that we can construct a smoother perturbation with a similar effect as the worst-case, as we did in Section 7.2. Importantly, as we will note in the next subsection, the influence function is cheap to compute relative to refitting. For a further discussion of the limitations of the linear approximation, see Supplement F.

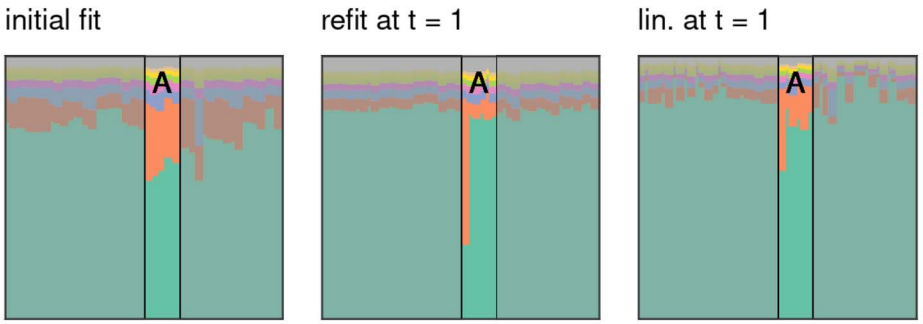

Figure 10: Inferred admixtures after the worst-case perturbation to individuals "A" (see Figure 9 for perturbation).

7.4 Computation Time

The relative computational costs of the approximation and re-fitting for our three experiments are shown in Table 1. The data sets we considered in our experiments had varying degrees of complexity, and the computational of cost of fitting the variational approximation thus also varies accordingly. However, the cost of forming the linear approximation—the step that requires computing and inverting the Hessian matrix—was consistently roughly an order of magnitude faster than refitting.

Recall from Section 6 that the solution of a linear system involving $\hat{H}^{-1}$ is the computationally intensive part of the linear approximation, and that the linear system needs to be solved only once for a given perturbation, as described in Section 6. Consistent with this observation, in all the examples, after the linear approximation is formed, extrapolating to any new prior parameter $\alpha \neq \alpha_0$ or $t \neq 0$ takes only fractions of a

	iris	mice	thrush
Initial fit	1	30	7
Hessian solve for α sensitivity	0.02	3	0.3
Linear approx. $\hat{\eta}^{\text{lin}}(\alpha)$	0.0008	0.001	0.0008
Refits $\hat{\eta}(\alpha)$	0.5	30	5
The influence function (at 1000 grid points)	0.09	3	0.6
Hessian solve for ϕ	0.02	3	0.4
Linear approx. $\hat{\eta}^{\text{lin}}(\phi)$	0.001	0.001	0.0008
Refit $\hat{\eta}^{\text{lin}}(\phi)$	0.6	20	10

Table 1: Compute time in seconds of various quantities on each data set. Reported times for $\hat{\eta}(\alpha)$ and $\hat{\eta}^{\text{lin}}(\alpha)$ are median times over the set of considered α's. The reported influence function time is the time required to evaluate the influence function on a grid of 1000 points.

second. For example, in the thrush data and fastSTRUCTURE model, the initial fit took seven seconds, with subsequent refits (which we warm-started with the initial fit) taking between five and ten seconds. Solving a linear system to form the linear approximation for a particular perturbation ϕ took less than a second, and evaluating $\hat{\eta}(\phi)$ was essentially free.

8 Conclusion

We provide a method to approximate the effect of changing a BNP prior on a posterior quantity of interest in a VB approximation. Our method is generally applicable, straightforward to implement, and computationally efficient. In our experiments, we show by refitting that the predictions of the approximation are typically qualitatively accurate. Over the range of situations and quantities of interest we considered, we discovered robustness (co-clustering in the mice dataset), non-robustness (the predictive number of clusters in the iris dataset), and intermediate cases where some closely related posterior quantities were robust and others not (the population memberships of the thrush dataset). Given such a variety of outcomes, the authors hesitate to draw any generalizable conclusions about the robustness of DPM models, much less generic BNP posteriors. On the contrary, we hope that our results motivate users to check for robustness directly, for their particular datasets and models of interest, rather than trying to rely too heavily on intuition or general principles. Indeed, we hope that the ease with which one can compute our prior sensitivity measures, combined with the possibility of uncovering materially important non-robustness, will encourage the widespread adoption of routine prior robustness checks.

In the present work, we have focused only on the task of detecting and characterizing non-robustness. We have not attempted to address the critical questions of what to do when a conclusion is non-robust, nor how to make robust modeling choices. We hope that routine, widespread robustness checking in a variety of real-world problems will

further inform and motivate solutions to the important question of how to deal with, and even prevent, non-robustness in practice.

Though the need for prior robustness checking seems particularly well-motivated in discrete BNP problems, where the DPM prior is often chosen for computational convenience rather than considered subjective belief, prior robustness checks are relevant to almost all Bayesian analysis. Despite our present focus on BNP models and the DPM prior in particular, our methodology extends not only to other truncated approximations of discrete BNP priors, but to any VB approximation based on reverse KL divergence. The best evidence for the usefulness of our methodology of linear approximation will come from widespread adoption and verification in many different applications and modeling environments, and we hope that the present work is only the beginning.

Supplementary Material

Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics: Supplementary Materials (DOI: 10.1214/22-BA1309SUPP; .pdf). The supplementary materials contain detailed proofs, extended theoretical discussion, and more details on the experiments.

References

Ambrogioni, L., Güçlü, U., Güçlütürk, Y., Hinne, M., van Gerven, M., and Maris, E. (2018). "Wasserstein Variational Inference." *Advances in Neural Information Processing Systems*, 31: 2473–2482. 291

Anderson, E. (1936). "The species problem in iris." *Annals of the Missouri Botanical Garden*, 23(3): 457–509. 303

Averbukh, V. and Smolyanov, O. (1967). "The theory of differentiation in linear topological spaces." *Russian Mathematical Surveys*, 22(6): 201–258. MR0223886. 299, 300

Barrios, E., Lijoi, A., Nieto-Barajas, L., and Prünster, I. (2013). "Modeling with normalized random measure mixture models." *Statistical Science*, 28(3): 313–334. MR3135535. doi: https://doi.org/10.1214/13-STS416. 291

Basu, S. (2000). *Bayesian Robustness and Bayesian Nonparametrics*, 223–240. New York, NY: Springer New York. MR1795218. doi: https://doi.org/10.1007/978-1-4612-1306-2_12. 291

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). "Local posterior robustness with parametric priors: Maximum and average sensitivity." In *Maximum Entropy and Bayesian Methods*, 97–106. Springer. 291

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2018). "Automatic differentiation in machine learning: A survey." *Journal of Machine Learning Research*, 18. MR3800512. 296

Blei, D. and Jordan, M. I. (2006). "Variational inference for Dirichlet process mixtures." *Bayesian Analysis*, 1(1): 121 – 143. MR2227367. doi: https://doi.org/10.1214/06-BA104. 292

Blei, D., Kucukelbir, A., and McAuliffe, J. (2017). "Variational inference: A review for statisticians." *Journal of the American Statistical Association*, 112(518): 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 293

Campbell, T., Huggins, J., How, J., and Broderick, T. (2019). "Truncated random measures." *Bernoulli*, 25(2): 1256–1288. MR3920372. doi: https://doi.org/10.3150/18-bej1020. 291

Canale, A., Lijoi, A., Nipoti, B., and Prünster, I. (2017). "On the Pitman–Yor process with spike and slab base measure." *Biometrika*, 104(3): 681–697. MR3694590. doi: https://doi.org/10.1093/biomet/asx041. 291

Cook, D. (1986). "Assessment of local influence." *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2): 133–155. MR0867994. 290

Doshi, F., Miller, K., Van Gael, J., and Teh, Y. (2009). "Variational inference for the Indian buffet process." In *Artificial Intelligence and Statistics*, 137–144. PMLR. 291

Dudley, R. (2018). *Real Analysis and Probability*. CRC Press. MR1932358. doi: https://doi.org/10.1017/CBO9780511755347. 299

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 209–230. MR0350949. 287

Fisher, R. (1936). "The use of multiple measurements in taxonomic problems." *Annals of eugenics*, 7(2): 179–188. 303

Galbusera, P., Lens, L., Schenck, T., Waiyaki, E., and Matthysen, E. (2000). "Genetic variability and gene flow in the globally, critically-endangered Taita Thrush." *Conservation Genetics*, 1: 45–55. 310

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. MR3235677. 288, 290

Ghaderinezhad, F. and Ley, C. (2019). "Quantification of the impact of priors in Bayesian statistics via Stein's method." *Statistics & Probability Letters*, 146: 206–212. MR3884714. doi: https://doi.org/10.1016/j.spl.2018.11.012. 290

Giordano, R., Broderick, T., and Jordan, M. I. (2018). "Covariances, robustness and variational Bayes." *Journal of Machine Learning Research*, 19(51). MR3874159. 289, 290, 291

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2022). "Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics: Supplementary Materials." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1309SUPP. 293

Gustafson, P. (1996a). "Local sensitivity of inferences to prior marginals." *Journal of the American Statistical Association*, 91(434): 774–781. MR1395744. doi: https://doi.org/10.2307/2291672. 290

Gustafson, P. (1996b). "Local sensitivity of posterior expectations." *Annals of Statistics*, 24(1): 174–195. 289, 290, 291, 294, 295, 297, 300

Gustafson, P. (2000). *Local Robustness in Bayesian Analysis*, 71–88. New York, NY: Springer New York. 290

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (2011). *Robust Statistics: The Approach Based on Influence Functions*, volume 196. John Wiley & Sons. MR0829458. 290

Insua, D. R. and Ruggeri, F. (2000). *Robust Bayesian Analysis*. Springer. MR1795206. doi: https://doi.org/10.1007/978-1-4612-1306-2. 290

Jaeckel, L. (1972). "The Infinitesimal Jackknife, Memorandum." Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ. 290

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, 20(1): 50 – 67. MR2182987. doi: https://doi.org/10.1214/088342305000000016. 289

Krantz, S. and Parks, H. (2012). *The Implicit Function Theorem: History, Theory, and Applications*. Springer Science & Business Media. MR2977424. doi: https://doi.org/10.1007/978-1-4614-5981-1. 291

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). "Automatic differentiation variational inference." *The Journal of Machine Learning Research*, 18(1): 430–474. MR3634881. 289

Li, Y. and Turner, R. (2016). "Variational inference with Rényi divergence." *stat*, 1050: 6. 291

Lijoi, A., Mena, R., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 291

Liu, Q. and Wang, D. (2016). "Stein variational gradient descent: A general purpose Bayesian inference algorithm." *Advances in Neural Information Processing Systems*, 29: 2378–2386. 291

Luan, Y. and Li, H. (2003). "Clustering of time-course gene expression data using a mixed-effects model with B-splines." *Bioinformatics*, 19(4): 474–482. 307

Nieto-Barajas, L. and Prünster, I. (2009). "A sensitivity analysis for Bayesian nonparametric density estimators." *Statistica Sinica*, 19(2): 685–705. MR2514182. 291

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Science & Business Media. MR2244940. 302, 303

Pritchard, J., Stephens, M., and Donnelly, P. (2000). "Inference of population structure using multilocus genotype data." *Genetics*, 155(2): 945–959. 311

Raj, A., Stephens, M., and Pritchard, J. K. (2014). "fastSTRUCTURE: Variational inference of population structure in large SNP data sets." *Genetics*, 197(2): 573–589. 311

Ranganath, R., Gerrish, S., and Blei, D. (2014). "Black box variational inference." In *Artificial intelligence and statistics*, 814–822. PMLR. 289

Reeds, J. (1976). "On the definition of von Mises functionals." Ph.D. thesis, Statistics, Harvard University. MR2940746. 299

Roos, M., Martins, T., Held, L., and Rue, H. (2015). "Sensitivity analysis for Bayesian hierarchical models." *Bayesian Analysis*, 10(2): 321–349. MR3420885. doi: https://doi.org/10.1214/14-BA909. 290

Roychowdhury, A. and Kulis, B. (2015). "Gamma processes, stick-breaking, and variational inference." In *Artificial Intelligence and Statistics*, 800–808. PMLR. 291

Saha, A. and Kurtek, S. (2019). "Geometric sensitivity measures for Bayesian nonparametric density estimation models." *Sangkhya Series A.*, 81: 104–143. MR3982193. doi: https://doi.org/10.1007/s13171-018-0145-7. 291

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 639–650. MR1309433. 287

Shoemaker, J., Fukuyama, S., Eisfeld, A., Zhao, D., Kawakami, E., Sakabe1, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). "An ultrasensitive mechanism regulates influenza virus-induced inflammation." *PLoS Pathogens*, 11(6): 1–25. 306

Sivaganesan, S. (2000). "Global and local robustness approaches: Uses and limitations." In *Robust Bayesian Analysis*, 89–108. Springer. MR1795211. doi: https://doi.org/10.1007/978-1-4612-1306-2_5. 290

Teh, Y., Jordan, M. I., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 288, 312

Virtanen, P., Gommers, R., Oliphant, T., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S., Brett, M., Wilson, J., Millman, J., Mayorov, N., Nelson, A., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E., Harris, C., Archibald, A., Ribeiro, A., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). "SciPy 1.0: Fundamental algorithms for scientific computing in Python." *Nature Methods*, 17: 261–272. 302

von Luxburg, U. (2007). "A tutorial on spectral clustering." *Statistics and Computing*, 17: 395–416. MR2409803. doi: https://doi.org/10.1007/s11222-007-9033-z. 308

Zeidler, E. (1986). *Nonlinear Functional Analysis and Its Applications I: Fixed point theorems*. Springer Verlag New York, Inc. MR0816732. doi: https://doi.org/10.1007/978-1-4612-4838-5. 299, 300

Acknowledgments

We are indebted to helpful discussions with Nelle Varoquaux, Matthew Stephens, Michael C. Hughes, Eric Sudderth, and Jake Soloff, and to useful suggestions from anonymous reviewers.

Invited Discussion*

Steven N. MacEachern† and Juhee Lee‡

First, we would like to congratulate the authors for their development of a fast and efficient method of assessing sensitivity to the prior specification of the Dirichlet process mixture (DPM) and related models. Their techniques are widely applicable and will see much use. Their examples give us a taste of what can be done with their method in models that move beyond the DPM. We greatly enjoyed reading the paper.

The past 30-plus years have seen incredible growth in the use of Bayesian methods for data analysis. The initial growth was driven by the development of Markov chain Monte Carlo (MCMC) methods that allowed one to fit models of substantial complexity. This was quickly followed by a realization in the entire statistics community that Bayesian methods simply work better than classical methods in "high information" settings with clean data – settings where one can reliably write down a complete Bayesian model and can clearly specify the inference problem. In practice, this translates to settings where different analysts (or the same analyst on different days) would select similar sampling densities for the data and similar prior distributions for the parameters, including latent structures, to arrive at similar models, and would also choose similar loss functions to formalize the inference problem. It also relies on realism in the model, with choices based on an understanding of the phenomenon under study rather than computational convenience. In these settings, the mathematical theory that links optimal inference to Bayesian methods is borne out. Bayesian methods simply work better than methods that are distant from Bayes.

The success of Bayesian methods is not uniform. In "low information" settings, specification of the sampling density and form of the model are every bit as challenging for the Bayesian as for the classical statistician. For high or infinite dimensional models, specification of the prior distribution remains a challenge. Data of dubious quality have the potential to dramatically impact the final inference. MCMC methods are often slow and sometimes exhibit poor convergence. And analysts commonly adjust their model to make fitting it easier and quicker. These difficulties are not shortcomings of the analyst, but rather features of the low information-complex model setting. They set an agenda for research on Bayesian data analysis.

The centerpiece of this agenda is how to improve Bayesian data analysis. Here, we see three main threads. One focuses on diagnostics through the development of techniques to identify deficiencies in the model (whether sampling density, prior distribution or a combination of the two), to identify cases that do not accord with the model (as in outliers), and to identify cases that have a large impact on inference (influential cases).

*Supported by the NSF under grant numbers SES-1921523, DMS-2015428, and DMS-2015552.

†Department of Statistics, The Ohio State University, ORCID 0000-0003-4106-1232, snm@stat.osu.edu

‡Department of Statistics, University of California, Santa Cruz, ORCID 0000-0002-9787-3830, juheelee@soe.ucsc.edu

The second focuses on computational implementation. The third focuses on strategies to improve model specification and inference, with particular attention to robustness in the low information setting.

Giordano et al.'s delightful paper focuses on the first thread and is informed by the second. The authors' insight into the (in)effective use of Bayesian methods shines sharply through their paper. They consider a high/infinite dimensional setting where the prior distribution is specified through a rule-based strategy and where it would be difficult to place full confidence in any chosen rule. In this same setting, variational methods for fitting the model are far quicker than MCMC methods. Variational Bayes (VB) methods also generate the derivatives needed to examine the local sensitivity of features of the posterior distribution to changes in the prior distribution.

The developments in Giordano et al.'s paper parallel the development of local influence as a diagnostic method in classical statistics. Cook (1977)'s initial development of case influence examined the impact of individual cases on inference in the linear model. The main technique was case deletion. It led to Cook's distance, now a standard diagnostic summary in regression. Cook (1986) subsequently extended measures of influence to classical nonlinear models which, in the early-to-mid 1980's, were subject to the difficulties more recently experienced by MCMC methods. The models were slow to fit (via maximum likelihood) and one needed to be concerned with the numerical accuracy of the fits. With no clean analytical form, these difficulties rendered case deletion methods ineffective for the interactive data analysis that was being developed at the time. Instead, Cook turned to local influence, considering infinitesimal perturbations of case weights, and looked for big directional derivatives.

Cook's methods have been extended to the Bayesian setting, initially with pre-MCMC computation (e.g., Johnson and Geisser (1983)) and then with MCMC (e.g., Weiss (1996), Bradlow and Zaslavsky (1997), MacEachern and Peruggia (2000)). The approaches include both full case deletion and local influence (viz. Thomas et al. (2018)), and they cover a variety of inferences, from impact on the full posterior distribution to impact on marginal summaries of the posterior. While these methods are successful for linear models and low-dimensional nonlinear models the techniques are less effective for high-dimensional problems of the sort considered by Giordano et al.

Our first question for the authors is whether the techniques they develop can be adapted to assess local case influence. If so, is such an adaptation computationally feasible? The extension would provide the analyst with an additional tool to identify cases or sets of cases that have a large impact on inference.

Our second question concerns robust forms of the prior distribution. As described in the paper, DPM models are often used for clustering problems. Inferences on clustering, such as the number of clusters and co-clustering probabilities, are influenced by the prior specification. The parameter of the Dirichlet process (DP) may be split into two parts, the total mass parameter α and a base probability measure, $\mathcal{P}_{\text{base}}$. The distribution $\mathcal{P}_{\text{base}}$ generates cluster specific parameters, i.e., $\beta_k \overset{iid}{\sim} \mathcal{P}_{\text{base}}(\beta \mid \xi)$, where the β_k's are cluster specific parameters and ξ is the hyperparameter vector for $\mathcal{P}_{\text{base}}$. Jointly

with α, $\mathcal{P}_{\text{base}}$ influences inference on the clustering structure. For example, Bush et al. (2010) and Lee et al. (2014) studied the joint impact of α and the dispersion of $\mathcal{P}_{\text{base}}$ on posterior inference. In particular, for fixed α and a given dataset of size n, a DPM model tends to produce fewer clusters with larger sizes under more dispersed $\mathcal{P}_{\text{base}}$. In response to this phenomenon, Lee et al. (2014) developed a local-mass preserving prior distribution for Bayesian nonparametric (BNP) models that produces more stable inference for clustering. The central idea is to define "local mass" as the mass assigned by a measure to a region $\mathcal{L}$ of interest in the parameter space Ω_β prior to analysis and to jointly elicit $\mathcal{P}_{\text{base}}$ and α by holding the mass in $\mathcal{L}$ constant. In addition to providing more stable inference about clustering, this form of prior distribution stabilizes inference for other quantities such as estimation of cluster locations.

It would be of great interest to see whether the authors' method can be extended to perform a more comprehensive examination of model sensitivity, including sensitivity to the local mass $\mathcal{L}$. Does the authors' quick and automated tool to assess sensitivity of inference to the specification of α extend to prior distributions with a local mass structure? If so, does one form of prior distribution show greater robustness than the other?

Our final comment returns to data analysis. The standard Bayesian analysis requires a rigorous determination of three components; (i) the sampling distribution (likelihood function) for the observations, (ii) the prior distribution of the parameters and (iii) loss function for making inference (decision). Bayesian analysis strongly depends on the choice of these components, and it is essential to investigate the sensitivity of the procedure to perturbation of all three components. The loss function has traditionally received less attention than the prior distribution and sampling density, as it often has little impact in a low-dimensional parametric setting. However, the choice of the inference (loss) function is critical for BNP models due to their flexibility. For example, DPMs are good at accommodating local features of the data such as outliers. These cases may be captured as one or more clusters that depart from the general pattern of the data. Thus, an inference function that discounts the impact of the outliers on the overall analysis can be more desirable than traditional inference functions (e.g., the quadratic loss function and the 0-1 loss function) for robust decision making (Lee and MacEachern, 2014). The inference function does not "wash out" as the sample size grows. This is similar to the parameter α not washing out for clustering in DPM models. We see scope for the development of inference functions that target a sensible summary of the posterior (or predictive) distribution and that lead to stable inference as the prior and sampling density are varied. Do the authors have a sense of whether a summary such as "number of clusters exceeding a given size" tends to be more robust than the simple "number of clusters"? Do the authors know of systematic ways to create more robust summaries of clusters?

In keeping with the level innovation in this work, the paper opens a host of questions. Those above seem, to us, to walk the tightrope of computational feasibility that leads from the questions we can answer with our written model to those we would answer with infinite resources. We close our discussion here, congratulating the authors on an interesting paper that develops a technique that will undoubtedly see heavy use.

References

Bradlow, E. T. and Zaslavsky, A. M. (1997). Case influence analysis in Bayesian inference. *Journal of Computational and Graphical Statistics*, 6(3):314–331. 322

Bush, C. A., Lee, J., and MacEachern, S. N. (2010). Minimally informative prior distributions for non-parametric Bayesian analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):253–268. MR2830767. doi: https://doi.org/10.1111/j.1467-9868.2009.00735.x. 323

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19(1):15–18. MR0436478. doi: https://doi.org/10.2307/1268249. 322

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 48(2):133–169. MR0867994. 322

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association*, 78(381):137–144. MR0696858. doi: https://doi.org/10.1080/01621459.1983.10477942. 322

Lee, J. and MacEachern, S. N. (2014). Inference functions in high dimensional Bayesian inference. *Statistics and Its Interface*, 7(4):477–486. MR3302376. doi: https://doi.org/10.4310/SII.2014.v7.n4.a5. 323

Lee, J., MacEachern, S. N., Lu, Y., and Mills, G. B. (2014). Local-mass preserving prior distributions for nonparametric Bayesian models. *Bayesian Analysis*, 9(2):307–330. MR3216998. doi: https://doi.org/10.1214/13-BA857. 323

MacEachern, S. N. and Peruggia, M. (2000). Importance link function estimation for Markov chain Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 9(1):99–121. MR1819867. doi: https://doi.org/10.2307/1390615. 322

Thomas, Z. M., MacEachern, S. N., and Peruggia, M. (2018). Reconciling curvature and importance sampling based procedures for summarizing case influence in Bayesian models. *Journal of the American Statistical Association*, 113(524):1669–1683. MR3902237. doi: https://doi.org/10.1080/01621459.2017.1360777. 322

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(4):739–750. MR1410188. 322

Invited Discussion

Jim Griffin* and Maria Kalli†

Giordano *et al.* discuss the use of (local) sensitivity analysis in Bayesian nonparametric (BNP) models applied to classification and clustering. Understanding the sensitivity of BNP models to the choice of hyperparameters is an important topic. Surprisingly little work has been done to date (see the references in the paper for some notable exceptions). Perhaps, this is partly due to a reliance on "large support" arguments for BNP methods which downplay the importance of hyperparameter choices. In practice, conclusions can be affected by the choice of the underlying nonparametric prior and the paper describes tools based on variational Bayes (VB) and automatic differentiation (AD) to build approximations of the posterior distribution under a perturbed prior. The use of AD methods is particularly important in this problem since it avoids the laborious calculation of large numbers of derivatives which have made local sensitivity methods hard to apply to complex models. This work also complements recent work by Jacobi et al. (2018) and Chan et al. (2019) who apply AD to Markov chain Monte Carlo methods applied to parametric models in econometrics (such as vector autoregressions) to understand local sensitivity.

The paper focuses on estimating the number of clusters underlying a data set. It shows that this number can be sensitive to both the choice of the mass/precision parameter in a Dirichlet process or the choice of the "breaks" distribution in a stick-breaking construction. Their method allows the derivation of "worse-case" perturbation distributions, which is important in understanding the extent of sensitivity to these choices.

A key assumption made by the authors is that the distribution $\mathcal{P}_{stick}$ is the same for all ν_k. This assumption has an effect on the number of clusters. Relaxing it so that ν_k depends on k would be an interesting future direction for local sensitivity with stick-breaking processes, perhaps through a simple dependence of the distribution of ν_k on k such as in the Pitman-Yor process (Pitman and Yor, 1997). More generally, the Dirichlet process could be embedded within a more general class of processes such as normalized random measures with independent increments (James et al., 2009) or Gibbs-type priors (Gnedin and Pitman, 2006). These type of priors would allow more flexibility in properties such as the distribution of the number of distinct values (see *e.g.* De Blasi et al., 2015).

For the rest of our discussion, we will consider how some of the methods developed in the paper can be applied to density estimation. This is another important application of Bayesian nonoparametrics methods, in particular, the use of Bayesian nonparametric mixture models. In this case, the posterior distribution can be sensitive to both the

*Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, U. K., j.griffin@ucl.ac.uk

†Department of Mathematics, King's College London, Strand, London, WC2R 2LS, U. K., maria.kalli@kcl.ac.uk

choice of the nonparametric prior and the prior on the mixture components. We focus on the model of Griffin (2010) where a univariate sample $x_1, \ldots, x_n$ is modelled by

$$x_i \sim \mathrm{N}(\mu_i, a\sigma^2), \qquad \mu_i \overset{i.i.d.}{\sim} F, \qquad F \sim \mathrm{DP}(MH),$$

where $\mathrm{DP}(MH)$ represents a Dirichlet process with a concentration/precision parameter M and base measure $H = \mathrm{N}(\mu_0, (1-a)\sigma^2)$. The parameter a controls the flexibility of the marginal distribution of x_i with larger values shrinking the distribution towards a normal distribution.

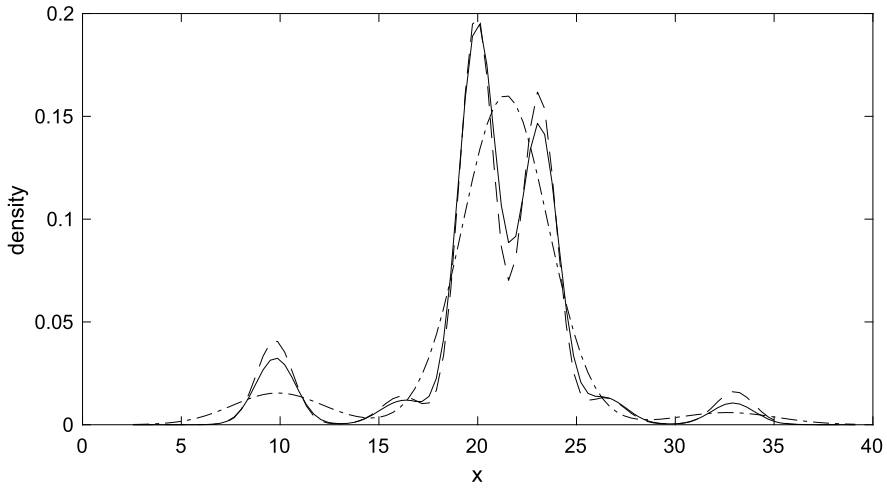

Figure 1: Fitted densities for the galaxy data with $M = 1$ and $a = 0.001$ (dashed line), $a = 0.034$ (solid line), and $a = 0.12$ (dot-dashed line).

This model was fitted using VB to the galaxy data (Escobar and West, 1995) for a range of values of a and M. We find that the fitted density is insensitive to the choice of M but not to the choice of a. To further investigate the effect of a, we plot the fitted densities for three values of a. These are displayed in Figure 1. There is substantial variation in the estimated density, The value of $a = 0.12$ led to a single mode around $x = 21.56$ whereas $a = 0.0001$ and $a = 0.034$ each have two modes at $x = 20.1$ and $x = 23.0$.

Giordano *et al*'s approach can be easily extended to other hyperparameters, such as a, by suitably adapting Theorem 1 (using a suitable choice of θ and $\mathcal{P}$ in the development in the supplementary material) if the regularity conditions are met. In our experience, the main computational overhead of the method is the VB and AD steps for evaluating the second derivatives of the Kullback-Leibler divergences with respect to the variational parameters. However, once these have been calculated, the sensitivity with respect to different hyperparameters is cheap to compute.

We apply this extended approach to understanding local sensitivity to the value of a. We choose a single value of a which is denoted by a_0 and run the VB algorithms 100 times to find local modes. We choose the local mode with the highest value of the Evidence

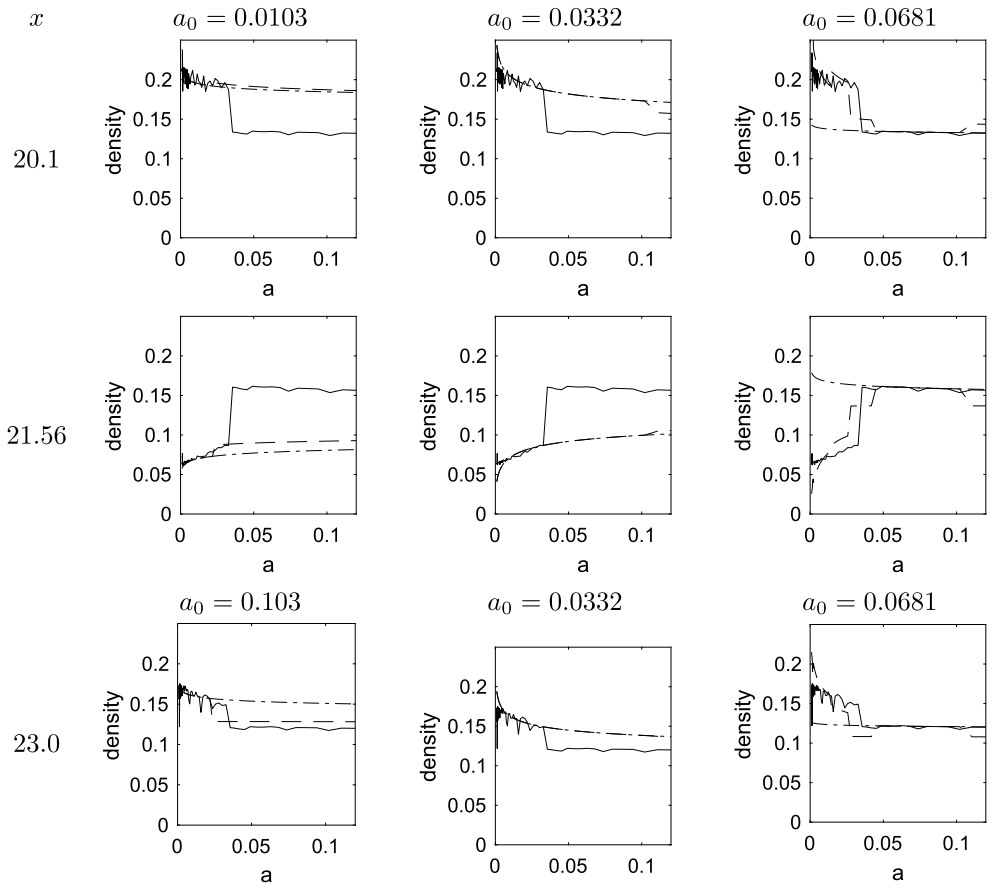

Figure 2: The density at $x = 20.1$, $x = 21.56$ and $x = 23.0$ as a function of a. The lines are: refitting (solid line), approximation using all 100 VB estimates evaluated with a_0 (dashed line), approximation using the optimal VB estimate evaluated with a_0 (dot-dashed line). The values of a_0 are 0.0103, 0.0332 and 0.0681. You can see in all plots a clear change point around 0.035.

Lower Bound (ELBO) as the final VB parameter values. This addresses the large number of local modes that can be found during the VB optimisation process. Figure 2 displays the results of using the approximation for the fitted density at $x = 20.1$, $x = 21.56$ and $x = 23.0$ for different values of a_0 with $M = 1$. These values of x correspond to the two central peaks and the central trough in the density estimate for small values of a. There is a clear change point in the fitted density value around $a = 0.035$. The approximation (shown as the dot-dashed line) is good locally and the fitted density value for $a < 0.035$ can be well-approximated if $a_0 < 0.035$. Similarly, fitted density values for $a > 0.035$ can be well-approximated if $a_0 > 0.035$. We also consider retaining all 100 local modes and using the linear approximation to calculate approximate local modes for other values of

a. For each value of a, the local modes with the highest ELBO is used as the final VB parameters. This approach is more expensive since AD is performed 100 times but has the potential to allow the chosen local mode to change with a. The results are shown as the dashed line. These work well when $a_0 = 0.0681$ where the dashed line is able to follow the change point at $a = 0.035$ for all three values of x but poorly for the other values of a_0 with only $a = 0.103$ and $x = 23.0$ showing the change point. This suggests that the approximation can be made more robust by considering many local modes. This idea could be further developed to create a set of local modes for different values of a_0 and the creation of a smaller, diverse set of local modes which could control the number of times that AD is run.

Our application of the method developed in the paper shows that these methods can be used to evaluate the local sensitivity to parts of the model beyond the distribution of the weights. We find that the approximation can work well locally. However, in the presence of multi-modality, it can break down. This could be addressed by building a representative sets of pairs of the values of the hyperparameters and VB parameters. We believe that the authors methods for functional perturbation could also be applied to wider class of parameters in nonparametric models. We congratulate the authors on a thought-provoking paper which describes a promising approach to a difficult problem. At the heart of the paper is an approach to find approximations of the VB parameters and various derivatives for different values of hyperparameters. This could be used to drive optimisation algorithms to find optimal values of hyperparameters using empirical Bayes. Perhaps, more interestingly this would allow hyperparameters to be chosen using a particular loss function (perhaps with cross-validation), which would be consistent with recent thinking around generalized Bayes estimation.

References

Chan, J. C. C., Jacobi, L., and Zhu, D. (2019). "How sensitive are VAR forecasts to prior hyperparameters? An automated sensitivity analysis." *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A, Advances in Econometrics*, 40: 229–248. 325

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37: 212–229. 325

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 326

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, 138: 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 325

Griffin, J. E. (2010). "Default priors for density estimation with mixture models." *Bayesian Analysis*, 5: 45–64. MR2596435. doi: https://doi.org/10.1214/10-BA502. 326

Jacobi, L., Joshi, M. S., and Zhu, D. (2018). "Automated sensitivity analysis for Bayesian inference via Markov chain Monte Carlo: Applications to Gibbs sampling." Technical report. doi: https://doi.org/10.2139/ssrn.2984054. 325

James, L., Lijoi, A., and Prünster, I. (2009). "Posterior analysis of normalized random measures with independent increments." *Scandinavian Journal of Statistics*, 36: 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 325

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *Annals of Probability*, 25: 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 325

Invited Discussion*

María F. Gil–Leyva† and Ramsés H. Mena‡

In Bayesian nonparametric models it is a difficult task to track analytically how prior choices affect posterior inference. As a consequence, hyperparameters, which typically do not enjoy a straightforward interpretation, are often chosen somewhat arbitrarily, blindly or only relying on empirical guidelines. For this reason, we very much appreciate the topic choice and congratulate the authors for the solid mathematical framework developed in their paper.

1 Overview

Consider a Bayesian nonparametric (BNP) model where data points $x = (x_n)_{n=1}^N$, taking values in a Borel space, $\mathbb{X}$, are modeled by

$$\mathcal{P}(x \mid \beta, \pi) = \prod_{n=1}^{N} \mathcal{P}(x_n \mid \beta, \pi) = \prod_{n=1}^{n} \sum_{k=1}^{\infty} \pi_k \, \mathcal{G}(x_n \mid \beta_k), \tag{1}$$

where the frequencies $\pi = (\pi_k)_{k=1}^{\infty}$ are non negative random variables that sum up to one, $\beta = (\beta_k)$ are the random component parameters, and $\mathcal{G}(\cdot \mid \beta_k)$ stands for a distribution over $\mathbb{X}$ for each fixed value of β_k in the parameter space Ω_β. This description of the model is equivalent to that stated in the paper, in which data points are modeled via

$$\mathcal{P}(x \mid z, \beta) = \prod_{n=1}^{N} \mathcal{P}(x_n \mid z_n, \beta) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} \mathcal{G}(x_n \mid \beta_k)^{z_{nk}}, \tag{2}$$

where $z = (z_n)_{n=1}^N$ and $z_n = (z_{n1}, z_{n2}, \ldots)$ is an indicator vector such that $z_{nk} = 1$ occurs with probability π_k and indicates x_n is assigned to the kth component, with $z_{ni} = 0$ for all $i \neq k$. It is further assumed that π and β are independent, and the β_k are iid from a diffuse prior $\mathcal{P}_{\text{base}}$. This means that the underlying discrete probability measure over Ω_β, defined by

$$\mathbf{P} = \sum_{j=1}^{\infty} \pi_k \delta_{\beta_k}, \tag{3}$$

is a *species sampling process* as studied by Pitman (1996, 2006). Once the base measure, $\mathcal{P}_{\text{base}}$, is chosen, it only remains to determine the distribution of the frequencies, π, to fully specify the prior distribution of the BNP model. Being that π takes values in the infinite dimensional simplex $\Delta_\infty = \{(p_1, p_2, \ldots) : p_k \geq 0, \sum_{k=1}^{\infty} p_k = 1\}$, it is

*The authors thankfully acknowledge the support of PAPIIT-UNAM project number IG100221.
†IIMAS-UNAM, México
‡IIMAS-UNAM, México, ramses@sigma.iimas.unam.mx

not straightforward to choose a prior for π directly. The stick-breaking construction (Sethuraman, 1994; Ishwaran and James, 2001) suggest to decompose

$$\pi_1 = \nu_1, \quad \pi_k = \nu_k \prod_{j=1}^{k-1}(1-\nu_j), \quad k \geq 2, \tag{4}$$

and instead specify the law of the *sticks* $(\nu_k)_{k=1}^{\infty}$. At the outset, one could consider $\nu_k \overset{iid}{\sim} \mathcal{P}_{\text{stick}}$, in particular the choice $\mathcal{P}_{\text{stick}} = \mathsf{Beta}(1,\alpha)$ yields $\mathbf{P}$ is a Dirichlet process (Ferguson, 1973; Sethuraman, 1994) with concentration parameter $\alpha > 0$. The tractability of this process and its distinct representations (Blackwell and MacQueen, 1973; Regazzini et al., 2003; Hjort et al., 2010; Ghosal and van der Vaart, 2017) have positioned it as the canonical example of species sampling processes in BNP literature.

One way to perform posterior inference is to design a Markov Chain Monte Carlo (MCMC) algorithm (cf. Escobar and West, 1995; Neal, 2000; Walker, 2007; Papaspiliopoulos and Roberts, 2008; Kalli et al., 2011) where the invariant measure of the constructed Markov chain is the posterior $\mathcal{P}(z,\beta,\pi \mid x)$. Then, after suitably initializing the Markov chain and allowing it to evolve long enough, one obtains samples from $\mathcal{P}(z,\beta,\pi \mid x)$ which can be used to estimate quantities of interest. The most important advantage of MCMC methods is that, at least theoretically, posterior inference can be performed with arbitrary precision if a sufficiently large number of iterations is considered. However, these algorithms tend to be computationally expensive, specially for most complex models and large datasets. Another approach is Variational Bayes (VB) or Variational Inference (VI) (Wainwright and Jordan, 2008; Bishop, 2006; Blei and Jordan, 2006), it consist in approximating the posterior, $\mathcal{P}(z,\beta,\pi \mid x)$, through a distribution, $\mathcal{Q}(\cdot \mid \hat{\eta})$, that minimizes the Kullback-Leiber divergence, among a class of proposals $\{\mathcal{Q}(\cdot \mid \eta) : \eta \in \Omega_\eta\}$ enjoying a simplified tractable form. Quantities of interest can then be estimated by a function, say g, of the optimized variational parameters $\hat{\eta}$: $g(\hat{\eta})$. VB methods are typically much faster to implement than MCMC algorithms. Hence they are well suited for large datasets or when one wants to explore and compare many prior assumptions. The drawback is that there is no guarantee that the posterior distribution will be approximated with arbitrary precision. In general less restricted approximating classes, $\{\mathcal{Q}(\cdot \mid \eta) : \eta \in \Omega_\eta\}$, will lead to a more precise approximations, nonetheless the optimization problem can also be harder to solve for them (if not intractable).

The paper focuses in a mean-field VB variant, where approximating distributions factorize over the parameters. In particular, for the models in (2), it is assumed they take the form:

$$Q(\zeta \mid \eta) = \left(\prod_{k=1}^{K_{\max}} \mathcal{Q}(\beta_k \mid \eta)\right)\left(\prod_{k=1}^{K_{\max}-1} \mathcal{Q}(\nu_k \mid \eta)\right)\left(\prod_{n=1}^{N} \mathcal{Q}(z_n \mid \eta)\right), \tag{5}$$

where ζ collects the first $K_{\max}$ elements of β, π and of z_n, and $\nu_{K_{\max}} = 1$ for all distributions in the variational class. In general, $\hat{\eta}$ will depend on the initial assumptions of the model, and thus the estimation of quantities of interest, $g(\hat{\eta})$, may change if

distinct priors are chosen. The main contribution of the authors is to establish a solid methodology for quantifying the change in $g(\hat{\eta})$ under "mild" changes of

(a) the concentration parameter α in a Dirichlet model, and

(b) the stick-breaking density, $\mathcal{P}_{\text{stick}}$, in models where $\nu_k \overset{iid}{\sim} \mathcal{P}_{\text{stick}}$.

To tackle (a) the mapping $\alpha \mapsto g(\hat{\eta}) = g(\hat{\eta}(\alpha))$ is considered, and under smoothness and differentiability conditions it is locally approximated by

$$g(\hat{\eta}(\alpha)) \approx g(\hat{\eta}^{\text{lin}}(\alpha)), \quad \text{with} \quad \hat{\eta}^{\text{lin}}(\alpha) = \hat{\eta}(\alpha_0) + \left.\frac{d\hat{\eta}(\alpha)}{d\alpha}\right|_{\alpha_0} (\alpha - \alpha_0), \tag{6}$$

in a neighborhood of α_0. Although the treatment of (b) is more delicate and challenging, the idea is the same: evaluate the mapping $\mathcal{P}_{\text{stick}} \mapsto g(\hat{\eta}) = g(\hat{\eta}(\mathcal{P}_{\text{stick}}))$ in a neighborhood of a density $\mathcal{P}_0$, using first-order Taylor series approximations. To this aim the authors consider the perturbations, $\mathcal{P}(\cdot \mid t)$, of $\mathcal{P}_0$, obtained by normalizing $\tilde{\mathcal{P}}(\cdot \mid t)$ where

$$\log \tilde{\mathcal{P}}(\cdot \mid t) = \log \mathcal{P}_0 + t\phi, \tag{7}$$

and ϕ is a Lebesgue-measurable function on $[0, 1]$. The function ϕ can be interpreted as the "direction" towards which $\mathcal{P}_0$ is being perturbed. In particular if $\mathcal{P}_1$ is another density (absolutely continuous with respect to $\mathcal{P}_0$) the choice $\phi = \log(\mathcal{P}_1/\mathcal{P}_0)$ yields

$$\log \tilde{\mathcal{P}}(\cdot \mid t) = (1 - t) \log \mathcal{P}_0 + t \log \mathcal{P}_1, \tag{8}$$

allowing to continuously transform $\mathcal{P}_0$ into $\mathcal{P}_1$. In general, if $\hat{\eta}(t)$ represents the value of $\hat{\eta}$ for a particular choice $\mathcal{P}(\cdot \mid t)$ (once ϕ is chosen), under smoothness and differentiability conditions, an approximation analogous to (6) is obtained through

$$g(\hat{\eta}(t)) \approx g(\hat{\eta}^{\text{lin}}(t)), \quad \text{with} \quad \hat{\eta}^{\text{lin}}(t) = \hat{\eta}(0) + \left.\frac{d\hat{\eta}(t)}{dt}\right|_{t=0} (t). \tag{9}$$

Since the mapping, $t \mapsto g(\hat{\eta}(t))$, depends on the choice of ϕ, this function can also be understood as the "direction" over which $\mathcal{P}_{\text{stick}} \mapsto g(\hat{\eta}(\mathcal{P}_{\text{stick}}))$ is being studied in a neighborhood of $\mathcal{P}_0$. Naturally, some choices of ϕ yield larger (positive or negative) gradients of $t \mapsto g(\hat{\eta}(t))$. In this sense, the authors also provide guidelines to understand the effect of ϕ and discriminate "directions" over which the mapping $\mathcal{P}_{\text{stick}} \mapsto g(\hat{\eta}) = g(\hat{\eta}(\mathcal{P}_{\text{stick}}))$ is more steep. Performing the analysis for such choices of ϕ allows an overall assessment of the model's local robustness at $\mathcal{P}_0$, for a particular quantity of interest.

In the experimental analysis the methodology is applied to evaluate sensitivity and model robustness for clustering estimation. Although, the experimental results do not include evaluation for density estimation, BNP models are widely used for this task. To the best of our understanding, the methodology developed by the authors readily applies to this context, where different BNP models have been found less sensitive to distinct choices of sticks distributions (cf. see Figure 1).

2 Extensions

The remainder of this discussion attempts to motivate extensions of the solid mathematical framework developed in the paper for assessing local sensitivity in stick-breaking BNP models. The methodology developed by the authors considers BNP stick-breaking models with iid sticks, $\nu_k \overset{iid}{\sim} \mathcal{P}_{\text{stick}}$. Two natural directions to extend them are (a) to relax the independence hypothesis or (b) to relax the identical distribution assumption. In particular, a model that arises under (a) is the Geometric model (Fuentes-García et al., 2010) where $\nu_k = v$ for every $k \geq 1$, and some random variable $0 < v < 1$ almost surely. A common feature between these sticks and iid sticks is that in both cases $(\nu_k)_{k=1}^{\infty}$ forms an exchangeable sequence. The models resulting from this general assumption have been recently studied by Gil-Leyva and Mena (2021), under the name exchangeable stick-breaking (ESB) models. In particular, they provide a way to modulate dependence among elements in $(\nu_k)_{k=1}^{\infty}$ by tuning a single $[0,1]$-valued hyperparameter, ρ, under the assumption that the directing random measure, of the exchangeable sequence $(\nu_k)_{k=1}^{\infty}$, is a Dirichlet process (Ferguson, 1973) over $[0,1]$. Formally, in their Theorem 3.3 and Corollary 4.2 they derived the following result:

Theorem 1. *For each $\rho \in (0,1)$ let $\mathbf{P}_\rho$ be a species sampling process, as in (3), with stick-breaking weights, as in (4), where $\nu_k \mid \mathbf{P}_{\text{stick}} \overset{iid}{\sim} \mathbf{P}_{\text{stick}}$, and $\mathbf{P}_{\text{stick}}$ is a Dirichlet process with concentration parameter $\alpha = (1-\rho)/\rho$. Then as $\rho \to 0$, $\mathbf{P}_\rho$ converges in distribution to a stick-breaking process with iid sticks, $\mathbf{P}_0$; and as $\rho \to 1$, $\mathbf{P}_\rho$ converges in distribution to a Geometric process, $\mathbf{P}_1$.*

Whenever the mapping $\beta \mapsto \mathcal{G}(\cdot \mid \beta)$ is continuous with respect to the weak topology, which is the case of Gaussian kernels, Theorem 1 provides a way to continuously warp BNP models with iid sticks into Geometric models. Or in other words, continuously transform completely independent sticks into completely dependent ones while keeping the identical distribution assumption. Another common feature between the identical sticks of Geometric processes and iid sticks is that both form Markov processes. In particular, if one considers Markov sticks, $(\nu_k)_{k=1}^{\infty}$, with transition probability kernel

$$\mathbb{P}[\nu_k \in \cdot \mid \nu_{k-1}] = \rho\, \delta_{\nu_{k-1}} + (1-\rho)\mathcal{P}_{\text{stick}}, \tag{10}$$

and initial distribution $\mathcal{P}_{\text{stick}}$, then $\nu_k \sim \mathcal{P}_{\text{stick}}$, marginally, for every $k \geq 1$. Furthermore, the choice $\rho = 0$ yields $\nu_k \overset{iid}{\sim} \mathcal{P}_{\text{stick}}$ and for $\rho = 1$ we recover the sticks $\nu_k = \nu_1 \sim \mathcal{P}_{\text{stick}}$, of Geometric processes. Hence (10), also establishes a way to continuously transform iid sticks into identical sticks, but in this case by means of a Markovian symmetry instead of exchangeability. This suggests that a first step towards evaluating sensitivity of stick-breaking BNP models with respect to the dependence amongst sticks could be to study the mapping $\rho \mapsto g(\hat{\eta}(\rho))$, by extending the present work to the models in Theorem 1 or to models with sticks described by (10). It would be particularly appealing to evaluate $\rho \mapsto g(\hat{\eta}(\rho))$ at a neighborhood of $\rho = 0$, as this could give insight on how much does the estimated quantity of interest, $g(\hat{\eta})$, changes if the independence of the sticks is perturbed towards exchangeability or a Markov dependence.

Another interesting direction to extend the work of the authors is to keep the independence assumption on the sticks and relax the identical distribution one. In this

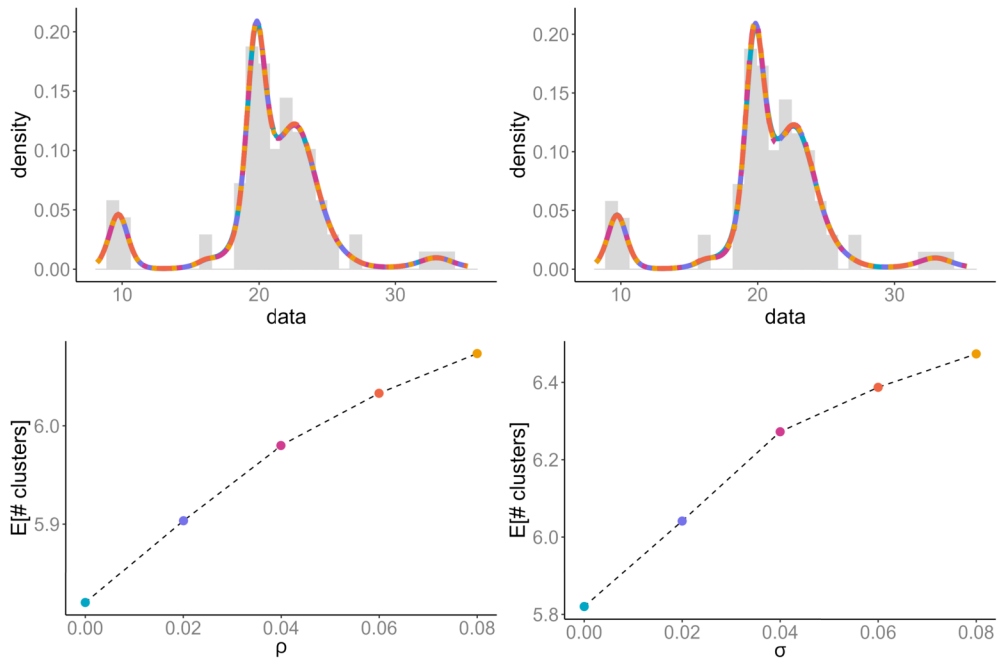

Figure 1: Estimated density of the galaxy dataset (upper row) and sensitivity analysis of the expected number of clusters (bottom row), for ESB (left column) and PY (right column).

scenario we find the well-known Pitman-Yor (PY, Pitman and Yor, 1997; Ishwaran and James, 2001) model, whose frequencies admit a stick-breaking decomposition with independent $\nu_k \overset{ind}{\sim} \mathsf{Beta}(1-\sigma, \alpha + k\sigma)$, for some $\sigma \in [0,1)$ and $\alpha > -\sigma$. Evidently, the choice $\sigma = 0$ recovers a Dirichlet process for which the expected frequencies, $\mathbb{E}[\pi_k]$, decay exponentially fast. In contrast, for $\sigma > 0$, the expected weights decrease much slower. Hence, it also seems very appealing to extend the proposed methodology so to (at least) cover Pitman-Yor processes. Then, the analysis of the mapping $(\sigma, \alpha) \mapsto g(\hat{\eta}(\sigma, \alpha))$ on a neighborhood of $(0, \alpha)$ could be starting point of a more rigorous understanding about the implications of biasing the identical distribution assumption on the sticks towards working with ν_k's that decrease in expectation, and thus lead to heavier-tailed frequencies π.

Inhere we have performed a small simulation study where five ESB models, as in Theorem 1, and five PY models were implemented with distinct values of ρ and σ, respectively, in order to estimate the density of the well-known `galaxy` dataset as well as the posterior expected number or clusters

$$\mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}\left\{\sum_{n=1}^{N} z_{n,k} > 0\right\} \middle| x\right]. \tag{11}$$

In the first row of Figure 1 we observe that the estimated density is not very sensitive to different prior assumptions on $(\nu_k)_{k=1}^{\infty}$, in fact it is very similar for all considered ESB and PY models. In the bottom row we see that the expected number of clusters does changes even under mild increments of ρ or σ, for both ESB and PY models. We also see that if these quantities are closer to zero, increments affect more the expected number of clusters. Comparing the two graphs in the bottom row we observe that PY models are more sensitive towards increments of σ than ESB models are when increasing ρ. Recalling the role that ρ and σ play, this small analysis suggests that stick-breaking models are less sensitive when biasing the iid assumption on the sticks towards exchangeable sticks than when biasing the same assumption towards the non identically distributed PY sticks.

It is worth highlighting that for this study we used a Gibbs sampling method, thus the algorithm had to be run separately for each different model. Additionally, due to the random nature of the algorithm we observed relatively large variance for the estimates of (11) when running the code multiple times with few iterations, therefore many iterations had to be considered in order to provide good estimates of (11). This is computationally expensive, even for such a simple dataset and few hyper-parameter values. The framework developed by the authors to analyze local sensitivity for iid stick-breaking models can be implemented significantly faster, and it would be very appealing if their methodology could be extend to account for experiments such as the one presented inhere. We believe that the gain in terms of computational time would be specially significant for ESB models, where Gibbs samplers convergence to the stationary distribution is slow.

2.1 Variational Inference for Dirichlet driven ESB models

In order to extend the present methodology to ESB models, the first step would be to derive a VB approach to implement them. Here we discuss possible paths that could be pursued to derive a mean-field VB implementation of ESB models. To this aim, we first recall how does mean-field VB operates, so say we are seeking to approximate the posterior distribution $\mathcal{P}(\zeta \mid x)$. The mean-field VB approach suggests to consider variational distributions that factorize as

$$\mathcal{Q}(\zeta) = \prod_{i=1}^{m} \mathcal{Q}(\zeta_i), \tag{12}$$

for some partition $\{\zeta_i\}_{i=1}^{m}$ of ζ. To solve the optimization problem we can guess initial values of $\mathcal{Q}^*(\zeta_i)$ and iteratively update the factors through

$$\log \mathcal{Q}^*(\zeta_j) = \int \log \mathcal{P}(x, \zeta) \prod_{i \neq j} \mathcal{Q}^*(d\zeta_i), \quad j \in \{1, \ldots, m\} \tag{13}$$

(cf. Equation (10.9) in Bishop, 2006).

For the BNP models in (2) with stick-breaking weights as in (4), at the outset, one can assume that the approximating distributions take the form

$$\mathcal{Q}(\zeta) = \mathcal{Q}(z_{[K]})\mathcal{Q}(\beta_{[K]}, \nu_{[K]}), \tag{14}$$

where $\zeta = \{z_{[K]}, \beta_{[K]}, \nu_{[K]}\}$, $z_{[K]}$ collects the first K elements of each z_n, $\beta_{[K]} = (\beta_k)_{k=1}^K$ and $\nu_{[K]} = (\nu_k)_{k=1}^K$, with $\nu_K = 1$ for all distributions in the variational class, and some truncation level $K = K_{\max}$. Noting that

$$\mathcal{P}(x, z, \beta, \nu) = \mathcal{P}(x \mid z, \beta)\mathcal{P}(z \mid \nu)\mathcal{P}(\beta)\mathcal{P}(\nu),$$

from (13) we find

$$\log \mathcal{Q}^* \left(z_{[K]}\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \lambda_{nk} + \text{const.}, \tag{15}$$

where $\log \lambda_{nk} = \mathbb{E}^*[\log \mathcal{G}(x_n \mid \beta_k)] + \mathbb{E}^*[\log \nu_k] + \sum_{i=1}^{k-1} \mathbb{E}^*[\log(1 - \nu_i)]$, and

$$\log \mathcal{Q}^* \left(\beta_{[K]}, \nu_{[K]}\right) = \log \mathcal{Q}^* \left(\beta_{[K]}\right) + \log \mathcal{Q}^* \left(\nu_{[K]}\right), \tag{16}$$

where

$$\begin{aligned} \log \mathcal{Q}^* \left(\beta_{[K]}\right) &= \sum_{k=1}^{K} \left\{ \log \mathcal{P}_{\text{base}}(\beta_k) + \sum_{n=1}^{N} \mathbb{E}^*[z_{nk}] \log \mathcal{G}(x_n \mid \beta_k) \right\} + \text{const.}, \\ \log \mathcal{Q}^* \left(\nu_{[K]}\right) &= \log \mathcal{P}(\nu_{[K]}) + \sum_{k=1}^{K-1} a_k \log \nu_k + b_k \log(1 - \nu_k) + \text{const.} \end{aligned} \tag{17}$$

Here $\mathbb{E}^*[f(\cdot)] = \mathbb{E}_{\mathcal{Q}^*(\cdot)}[f(\cdot)]$, $a_k = \sum_{n=1}^{N} \mathbb{E}^*[z_{nk}]$, $b_k = \sum_{n=1}^{N} \sum_{i=k+1}^{K} \mathbb{E}^*[z_{nk}]$, and $\mathcal{P}(\nu_{[K]})$ refers to the joint prior density of $\nu_{[K]}$. In particular, if $\nu_k \overset{iid}{\sim} \mathcal{P}_{\text{stick}}$ apriori, we get $\log \mathcal{P}(\nu_{[K]}) = \sum_{k=1}^{K} \log \mathcal{P}_{\text{stick}}(\nu_k)$. In this case, (15)–(17) imply that considering the variational class determined by (14) is equivalent to assume that the approximating distributions factorize as

$$\mathcal{Q}(\zeta) = \left(\prod_{n=1}^{N} \mathcal{Q}(z_n) \right) \left(\prod_{k=1}^{K} \mathcal{Q}(\beta_k) \right) \left(\prod_{k=1}^{K-1} \mathcal{Q}(\nu_k) \right), \tag{18}$$

identically as in (5). If instead, the ν_k are exchangeable and $\nu_k \mid \mathbf{P}_{\text{stick}} \overset{iid}{\sim} \mathbf{P}_{\text{stick}}$, where $\mathbf{P}_{\text{stick}}$ is a Dirichlet process with total mass parameter α and base measure $\mathcal{P}_{\text{stick}}$ we get

$$\mathcal{P}(\nu_{[K]}) = \frac{\alpha^M \prod_{m=1}^{M} (n_m - 1)!}{\prod_{i=0}^{K-1} (\alpha + i)} \times \prod_{m=1}^{M} \mathcal{P}_{\text{stick}}(\nu_m^*), \tag{19}$$

where $\nu_1^*, \ldots, \nu_M^*$ are the distinct values that $\nu_{[K]}$ exhibits and $n_m = |\{k \leq K : \nu_k = \nu_m^*\}|$ (Pitman, 1996). For this type of sticks we find that assuming the class in (14) is equivalent to consider approximating distributions of the form

$$\mathcal{Q}(\zeta) = \left(\prod_{n=1}^{N} \mathcal{Q}(z_n) \right) \left(\prod_{k=1}^{K} \mathcal{Q}(\beta_k) \right) \mathcal{Q}\left(\nu_{[K]}\right). \tag{20}$$

However, since $\log \mathcal{P}(\nu_{[K]})$ does not decomposes into terms that only involve one ν_k, the factorization in (18) is not equivalent to that in (14). One option would be to work with approximating distributions as in (20), thus avoiding additional restrictions to the variational class. In this setting, to fully derive the mean-field VB implementation it would be necessary to compute explicitly $\mathcal{Q}^*(\nu_{[K]})$ as in (17) where $\mathcal{P}(\nu_{[K]})$ is given by (19). Although theoretically possible, in practice this is numerically very expensive as the computation of the normalization constant of $\mathcal{Q}^*(\nu_{[K]})$ requires to sum over all partitions of the first $K-1$ integers. Alternatively, one could assume approximating distributions as in (18), in which case it suffices to compute

$$\log \mathcal{Q}^*(\nu_k) = \log \mathcal{P}(\nu_k \mid \nu_{-k}) + \sum_{k=1}^{K-1} a_k \log \nu_k + b_k \log(1-\nu_k) + \text{const.},$$

for each $k \leq K-1$, where a_k and b_k are as in (17) and $\mathcal{P}(\nu_k \mid \nu_{-k})$ refers to the conditional distribution ν_k given $\nu_{-k} = (\nu_1, \ldots, \nu_{k-1}, \nu_{k+1}, \ldots, \nu_{K-1})$. Being that apriori elements in $\nu_{[K]}$ are exchangeable and driven by a Dirichlet process it is well-known that

$$\mathcal{P}(\nu_k \mid \nu_{-k}) = \sum_{m=1}^{M} \frac{n_m}{\alpha + K - 1} \mathbb{1}\{\nu_k = \nu_m^*\} + \frac{\alpha}{\alpha + K - 1} \mathcal{P}_{\text{stick}}(\nu_k),$$

where $\nu_1^*, \ldots, \nu_M^*$ are the distinct values in ν_{-k} and $n_m = |\{j \neq k : \nu_j = \nu_m^*\}|$ (Blackwell and MacQueen, 1973; Pitman, 1996). In particular, if $\mathcal{P}_{\text{stick}}$ stands for a Beta distribution, computing $\mathcal{Q}^*(\nu_k)$ explicitly can be easily achieved. Unfortunately, assuming the factorization in (18) for ESB models does represents making an additional assumption that shortens the variational class. Hence, it is not obvious whether it is preferable assume (18), or to assume (20) and approximate the normalization constant of $\mathcal{Q}^*(\nu_{[K]})$ instead of computing it explicitly. A third option that could be worth investigating is to take into consideration the directing random measure, $\mathbf{P}_{\text{stick}}$, and include it in ζ. Perhaps by doing so one could find a middle point between considering the class in (18) or that in (20). Provided that we can find a variational class of distributions that is sufficiently flexible and tractable, and that generalizes the mean-field VB approach for iid sticks considered by the authors, it then makes sense to try and extend their methodology to ESB models.

References

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer. URL https://books.google.com.mx/books?id=qWPwnQEACAAJ. MR2247587. 331, 335

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Polya urn schemes." *Annals of Statistics*, 1(2): 353–355. MR0362614. 331, 337

Blei, D. M. and Jordan, M. I. (2006). "Variational inference for Dirichlet process mixtures." *Bayesian Analysis*, 1(1): 121–143. MR2227367. doi: https://doi.org/10.1214/06-BA104. 331

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 331

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1(2): 209–230. MR0350949. 331, 333

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2010). "A New Bayesian Nonparametric Mixture Model." *Communications in Statistics – Simulation and Computation*, 39(4): 669–682. MR2785657. doi: https://doi.org/10.1080/03610910903580963. 333

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. MR3587782. doi: https://doi.org/10.1017/9781139029834. 331

Gil-Leyva, M. and Mena, R. H. (2021). "Stick-breaking processes with exchangeable length variables." *Journal of the American Statistical Association*. doi: https://doi.org/10.1080/01621459.2021.1941054. 333

Hjort, N., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. MR2722988. doi: https://doi.org/10.1017/CBO9780511802478.002. 331

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 331, 334

Kalli, M., Griffin, J. E., and Walker, S. (2011). "Slice sampling mixtures models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 331

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 331

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95: 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 331

Pitman, J. (1996). "Some developments of the Blackwell-MacQueen urn scheme." In et al., T. F. (ed.), *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30 of *Lecture Notes-Monograph Series*, 245–267. Hayward, California: Institute of Mathematical Statistics. MR1481784. doi: https://doi.org/10.1214/lnms/1215453576. 330, 336, 337

Pitman, J. (2006). *Combinatorial Stochastic Processes*, volume 1875 of *École d'été de probabilités de Saint-Flour*. New York: Springer-Verlag Berlin Heidelberg, first edition. 330

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution de-

rived from a stable subordinator." *Annals of Probability*, 25(2): 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 334

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *Annals of Statistics*, 31(2): 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 331

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. MR1309433. 331

Wainwright, M. J. and Jordan, M. I. (2008). "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning*, 1(1–2): 1–305. 331

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics-Simulation and Computation*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 331

Invited Discussion*

Filippo Ascolani†, Marta Catalano‡ and Igor Prünster§

The specification of the prior and its impact on statistical analyses have received considerable attention since the early developments of Bayesian methodologies. In a seminal work (de Finetti, 1934) Bruno de Finetti writes (in our own translation): "... *subjective* does not mean *arbitrary*, we can not make events more or less likely as we please, but rather according to the degree of confidence we feel about them". In fact, a prior should be a truthful representation of our beliefs about the phenomenon of interest and needs to satisfy the *coherence* principle. See de Finetti (1937, 1970); Regazzini (1987). The following natural step consists in assessing how prior choices affect conclusions, leading immediately into the realm of robustness: does a small change in the prior specification lead to a big change in the inferential conclusions?

This is particularly relevant when dealing with Bayesian nonparametric (BNP) inference, which loosely speaking entails specifying a distribution over infinitely many parameters. This makes the study of the distributional properties of a nonparametric model and, a fortiori, its subjective elicitation challenging tasks. From this viewpoint, it is reasonable to focus on meaningful finite-dimensional functionals of the nonparametric prior and tune the parameters based on their behaviour. For instance, a popular recipe is to elicit the parameters of the nonparametric prior based on the induced behaviour of the number of clusters (Lijoi et al., 2007a,b). An alternative strategy consists in specifying the distribution of the mean (Kessler et al., 2015; Gaffi et al., 2022). In this spirit, the main contribution of this stimulating discussion paper is to develop a framework to conduct a sensitivity analysis for specific functionals, with theoretical guarantees and still retaining computational convenience: this works by combining a variational Bayes approach (Blei and Jordan, 2006) with automatic differentiation and the tools of local robustness (Gustafson, 1996). The result is an interesting and inspiring work, that will surely have an impact on both future research and practice.

The authors consider mixture models directed by a discrete nonparametric prior, with a special focus on Dirichlet process mixtures. Thus the exchangeable observations are modeled through a nonparametric random mixture of suitable kernels. Note that, when performing frequentist asymptotic validations of Bayesian models, one assumes the data not to be generated according to the Bayesian model, but rather to be i.i.d. from a "true" distribution. This corresponds to the so–called "what-if" approach of Diaconis and Freedman (1986). In this perspective, when analyzing the asymptotic distribution

*M. Catalano is partially supported by the Heilbronn Institute for Mathematical Research. I. Prünster acknowledges support from MIUR, PRIN Project 2015SNS29B.

†Department of Decision Sciences and BIDSA, Bocconi University, via Röentgen 1, 20136 Milan, Italy, filippo.ascolani@phd.unibocconi.it

‡Department of Statistics, University of Warwick, CV47AL Coventry, UK, marta.catalano@warwick.ac.uk

§Department of Decision Sciences and BIDSA, Bocconi University, via Röentgen 1, 20136 Milan, Italy, igor.pruenster@unibocconi.it

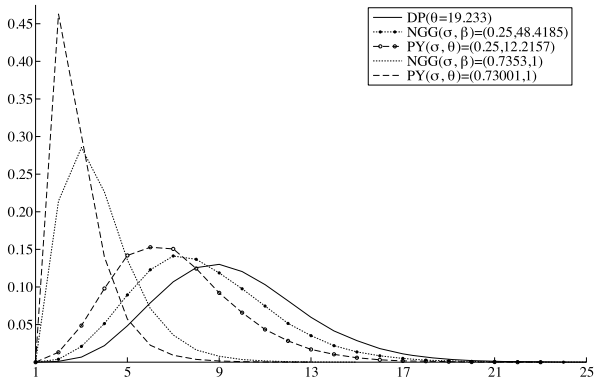

Figure 1: Posterior distributions on the number of components corresponding to mixtures of the Dirichlet (DP), the Pitman-Yor (PY) and the normalized generalized gamma (NGG) processes with $n = 50$ and parameters set so that $E[K_{50}] = 25$. Source: De Blasi et al. (2015).

of a functional of interest, there can be distinct behaviours depending on whether the data generating mechanism produces exchangeable observations from the model or i.i.d. data from a "true" distribution. Since it may well be the case that both assumptions are misspecified, particular care is needed before reaching general conclusions. In the context of mixture models, the (latent) mixing component is often of crucial interest, since it allows to perform model-based clustering and infer a probability distribution for the number of clusters in the population. It is therefore important to conduct a sensitivity analysis on these latent quantities, although in contrast with de Finetti's prescription to focus exclusively on observable quantities.

In fact, the latent structure within mixture models is a quite delicate object and different nonparametric priors may lead to very different posterior distributions on the number of clusters, while at the same time to almost identical density estimates (which are obtained by "integrating out the latent structure"). To illustrate this phenomenon we recall a toy example of De Blasi et al. (2015, Section 3): for 50 i.i.d. data from a mixture of two Gaussians (which plays the role of the "true" distribution), the performance of five different "misspecified" nonparametric priors is compared. More precisely, the five priors on the mixing measure (belonging to Dirichlet, Pitman-Yor or Normalized generalized gamma families) are all tuned such that the prior expectation on the number of mixture components K_{50} is 25, strongly misspecified with respect to the "true" number 2. The posterior distributions clearly move mass towards a smaller number of components as shown in Figure 1, but a significant variability persists with the posterior mode going from to 2 to 9 components. The Pitman-Yor and normalized generalized gamma processes with a high value of σ exhibit a better correction to the prior misspecification, whereas the Dirichlet process stands out as the less robust. It is important to note that none of the models assigns mass to a single cluster: in general, the lowest value with positive mass can be heuristically interpreted as the minimum number of mixture compo-

nents (of the given kernel, which in this case is Gaussian) needed to fit a given "true" distribution. In contrast, the density estimates are almost indistinguishable (De Blasi et al., 2015, Figure 4). This is not surprising since any density can be approximated arbitrarily well with a mixture density with number of components larger than the ground truth (possibly even infinite). In fact, the described phenomena in terms of clustering structure and density estimation are generic for misspecified priors and not confined to the toy example recalled here. As for the frequentist asymptotic evaluation, results about density estimation support the above findings: under mild assumptions on the "true" univariate density, the posterior densities corresponding to a Dirichlet process or a Pitman-Yor process mixing measure are proven to converge to the "truth" with nearly optimal contraction rates (Ghosal and van der Vaart, 2007b; Scricciolo, 2014), and we envisage this to hold for a wide range of BNP mixing measures and beyond exchangeability (see, e.g., Catalano et al. (2022)). Frequentist evaluation of the posterior distribution of K_n itself is more problematic, since for any finite sample the number of clusters is bounded by the sample size n leading to a distribution for the number of clusters on the integers $\{1, \ldots, n\}$, an appealing feature. Letting the sample size diverge, one automatically allows the number of clusters to diverge as well. Then, an identifiability issue takes over because of the above mentioned fact that a density can be approximated arbitrarily well with more mixture components than needed: what is usually framed in terms of inconsistent behaviour, should rather be interpreted as lack of identifiability. Finding the appropriate framework for a frequentist asymptotic evaluation of K_n within mixtures is, in our view, still an open problem. On the contrary, interesting results have been obtained for the frequentist evaluation of the whole mixing measure in the Wasserstein distance, a research line opened by the seminal paper of Nguyen (2013). The Wasserstein distance is ideal in this setting because it takes into account the geometry of the underlying space, thus identifying clusters that are arbitrarily close, irrespectively of their overall number.

In summary, inference on the latent structure is often a primary goal of the analysis. However, it is quite a delicate object and the inferential results may depend on both the chosen prior and the specification of its parameters. Hence, a careful sensitivity study is of paramount importance. This represents a challenging task and is typically carried out making some simplifying hypotheses. Two of them are also assumed in the paper.

First, the authors approximate the Dirichlet process by truncating its stick-breaking representation to $K_{\max} < \infty$ elements: this allows one to work with finitely many parameters and leads to a computationally convenient algorithm. Taking a sufficiently large $K_{\max}$ leads to "good" approximations of a Dirichlet process a priori. However, what happens a posteriori is less clear-cut. See Regazzini and Sazonov (2001) for early results in this direction. When the Dirichlet process is employed within a mixture model and the focus is on the clustering structure, assessing the quality of approximation in the posterior represents an important open question. Moreover, if one allows the number of observations n to vary, it is natural to wonder about the relationship between $K_{\max}$ and n. In fact, it seems intuitive to let $K_{\max}$ increase with n leading to a sample size dependent approximation. By the previous considerations it is apparent that $K_{\max} \geq n$ is desirable: $K_{\max} < n$ would imply a truncation also of the support of the posterior distribution of the number of clusters adding another layer of approximation. If, for computational convenience, one is willing to accept this additional approximation layer in

order to have $K_{\max}$ grow slower than n, what could a reasonable choice be? A possibility would be to use the *a priori* growth rate of K_n, which is often available (e.g., logarithmic and polynomial for, respectively, Dirichlet and Pitman-Yor processes). Clearly its posterior growth rate would be more appropriate, but no results are known for mixture models and current results are confined to prediction in species sampling problems (Favaro et al., 2009). Our intuition is that the growth may depend on the degree of closeness between the kernel and the data generating mechanism, but since at the present stage a formal treatment is missing, the value of $K_{\max}$ should be decided case by case.

Second, the authors consider a variational approximation, on which the sensitivity analysis is conducted: this is a main theme of the paper, since by carefully studying the variational distribution they are able to derive (Fréchet) differentiability of the involved functions (Theorems 1 and 2). However, it is not clear how close in reverse KL-divergence the distributions should be in order to be close in terms of the clustering structure. Indeed, it is not difficult to find priors that are close in distributional sense, but with very different asymptotic predictive growth, such as a Dirichlet process with parameter θ and a Pitman-Yor process with parameters $(\theta, \sigma \approx 0)$. See Favaro et al. (2009); De Blasi et al. (2015). Thus, taking into account also the dependence of $K_{\max}$ on n, we cannot exclude that the resulting approximation has a different latent behaviour than the original process, especially when n is large.

In conclusion, this inspiring paper combines results from different fields and provides probably the most effective tool for sensitivity analysis that we may hope for, with the current knowledge. Indeed, in our opinion, the biggest issue is our limited understanding of the behaviour of the posterior distribution on the latent space, even when n is large. The same holds for the distributional properties of other latent quantities of interest, such as the co-clustering matrix, which are currently still unknown. We hope to see a stream of work in this direction in the near future. We congratulate the authors for this seminal contribution to the study of sensitivity in BNP.

References

Blei, D. and Jordan, M. I. (2006). "Variational inference for Dirichlet process mixtures." *Bayesian Analysis*, 1: 121–143. MR2227367. doi: https://doi.org/10.1214/06-BA104. 340

Catalano, M., Blasi, P. D., Lijoi, A., and Pruenster, I. (2022). "Posterior asymptotics for boosted hierarchical Dirichlet process mixtures." *Journal of Machine Learning Research*, 23(80): 1–23. 342

De Blasi, P., Favaro, S., Lijoi, A., Mena, H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13: 212–229. doi: https://doi.org/10.1109/TPAMI.2013.217. 341, 342, 343

de Finetti, B. (1934). *L'invenzione della verità*. Raffaello Cortina. 340

de Finetti, B. (1937). "La prévision: ses lois logiques, ses sources subjectives." *Annales de l'institut Henri Poincaré*, 7(1): 1–68. MR1508036. 340

de Finetti, B. (1970). *Theory of Probability: A Critical Introductory Treatment*. New York: John Wiley. MR0440640. 340

Diaconis, P. and Freedman, D. (1986). "On the consistency of Bayes estimates." *The Annals of Statistics*, 14(1): 1–26. MR0829555. doi: https://doi.org/10.1214/aos/1176349830. 340

Favaro, S., Lijoi, A., Mena, H., and Prünster, I. (2009). "Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior." *Journal of the Royal Statistical Society Series B*, 71: 993–1008. MR2750254. doi: https://doi.org/10.1111/j.1467-9868.2009.00717.x. 343

Gaffi, F., Lijoi, A., and Prünster, I. (2022). "Random probability measures with fixed mean distributions." Technical report. 340

Ghosal, S. and van der Vaart, A. (2007b). "Posterior convergence rates of Dirichlet mixtures at smooth densities." *The Annals of Statistics*, 35(2): 697–723. MR2336864. doi: https://doi.org/10.1214/009053606000001271. 342

Gustafson, P. (1996). "Local sensitivity of posterior expectations." *Annals of Statistics*, 24: 174–195. MR1389886. doi: https://doi.org/10.1214/aos/1033066205. 340

Kessler, D., Hoff, P., and Dunson, D. (2015). "Marginally specified priors for nonparametric Bayesian estimation." *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 77: 35–58. MR3299398. doi: https://doi.org/10.1111/rssb.12059. 340

Lijoi, A., Mena, R. H., and Prünster, I. (2007a). "Bayesian nonparametric estimation of the probability of discovering new species." *Biometrika*, 94(4): 769–786. MR2416792. doi: https://doi.org/10.1093/biomet/asm061. 340

Lijoi, A., Mena, R. H., and Prünster, I. (2007b). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 715–740. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 340

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models." *Annals of Statistics*, 41: 370–400. MR3059422. doi: https://doi.org/10.1214/12-AOS1065. 342

Regazzini, E. (1987). "De Finetti's coherence and statistical inference." *The Annals of Statistics*, 15(2): 845–864. MR0888444. doi: https://doi.org/10.1214/aos/1176350379. 340

Regazzini, E. and Sazonov, V. V. (2001). "Approximation of laws of random probabilities by mixtures of Dirichlet distributions with applications to nonparametric Bayesian inference." *Theory of Probability and Its Applications*, 45: 93–110. MR1810976. doi: https://doi.org/10.1137/S0040585X97978063. 342

Scricciolo, C. (2014). "Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures." *Bayesian Analysis*, 9(2): 475–520. MR3217004. doi: https://doi.org/10.1214/14-BA863. 342

Contributed Discussion

Giovanni Rebaudo*, Augusto Fasano†, Beatrice Franzolini‡, and Peter Müller§

We congratulate the authors for a very interesting paper, which provides a concrete contribution to the Bayesian nonparametric (BNP) literature. Their work provides an efficient method to evaluate the sensitivity of posterior quantities of interest – computed through variational Bayes approximations – to the prior distribution of the mixing weights in Bayesian discrete mixture models. The authors argue for sensitivity checks to uncover possible non-robustness of the results to prior settings. Importantly, they develop an easy-to-use and efficient method to do it in BNP mixtures. In the following, we illustrate our comments on the most widely used construction discussed by the authors: the Dirichlet process mixture (DPM) model (Lo, 1984), namely

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} \mathcal{P}(\cdot|\theta_i), \quad \theta_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \sim \text{DP}(\alpha, \mathcal{P}_{\text{base}}). \tag{1}$$

Following the authors, in such a setting one of the goals is to perform inference about the random number of clusters, G_{cl}, that is the number of occupied mixture components $\mathcal{P}$ in a sample of size N, with a particular focus on its expected value. In this regard, it is worth noticing that robustness is relevant just in terms of the specific quantity of interest or decision of the analysis.

First, we agree with the authors that sensitivity analysis to prior assumptions should be done routinely if the prior specification is driven by mathematical convenience or heuristics, as often is in BNP models. This would highlight the influence of the specific assumptions on the results of the analysis, pointing to which of them should be justified more strongly. After assessing sensitivity, the next fundamental question is: *how can we justify probabilistic assumptions in the challenging infinite/high-dimensional Bayesian settings?* In principle, one possibility could be provided by the subjective Bayesian paradigm. According to it, the prior should reflect *a priori* opinions on quantities of interest. Those are more easily elicited when expressed directly in terms of observable values (see e.g., Fortini and Petrone, 2016). However, this approach can be particularly challenging in the BNP world due to the infinite/high-dimensionality of the parameter space. One way to tackle this issue and elicit prior assumptions consists in working with prior predictive distributions or with the *a priori* expected value of the number of clusters (see e.g., De Blasi et al., 2015). However, posterior inference strongly depends also on other *a priori* assumptions, such as the choice of the mixture kernel and the base measure in DPM. Another way to justify the choice of the prior is in terms of the properties of the summaries of interest (e.g., consistency of G_{cl}) assuming an ideal frequentist truth (Nobile, 1994; Miller and Harrison, 2018; Ascolani et al., 2022). Furthermore, different prior settings can be also specified in a given dataset by tuning

*University of Texas at Austin, TX, USA, giovanni.rebaudo@austin.utexas.edu
†Collegio Carlo Alberto, Torino, Italy, augusto.fasano@carloalberto.org
‡Agency for Science, Technology and Research, Singapore, franzolini@pm.me
§University of Texas at Austin, TX, USA, pmueller@math.utexas.edu

the hyperparameters in terms of predictive accuracy, e.g. via cross-validation. All of the above methods – as well as other possible alternatives, including popular empirical-based approaches (Liu, 1996; McAuliffe et al., 2006) – require to specify some subjective assumptions (e.g., homogeneity assumptions, prior distribution, data generating process, loss function). The implications of such assumptions are challenging to assess, pointing toward the need for further research, especially in the BNP mixture framework.

Second, considering the sensitivity of the stick-breaking prior to values of α in (1), it would be interesting to assess how the specification of a prior for the concentration parameter could increase robustness. The use of a prior on α leads to a mixing measure that is itself a mixture in the sense of Antoniak (1974). Ideally, this would allow learning from the data which values of α are most appropriate for the data at hand. Consequently, it would be very interesting to investigate how the results and the sensitivity checks proposed by the authors could be embedded in such a framework.

Another useful extension of the idea and techniques developed by the authors is to provide a toolkit that assesses sensitivity to the choice of the kernel or of the base measure of the DP. A common choice of kernel $\mathcal{P}$ and base measure $\mathcal{P}_{\text{base}}$ in (1) are the Gaussian kernel and the conjugate normal-inverse-Wishart base measure, respectively. Such assumptions are typically motivated by mathematical convenience and the choice of the hyperparameters of the base measure is mainly carried out following heuristics. However, posterior inference on the number of clusters strongly depends on such assumptions as shown empirically and theoretically (see e.g., Petralia et al., 2012; Cai et al., 2021; Chandra et al., 2020).

Finally, as anticipated by the authors, it would be very interesting to exploit the general results developed in the work to obtain an easy-to-use tool to check the sensitivity also for mixture models under different prior distributions for the mixing measure in the popular class of Gibbs-type priors (Gnedin and Pitman, 2006; De Blasi et al., 2015) as well as for other approximations arising from different divergences or distances.

To conclude we believe the work by Giordano, Liu, Jordan, and Broderick can stimulate further computational research in the Bayesian community and be applied in many practical situations. We commend them one more time for a remarkable paper.

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *Annals of Statistics*, 2: 1152–1174. MR0365969. doi: https://doi.org/10.1214/aos/1176342871. 346

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022). "Clustering consistency with Dirichlet process mixtures." *Biometrika*, in press. doi: https://doi.org/10.1093/biomet/asac051. 345

Cai, D., Campbell, T., and Broderick, T. (2021). "Finite mixture models do not reliably learn the number of components." In *International Conference on Machine Learning*, PMLR, volume 139, 1158–1169. 346

Chandra, N. K., Canale, A., and Dunson, D. B. (2020). "Escaping the curse of dimensionality in Bayesian model based clustering." *Preprint at arXiv:2006.02700*. doi: https://doi.org/10.48550/arXiv.2006.02700. 346

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37: 212–229. doi: https://doi.org/10.1109/TPAMI.2013.217. 345, 346

Fortini, S. and Petrone, S. (2016). *Predictive distribution (de Finetti's view)*, 1–9. Wiley StatsRef: Statistics Reference Online. doi: https://doi.org/10.1002/9781118445112.stat07831. 345

Gnedin, A. V. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences* , 138: 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 346

Liu, J. S. (1996). "Nonparametric hierarchical Bayes via sequential imputations." *Annals of Statistics*, 24: 911–930. MR1401830. doi: https://doi.org/10.1214/aos/1032526949. 346

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. density estimates." *Annals of Statistics*, 12: 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 345

McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). "Nonparametric empirical Bayes for the Dirichlet process mixture model." *Statistics and Computing*, 16: 5–14. MR2224185. doi: https://doi.org/10.1007/s11222-006-5196-2. 346

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113: 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 345

Nobile, A. (1994). "Bayesian Analysis of Finite Mixture Distributions." Ph.D. thesis, Carnegie Mellon University. MR2692049. 345

Petralia, F., Rao, V., and Dunson, D. B. (2012). "Repulsive mixtures." In *Advances in Neural Information Processing Systems*, volume 25, 1889–1897. 346

Contributed Discussion

Xenia Miscouridou* and Francesca Panero†

We congratulate the authors for the interesting and novel work on the evaluation of the sensitivity of a set of stick-breaking priors via mean-field variational Bayes. The paper focuses on Dirichlet process mixtures (DPM), a popular prior distribution in many applications. The widespread use of DPMs makes it vital to be able to understand their properties and the implications of their use. This method has the potential to become part of the toolkit of statisticians who would like to pursue applications under the Bayesian nonparametric (BNP) framework.

Our discussion focuses on possible extensions of the current work to non stick-breaking priors. We motivate why these random probability measures deserve to be considered for a similar sensitivity analysis and suggest a possible way to adapt the framework of Giordano et al.'s using some recently developed finite approximations of completely random measures.

The authors provide a computational tool to quickly and automatically assess the sensitivity to prior specification of variational Bayes (VB) approximations in the particular case of some stick-breaking priors. They focus on the canonical Dirichlet process mixture model, heavily used in topic modelling and clustering, and suggest that the methods apply directly to any discrete BNP model that admits a truncated stick-breaking construction with independent and identically distributed (iid) proportions $(\nu_k)_k$. The Dirichlet process (DP) is arguably the most widely used discrete random probability measure admitting a stick-breaking representation. It belongs to a wider family of species sampling models known as Gibbs-type priors (see, for example, De Blasi et al. (2013)), which are characterised by a particular form of the exchangeable partition probability function. Other models within this family are the Pitman-Yor (PY) process, the normalised σ-stable process, the normalised generalised gamma process (NGGP) and the uniform process (Wallach et al. (2010)).

Gibbs-type priors do not necessarily admit a stick-breaking representation. As already mentioned in the Invited Discussion by J. Griffin and M. Kalli, an exciting result would be to obtain an extension of the sensitivity analysis provided by Giordano et al.'s to the broader family of Gibbs-type models. Expanding on this, we will highlight in the following paragraphs the notable clustering properties of some random probability measures other than the Dirichlet process, and suggest a possible way to address the sensitivity analysis despite the lack of stick-breaking representations.

The generalised gamma process (GGP) (Hougaard, 1986; Brix, 1999), also known

*Department of Mathematics, Imperial College London, 180 Queen's Gate, South Kensington, London SW7 2AZ, x.miscouridou@imperial.ac.uk

†Department of Statistics, London School of Economics and Political Science, 69 Aldwych, London WC2B 4RR, f.panero@lse.ac.uk

as (exponentially) tilted stable process, has mean measure

$$\rho(dw) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\tau w} dw, \tag{1}$$

where $\sigma \in (0,1)$ and $\tau \geq 0$, or $\sigma \leq 0$ and $\tau > 0$.

Clustering models based on the DP or PY priors can only describe clusters whose size grows linearly with the sample size n. Di Benedetto et al. (2021) propose a class of random partition models based on the GGP which is able to generate partitions whose cluster sizes grow sublinearly with n, a property known as microclustering. In particular, their model offers a power-law growth of cluster sizes with exponent in $(0,1)$. While Di Benedetto et al. (2021) employed an MCMC approach for inference, a variational approach has computational and practical advantages. For a variational framework one needs to consider approximations to these distributions. Lee et al. (2016, 2017) proceed in this direction by introducing and using finite dimensional approximations of the GGP (and other infinite measures). Precisely, they use the BFRY (Devroye and James, 2014) distributions to approximate the infinite measures for power-law mixture models and graphs with power-law degree distribution within a mean-field variational inference framework. For some context on BFRY distributions,[1] recall that a BFRY$(\tau), \tau \in (0,1)$ random variable X is characterised by a density function $f_\tau(x) = \tau(1-e^{-x})/(\Gamma(1-\tau)x^{1+\tau})$, $x > 0$. It is infinitely divisible and can be conveniently sampled as a ratio of gamma and uniform random variables. Heading a step further, Lee et al. (2022) generalise the BFRY priors giving more generic series representations and iid approximations for both the GGP and stable beta process. This suggests the following question: can we adapt the proposed sensitivity toolbox to the case of series representations and iid approximations proposed in Lee et al. (2017, 2022) to cover these interesting applications? In this way one would obtain a sensitivity analysis for microclustering or other applications of the GGP which have a power-law behaviour.

Recently, there was a lot of attention on graph modelling with power-law behaviour as these can model real-world graphs with node heterogeneity. Sparse random graphs with power-law degree distributions were originally introduced by Caron and Fox (2017) who used a GGP process prior on the network node parameters. A series of papers (Herlau et al. (2016); Miscouridou et al. (2018); Todeschini et al. (2020); Naik et al. (2021, 2022)) followed, expanding the properties and types of graphs. In all of these works, the parameter σ in Eq (1) is crucial as it tunes the sparsity of the graph and the degree heterogeneity (power-law), therefore a desirable result would be to come up with a similar computational toolbox to evaluate the sensitivity to the GGP process prior focusing on σ for graph modelling.

References

Bertoin, J., Fujita, T., Roynette, B., and Yor, M. (2006). "On a particular class of self-decomposable random variables: the durations of Bessel excursions straddling

[1]The name BFRY was coined in Devroye and James (2014) after the work of Bertoin et al. (2006), who used this random variable in the study of excursion duration in Bessel processes.

independent exponential times." *Probability and Mathematical Statistics*, (26): 315–366. MR2325310. 349

Brix, A. (1999). "Generalized gamma measures and shot-noise Cox processes." *Advances in Applied Probability*, 31(4): 929–953. MR1747450. doi: https://doi.org/10.1239/aap/1029955251. 348

Caron, F. and Fox, E. (2017). "Sparse Graphs using Exchangeable Random Measures." *Journal of the Royal Statistical Society B*, 79: 1295–1366. Part 5. MR3731666. doi: https://doi.org/10.1111/rssb.12233. 349

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 212–229. 348

Devroye, L. and James, L. F. (2014). "On simulation and properties of the stable law." *Statistical Methods and Applications,*, 3(23): 307–343. MR3233961. doi: https://doi.org/10.1007/s10260-014-0260-0. 349

Di Benedetto, G., Caron, F., and Teh, Y. W. (2021). "Nonexchangeable random partition models for microclustering." *Annals of Statistics*, 49(4): 1931–1957. MR4319236. doi: https://doi.org/10.1214/20-aos2003. 349

Herlau, T., Schmidt, M. N., and Mørup, M. (2016). "Completely random measures for modelling lock-structured sparse networks." *Advances in Neural Information Processing Systems*, 29. 349

Hougaard, P. (1986). "Survival models for heterogeneous populations derived from stable distributions." *Biometrika*, 73(2): 387–396. MR0855898. doi: https://doi.org/10.1093/biomet/73.2.387. 348

Lee, J., Heaukulani, C., Ghahramani, Z., James, L. F., and Choi, S. (2017). "Bayesian inference on random simple graphs with power law degree distributions." In *Proceedings of the 34th International Conference on Machine Learning*. 349

Lee, J., James, L. F., and Choi, S. (2016). "Finite-dimensional BFRY priors and variational Bayesian inference for power-law models." In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. 349

Lee, J., Miscouridou, X., and Caron, F. (2022). "A unified construction of series representations and iid approximations of completely random measures." *Bernoulli*. 349

Miscouridou, X., Caron, F., and Teh, Y. W. (2018). "Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data." *Advances in Neural Information Processing Systems*, 31. 349

Naik, C., Caron, F., and Rousseau, J. (2021). "Sparse networks with core-periphery structure." *Electronic Journal of Statistics*, 15(1): 1814–1868. MR4255305. doi: https://doi.org/10.1214/21-ejs1819. 349

Naik, C., Caron, F., Rousseau, J., Teh, Y. W., and Palla, K. (2022). "Bayesian Nonparametrics for Sparse Dynamic Networks." *European Conference on Machine Learning and Data Mining (ECML PKDD)*. 349

Todeschini, A., Miscouridou, X., and Caron, F. (2020). "Exchangeable random measures for sparse and modular graphs with overlapping communities." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2): 487–520. MR4084173. doi: https://doi.org/10.1111/rssb.12363. 349

Wallach, H. M., Jensen, S. T., Dicker, L., and Heller, K. A. (2010). "An Alternative Prior Process for Nonparametric Bayesian Clustering." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 348

Contributed Discussion

David Banks* and Subarup Guha†

Abstract. What is the probability of realizing a distribution from a stick-breaking process that falls outside an ϵ-ball on the base measure?

The discussion paper is brilliant and important. It informs prior choice in the context of stick-breaking Bayesian analyses. Calculating derivatives in the variational Bayesian posterior approximations is especially cool. But we beg indulgence to discuss an adjacent problem whose solution would have implications for concentration of measure and the sensitivity of stick-breaking priors.

Let the vector of probabilities $(p_1, \ldots, p_n)$ have the n-dimensional Dirichlet distribution, $\mathcal{D}_n(\alpha/n, \ldots, \alpha/n)$. Let $\theta_1^*, \ldots, \theta_n^*$ be i.i.d. draws from the base distribution H that are independent of $p_1, \ldots, p_n$. Then $F_n \stackrel{L}{=} \sum_{j=1}^n p_j \delta_{\theta_j^*}$ where $\stackrel{L}{=}$ denotes "equal in distribution" and δ_θ is a point mass at θ. Ishwaran and Zarepour (2000) shows $F_n \stackrel{L}{\to} F$ as $n \to \infty$. What is the chance of a realization that is some specified distance from H?

Draw a distribution F from the Dirichlet process $DP(\alpha H)$, with H the base measure and $\alpha > 0$ the mass parameter. The sup norm of F and H is $d_K(F, H) = \sup_{x \in \mathbb{R}} |F(x) - H(x)|$. For $\epsilon \in (0, 1)$, we seek $\Pr[d_K(F, H) \leq \epsilon]$.

We know $\Pr[d_K(F_n, H) \leq \epsilon]$ approximates $\Pr[d_K(F, H) \leq \epsilon]$ for large n. So $F_n \stackrel{L}{=} \sum_{j=1}^n q_j \delta_{\theta_{(j)}}$, where $\theta_{(1)} \leq \theta_{(2)} \ldots \leq \theta_{(n)}$ are the order statistics of n i.i.d. draws from H and $(q_1, \ldots, q_n) \sim \mathcal{D}_n(\alpha/n, \ldots, \alpha/n)$. Since F_n is discrete, the sup norm equals $\max_{j=1}^n |F_n(\theta_j) - H(\theta_j)|$, which equals $\max_{j=1}^n |r_j - H(\theta_{(j)})|$ with $r_j = \sum_{k=1}^j q_k$. Since $H(\theta)$ is uniformly distributed whenever $\theta \sim H$, the distribution of $\max_{j=1}^n |r_j - H(\theta_j)|$ is the same distribution as that of $\max_{j=1}^n |r_j - U_j|$ provided $U_j \stackrel{i.i.d.}{\sim} U(0, 1)$ for $j = 1, \ldots, n$. Thus $\Pr[d_K(F_n, H) \leq \epsilon]$ equals $\Pr[\cap_{j=1}^n A_j]$ where the event $A_j = \left[|r_j - U_j| \leq \epsilon\right]$ and the probability for the Kolmogorov sup norm does not depend on H.

Let $\boldsymbol{r} = (r_1, \ldots, r_n)$. Because $\Pr[\cap_{j=1}^n A_j] = \int \Pr\left[\cap_{j=1}^n A_j \mid \boldsymbol{r}\right] [\boldsymbol{r}]\, d\boldsymbol{r}$, and since it is easy to generate samples of $\boldsymbol{r}$, one can use Monte Carlo to approximate $\Pr[\cap_{j=1}^n A_j]$. But direct calculation is impossible since evaluation of $\Pr\left[\cap_{j=1}^n A_j \mid \boldsymbol{r}\right]$ is equivalent to computing the volume of an n-dimensional simplex with vertices determined by $\boldsymbol{r}$:

$$\Pr\left[\cap_{j=1}^n A_j \mid \boldsymbol{r}\right] = n! \int_{L_n}^{M_n} \ldots \int_{L_1}^{M_1} dz_1 \ldots dz_n$$

where the integral limits (which correspond to the vertices) are $L_j = \max\{0, r_j - \epsilon\}$, $M_j = \min\{z_{j+1}, r_j + \epsilon\}$ for $j < n$ and $M_n = \min\{1, r_n + \epsilon\}$.

*Department of Statistical Science, Duke University, and Department of Biostatistics, University of Florida Gainesville, banks@stat.duke.edu

†Department of Statistical Science, Duke University, and Department of Biostatistics, University of Florida Gainesville, s.guha@ufl.edu

Let $G_\alpha(\cdot)$ be a gamma process with $\alpha \in [0,1]$. Then $G_\alpha(0) = 0$; for any two indices $0 \le t_1 < t_2 \le 1$, the process increment $G_\alpha(t_2) - G_\alpha(t_1)$ is gamma with shape parameter $\alpha(t_2 - t_1)$ and scale parameter 1; process increments for disjoint intervals are mutually independent. The $G_\alpha(\cdot)$ is a.s. continuous and increasing. Let $B_\alpha(t) = G_\alpha(t)/G_\alpha(1)$. Then $B_\alpha(t)$ is Beta $(\alpha t, \alpha(1-t))$ for $0 < t < 1$. It has a.s. continuous sample paths that monotonically increase from 0 to 1.

Theorem 1. *Let $F \sim DP(\alpha H)$. If H is continuous, then the sup norm is $d_K(F,H) \stackrel{L}{=} \sup_{0\le t\le 1} |B_\alpha(t) - U(t)|$ for U the uniform distribution on $[0,1]$.*

Proof. For $n \ge 2$, let $-\infty = x_0 < x_1 < \cdots < x_n = \infty$. The sets $A_i = (x_{i-1}, x_i]$, where $i = 1, \ldots, n$, form a partition of the real line. The vector $(F(A_1), \ldots, F(A_n))$ has the n-dimensional Dirichlet distribution, $\mathcal{D}_n\,(\alpha H(A_1), \ldots, \alpha H(A_n))$.

Consider the process $B^*_\alpha(x) = B_\alpha(H(x))$. Then $\lim_{x\to-\infty} B^*_\alpha(x) = 0$, $\lim_{x\to\infty} B^*_\alpha(x) = 1$, and $B^*_\alpha(x)$ is distributed as Beta $(\alpha H(x), \alpha[1 - H(x)])$. The vector $(B^*_\alpha(x_1), B^*_\alpha(x_2) - B^*_\alpha(x_1), \ldots, B^*_\alpha(x_n) - B^*_\alpha(x_{n-1}))$ has the distribution $\mathcal{D}_n\,(\alpha H(A_1), \ldots, \alpha H(A_n))$. Since $\cup_{j=1}^i A_j = (-\infty, x_i]$ for $i = 1, \ldots, n$, then $(F(x_1), \ldots, F(x_n)) \stackrel{L}{=} (B^*_\alpha(x_1), \ldots, B^*_\alpha(x_n))$. $B^*_\alpha(\cdot)$ has the same finite-dimensional distributions as $DP(\alpha H)$ but is a.s. continuous.

Since H is continuous, the mapping $H : R \to [0,1]$ is invertible. Thus

$$\max_{i=1,\ldots,n} \left|F(x_i) - H(x_i)\right| \stackrel{L}{=} \max_{i=1,\ldots,n} \left|B^*_\alpha(x_i) - H(x_i)\right| = \max_{i=1,\ldots,n} \left|B_\alpha(t_i) - H \circ H^{-1}(t_i)\right|$$

where $t_i = H(x_i)$, and the result follows from the beta process. For integers $n \ge 2$ and for $i = 0, \ldots, n$, set $t_i^{(n)} = i/n \in [0,1]$, let $x_i^{(n)} = H^{-1}(t_i^{(n)}) \in R$. We see

$$\max_{1\le i\le n} \left|F(x_i^{(n)}) - H(x_i^{(n)})\right| \stackrel{L}{=} \max_{1\le i\le n} \left|B_\alpha(t_i^{(n)}) - H \circ H^{-1}(t_i^{(n)})\right| \quad \text{for } n \ge 2. \tag{0.1}$$

For $i \ge 1$, $\lim_{n\to\infty}(t_{i+1}^{(n)} - t_i^{(n)}) = \lim_{n\to\infty} 1/n = 0$. So $\lim_{n\to\infty}(x_{i+1}^{(n)} - x_i^{(n)}) = 0$ because H^{-1} is continuous. Taking $n \to \infty$ in (0.1) gives $\sup_{x\in R} |F(x) - H(x)| \stackrel{L}{=} \sup_{0\le t\le 1} |B_\alpha(t) - H \circ H^{-1}(t)|$ from right-continuity of F, H, and B_α. □

Bayesian use of Dirichlet processes requires selection of H and α. That choice should account for the probability that a realization is near H. We hope that one of the super-smart readers of this journal will find a closed form solution for $\Pr[d_K(F,H) \le \epsilon]$.

References

Ishwaran, H. and Zarepour, M. (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." *Biometrika*, 87(2): 371–390. MR1782485. doi: https://doi.org/10.1093/biomet/87.2.371. 352

Acknowledgments

This work was partially supported by a grant from the Office of Naval Research, ONR 6000012277.

Rejoinder

Ryan Giordano[*,¶], Runjing Liu[†,¶], Michael I. Jordan[‡], and Tamara Broderick[§]

1 Introduction

We feel very grateful to have our work carefully read and commented on by so many insightful respondents. We would like to thank Professor Steel and the editorial board of Bayesian Analysis for selecting our work and making this discussion possible. Statistical robustness is a venerable topic of conversation and we have no doubt that the present discussion will continue far into the future.

We might roughly categorize points made in the responses as follows:[1]

1. Would a different model or summary statistic be more or less robust than the Dirichlet process (DP) and number of clusters?
2. Can (or should) one form variational Bayes (VB) approximations to different models from the BNP literature?
3. Can one form a local sensitivity metric to different quantities of interest, modeling or fitting parameters, different posterior approximation procedures, or some combination of all of these?

Questions in category (1) and (2) are natural and important, since our work is based on arguably the most canonical Bayesian nonparametric prior (the DP), and a fairly vanilla VB approximation (a mean field approximation to a truncated stick-breaking representation). The evaluation of our robustness ideas with respect to a wider range of priors and approximations is certainly warranted. Nevertheless, in the present rejoinder we will focus on questions in category (3), largely because we feel that our use of local sensitivity metrics constitutes our work's most distinctive contribution.

Can one form a local robustness metric for a particular problem? In Section 2 of the present rejoinder, we will argue: typically, yes, quite directly, in Markov chain Monte Carlo (MCMC) applications as well as VB. In Section 2.1 we derive local robustness metrics for a select few settings that were described in the responses. After reading

[*]Department of EECS, MIT, 77 Massachusetts Ave., 32-D762, Cambridge, MA 02139, rgiordan@gmail.com
[†]Department of Statistics, 367 Evans Hall, UC Berkeley, Berkeley, CA 94720
[‡]Department of Statistics, 367 Evans Hall, UC Berkeley, Berkeley, CA 94720
[§]Department of EECS, MIT, 77 Massachusetts Ave., 32-D762, Cambridge, MA 02139
[¶]Equal contribution.

[1]Please think of these categories as atom locations from an Indian buffet process, not a Chinese restaurant process; respondents engaged at times with multiple categories simultaneously. For example, **Gil-Leyva and Mena** ask how to form a VB approximation to an exchangeable stick breaking (ESB) prior (item 2) in order to form a local sensitivity metric (item 3) to assess whether the ESB prior is robust for pointwise density estimation (item 1).

Section 2, we hope that all readers of this rejoinder feel able and empowered to form and investigate local robustness metrics for their own particular problems.

However, in the subsequent Section 3, we will argue that simply forming a local robustness metric is not enough: the hard work is showing that it is useful. Computability, interpretability, and the ability of a local robustness metric to *extrapolate* well, are more important — and more difficult to establish — than mere computation of derivatives. It is this work of establishing usefulness that we have endeavored to undertake in the present paper, and to which we wish to call attention as a foundation for further work.

As might be expected in a topic as established as robustness, the points made in this rejoinder are not new, and have in fact been argued in the past by many of our own respondents. Nevertheless, we hope that by emphasizing the relative ease of computing derivatives and the relative difficulty of showing their utility in particular contexts, we can help advance the research agenda in this important and challenging area.

2 What does it take to do local robustness?

A great deal of statistical inference—including Bayesian statistics—fixes some hyperparameter ω and then performs posterior inference using some combination of two types of estimators:

- The solution to a system of estimating equations: $\hat{\theta}_{\text{opt}} := \theta$ such that $G(\theta, \omega) = 0$.
- A posterior moment from a density known up to a constant: $\hat{\theta}_{\text{samp}} := \underset{\mathcal{P}(\zeta|\omega)}{\mathbb{E}}[H(\zeta)]$.

An example of $\hat{\theta}_{\text{opt}}$ could be the parameter that sets the gradient of a VB loss function to zero (as in our paper), and an example of $\hat{\theta}_{\text{samp}}$ could be a posterior mean. In practice, we may not be able to compute either exactly: $\hat{\theta}_{\text{opt}}$ might be approximated using numerical optimization, and $\hat{\theta}_{\text{samp}}$ may be approximated using Markov chain Monte Carlo (MCMC). Below, we will briefly discuss the consequences of using approximations, but assume for the moment that we can compute $\hat{\theta}_{\text{opt}}$ and $\hat{\theta}_{\text{samp}}$ to a desired accuracy. A practitioner might then ask, "what would happen if ω had taken on some different value?" The techniques of local robustness approximately answer this question by forming a series approximation using derivatives of the maps $\omega \mapsto \hat{\theta}_{\text{opt}}(\omega)$ and $\omega \mapsto \hat{\theta}_{\text{samp}}(\omega)$.

There are simple, general formulas for both these derivatives, under certain common (but not universal) regularity conditions. For notational simplicity, take ω to be a scalar for the moment. Furthermore, let us take $\hat{\theta}_{\text{opt}}(\omega)$ to be finite-dimensional, $\mathcal{P}(\zeta|\omega)$ to be defined as a Radon-Nikodym derivative with respect to a common dominating measure for all ω, assume that we can exchange integration and differentiation as needed, and assume all needed partial derivatives exist. Then

$$\left.\frac{d\hat{\theta}_{\text{opt}}(\omega)}{d\omega}\right|_{\omega_0} = \left(\left.\frac{\partial G(\theta,\omega)}{\partial\theta}\right|_{\theta=\hat{\theta}_{\text{opt}}(\omega_0),\omega=\omega_0}\right)^{-1} \left.\frac{\partial G(\theta,\omega)}{\partial\omega}\right|_{\theta=\hat{\theta}_{\text{opt}}(\omega_0),\omega=\omega_0} \tag{1}$$

and

$$\left.\frac{d\hat{\theta}_{\text{samp}}(\omega)}{d\omega}\right|_{\omega_0} = \underset{\mathcal{P}(\zeta|\omega_0)}{\text{Cov}}\left(H(\zeta), \left.\frac{\partial \log \mathcal{P}(\zeta|\omega)}{\partial\omega}\right|_{\omega=\omega_0}\right). \tag{2}$$

These two formulas have been noted many times in the literature, though we feel it is worth calling attention to the simplicity of their form. The estimating equation derivative in eq. (1) is formed using the implicit function theorem (Krantz and Parks, 2012) and is used, explicitly or implicitly, in many local robustness works (Hampel, 1974; Thomas and Cook, 1989; Hattori and Kato, 2009; Shi et al., 2016), as well as our own present paper. The sampling derivative in eq. (2) is formed by differentiating under the integral using a dominated convergence theorem (Billingsley, 2008, Theorem 5.4) and appears widely, in various forms, in the local Bayesian robustness literature and beyond (Diaconis and Freedman, 1986; Ruggeri and Wasserman, 1993; Gustafson, 1996; Mohamed et al., 2020).

Importantly, eqs. (1) and (2) can be computed nearly automatically using automatic differentiation, using only the original solution $\hat{\theta}_{\text{opt}}(\omega_0)$ or the ability to compute moments of $\mathcal{P}(\zeta|\omega_0)$. Furthermore, higher-order derivatives can be computed mechanically by repeatedly applying eqs. (1) and (2) to themselves. For example, higher-order versions of eq. (1) can be found in Giordano et al. (2019a). In practice, one can use a corresponding numerical approximation to $\hat{\theta}_{\text{opt}}(\omega_0)$ or draws from $\mathcal{P}(\zeta|\omega_0)$ to approximate the derivatives. As observed by **Griffin and Kalli**, the key practical difficulty with eq. (1) is the solution of a linear system, and the key practical difficulty with eq. (2) is Monte Carlo error.

One might contrast eqs. (1) and (2) with approaches that differentiate the optimization procedure or the sampling procedure directly, as in chapter 6 of Maclaurin (2016) and Jacobi et al. (2018), respectively, both of which require considerable bespoke computational effort, even with automatic differentiation. The simplicity of eqs. (1) and (2) comes at a cost, however, of assuming (respectively) that $\hat{\theta}_{\text{opt}}$ actually solves the estimating equation, or that we are actually able to approximate draws from $\mathcal{P}(\zeta|\omega_0)$. Studying eqs. (1) and (2) in the presence of violations of these assumptions is exciting and ongoing work, most notably in the setting of optimization (Bae et al., 2022).

Although eqs. (1) and (2) apply to scalar ω, they extend readily to multivariate and even functional derivatives, since one can use scalar derivatives to differentiate along a path in a multivariate space. Different notions of "derivative" differ only in how they conceptually bundle these path derivatives together—as a basis for a gradient in finite dimensional vector spaces (Fleming, 2012), as a basis for a tangent plane in a geometric perspective (McInerney, 2013; Murray and Rice, 1993), as Hadamard or Fréchet derivatives in infinite-dimensional spaces according to the smoothness of the underlying function (Averbukh and Smolyanov, 1967; Zeidler, 1986). In each case, however, for a particular path, the derivative is always formally the same, and computable by eqs. (1) and (2).

Given estimators of the form $\hat{\theta}_{\text{opt}}$, $\hat{\theta}_{\text{samp}}$, or some smooth combination of them, the capacity to imagine a parameterized set of perturbations of interest (and a class of

univariate paths through it), one can form local robustness measures—even for infinite-dimensional perturbations—using little more than eqs. (1) and (2), univariate calculus, and the chain rule. In the following section, we will demonstrate this point in a few of the settings described by the respondents.

2.1 Some requested derivatives

We now demonstrate our claim that derivatives are typically straightforward to compute by doing so for several of the settings requested by our respondents: case influence measures (Cook, 1977), exchangeable stick breaking (ESB) processes (Gil-Leyva and Mena, 2021), empirical Bayes (EB) procedures (McAuliffe et al., 2006), and robustness to the loss function. For ESB processes and EB we will discuss some interesting challenges that arise from working with infinite-dimensional priors in BNP models.

Many respondents noted that our (classical) approach to deriving local robustness measures extends readily to other settings. Readers who are similarly convinced that it is not difficult to compute local robustness derivatives for a wide range of applications, for both MCMC and optimization-based statistical procedures, can safely skip to Section 3.

For this short rejoinder we have selected only a few settings from the responses to address in detail, and we have chosen to prioritize the settings that are most unlike the results in our paper. Unfortunately, doing so means forgoing discussion of some ideas which seem particularly promising to us, such as the proposal of **Miscouridou and Panero** to apply local robustness techniques to the VB approximations for generalized gamma processes given in Lee et al. (2016).

Throughout this section, we will take ζ to denote all parameters of a model and X to denote observed data, so that the posterior is $\mathcal{P}(\zeta|X)$. Let $\phi(\zeta)$ be some quantity of interest. Note that, in eq. (2), we need only differentiate $\log \mathcal{P}(\zeta, X|\omega)$ rather than $\log \mathcal{P}(\zeta|X, \omega)$, because the normalizing constant $\mathcal{P}(X|\omega)$ does not depend on ζ and does not contribute to the covariance.

Case influence **MacEachern and Lee** connect our work to a long history of frequentist and Bayesian "case influence" literature. This literature attempts to quantify the importance of individual datapoints or groups of datapoints on a particular inferential procedure. **MacEachern and Lee** point to a set of works, beginning with Cook (1977), which is particularly concerned with "outliers" or "gross errors" as popularized by Huber (1964).[2] Indeed, the idea of using local approximations to robustness under generic data perturbations goes back even farther—at least as far as von Mises (1947)—and has been employed for asymptotic theory (Serfling, 1980; Shao and Tu, 2012; van der Vaart and Wellner, 1996), design and analysis of robust estimators (Hampel, 1974), approximation of cross-validation in machine learning (Koh and Liang, 2017; Giordano et al., 2019b) and more.

To connect this broad literature to our work, we can augment each datapoint with a scalar-valued weight, w_n, in such a way that $w_n = 1$ represents no change, and $w_n = 0$

[2] A short historical account of this branch of robust statistics is given by Stigler (2010).

represents omitting the datapoint from the model. Specifically, letting $w = (w_1, \ldots, w_N)$ and $X = (X_1, \ldots, X_N)$, we can write the log likelihood in a Bayesian model as

$$\log \mathcal{P}(X, \zeta | w) = \sum_{n=1}^{N} w_n \log \mathcal{P}(X_n | \zeta) + \log \mathcal{P}(\zeta),$$

with $\mathcal{P}(\zeta | X, w)$ representing the corresponding posterior. With unit weights, we recover the original posterior: $\mathcal{P}(\zeta | X, w = (1, \ldots, 1)) = \mathcal{P}(\zeta | X)$. When $w_n = 0$ but all other weights are one, data point n is left out. Similarly, one can drop or replicate any set of data points using the appropriate configuration of zeros, ones, or other integers.

The advantage of writing $\log \mathcal{P}(\zeta | X, w)$ in this way is that we can take a particular w_n to be our hyperparameter ω and apply eqs. (1) and (2) to form a local approximation to leaving out (or replicating) sets of datapoints. The form of the derivative for estimating equations resulting from eq. (1) is the well-known empirical influence function for M-estimators (see, e.g., eq. 2.3.5 of Hampel et al. (1986)). Perhaps less widely known is the corresponding result for MCMC estimators, which is

$$\left. \frac{\partial \mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X,w)} [\phi(\zeta)]}{\partial w_n} \right|_{w=(1,\ldots,1)} = \mathop{\mathrm{Cov}}_{\mathcal{P}(\zeta|X)} \left(\phi(\zeta), \log \mathcal{P}(X_n | \zeta) \right). \tag{3}$$

The right-hand side of eq. (3) can be estimated from MCMC samples. Then the quantity given in eq. (3) is precisely the "Bayesian empirical influence function," evaluated at X_n, for the statistic $\mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X)} [\phi(\zeta)]$. As with the frequentist influence function, eq. (3) may be used to approximate all sorts of case deletion schemes from both the frequentist and Bayesian literature—as long as one can show that it provides a good approximation to the effect of actually removing the points.[3]

Dirichlet-driven ESB models **Gil-Leyva and Mena** ask about local sensitivity analysis for Dirichlet-driven exchangeable stick breaking (ESB) models (Gil-Leyva and Mena, 2021). The joint stick distribution in an ESB model is controlled by a parameter $\rho \in [0, 1]$ that smoothly transitions between stick-breaking priors with independent and identically distributed sticks and stick-breaking priors for which all the sticks take a common value. **Gil-Leyva and Mena** take the quantity of interest to be the posterior estimate of the density of the data generating distribution evaluated at a point—a quantity which we can call $\phi(\zeta)$—and ask how $\mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X,\rho)} [\phi(\zeta)]$ depends on ρ. **Gil-Leyva and Mena** run an MCMC chain, but then, in order to compute local robustness measures, attempt to construct a VB approximation to this posterior, observing that one would

[3]We should note that a first-order approximation is inadequate when taking some form of KL divergence from the original posterior as the quantity of interest, as is done in much of the literature cited by **MacEachern and Lee** (e.g., Johnson and Geisser, 1983; McCulloch, 1989; Carlin and Polson, 1991; Thomas et al., 2018). This KL divergence is minimized at $w = (1, \ldots, 1)$, so the first derivative with respect to the weights is zero, and one must form a local approximation using a second-order derivative. However, all our comments in the present rejoinder, particularly Section 3, apply to local second-order approximations as well as to first-order approximations.

need either to make a (potentially limiting) mean field assumption on the sticks or deal with a computationally intractable normalizing constant.

We will avoid the question of how to construct a VB approximation in their setting, and derive instead a local sensitivity measure that can be used with an MCMC chain—as long as the stick-breaking distribution can be effectively truncated at K sticks for some finite K. Let the truncated stick lengths be denoted by $v_1, \ldots, v_K$. We can imagine several ways to truncate an ESB model, but for the present purposes, one would need to be able to sample from the truncated model, and the prior $\mathcal{P}(v_1, \ldots v_K|\rho)$ would need to be tractable and smooth for any draw from the MCMC chain.[4] By eq. (2) we then have

$$\left.\frac{\partial \mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X,\rho)}[\phi(\zeta)]}{\partial \rho}\right|_{\rho=\rho_0} = \mathop{\mathrm{Cov}}_{\mathcal{P}(\zeta|X,\rho_0)}\left(\phi(\zeta), \left.\frac{\partial \log \mathcal{P}(v_1, \ldots, v_K|\rho)}{\partial \rho}\right|_{\rho=\rho_0}\right). \tag{4}$$

The preceding sampling covariance can in principle be approximated from MCMC samples, with no need to form a VB approximation.

Carefully considering the implications of truncating ESB models is beyond the scope of this rejoinder, but it is worth noting the challenges for local robustness if one does not truncate, especially since Gil-Leyva and Mena (2021) use a slice sampler and do not truncate the stick-breaking distribution. If the log prior contained an infinite number of terms, it is not obvious that one could apply the dominated convergence to derive eq. (2). There exist sampling schemes that in fact sample only a finite number of sticks without truncation; see, e.g., Gil-Leyva and Mena (2021) and Walker (2007). Similarly, one might hope that one could apply eq. (2) without truncation by conditioning on auxiliary random variables in the $\log \mathcal{P}(\zeta|\omega)$ term in eq. (2). But this term cannot be conditional on quantities that are random in $\mathcal{P}(\zeta|\omega)$. Additionally, the truncation will, in general, have an effect; here, the quantity of interest (the posterior estimate of the data density at a point) has nonzero correlation with *all* sticks. Although the posterior density at a point plausibly has diminishing correlation with sticks that come later in the process, one could design adversarial quantities of interest—e.g., the value of the 10,000-th stick—for which truncation would be quite inaccurate. Developing tractable sensitivity measures for infinite-dimensional posteriors is an interesting problem, though we suspect that eq. (4) will still be informative using straightforward finite truncation, especially given that any error in the linear approximation may well be larger than the error induced by truncation.

Empirical Bayes **Rebaudo, Fasano, Franzolini, and Müller** ask whether Empirical Bayes (EB) methods for setting a DP prior might be more robust. One might answer such a question empirically by forming local robustness measures for EB procedures, which we will now undertake.

[4]Note that for the un-truncated Dirichlet-driven ESB model, the density of any finite number of sticks is tractable and smooth as a function of ρ: see Appendix E, Section 5 of the supplementary material to Gil-Leyva and Mena (2021) where the needed density is derived as part of a Gibbs sampler for ρ.

Concretely, McAuliffe et al. (2006) rely on an EB procedure to choose the DP concentration parameter. Their EB procedure takes the following general form. Fix some hyperparameter ω, which might be a case weight (see above), a perturbation of the base measure, etc. The EB procedure then finds a prior parameter $\hat{\alpha}$ that satisfies, for some F and G,

$$G(\hat{\alpha}, m(\hat{\alpha}, \omega)) = 0 \quad \text{where} \quad m(\alpha, \omega) := \mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X,\alpha,\omega)}[F(\zeta)]\,. \tag{5}$$

Specifically, to set the concentration parameter of a DP prior using EB, McAuliffe et al. (2006) take α to be the DP concentration parameter, $F(\zeta)$ to denote the number of clusters observed for the dataset X of size N, and $G(\alpha, m) = \sum_{n=1}^{N} \frac{\alpha}{\alpha+n-1} - m$ (see their eq. 8).

EB procedures such as this one take the form of an estimating equation that depends on a posterior moment, which can be differentiated by eqs. (1) and (2) and the chain rule. Note that $\hat{\alpha}$ depends on ω, which we write as $\hat{\alpha}(\omega)$. Additionally, write $\hat{\alpha}_0 := \hat{\alpha}(\omega_0)$ and $m_0 := m(\hat{\alpha}_0, \omega_0)$. To compute a local robustness measure for the posterior expectation of $\phi(\zeta)$, we must compute

$$\left.\frac{\partial \mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X,\hat{\alpha}(\omega),\omega)}[\phi(\zeta)]}{\partial \omega}\right|_{\omega_0},$$

accounting for the ω dependence in both the empirical Bayes procedure and the final posterior expectation. This derivative can be readily computed by applying the chain rule and eqs. (1) and (2). The result, given below in eq. (6), is a bit tedious, but its computation is entirely mechanical (and automatable) and applies to any EB procedure of the form in eq. (5).[5]

$$\begin{aligned}
&\left.\frac{\partial \mathop{\mathbb{E}}_{\mathcal{P}(\zeta|X,\hat{\alpha}(\omega),\omega)}[\phi(\zeta)]}{\partial \omega}\right|_{\omega_0} \\
&\quad = \mathop{\mathrm{Cov}}_{\mathcal{P}(\zeta|X,\hat{\alpha}_0,\omega_0)}\left(\phi(\zeta), \left.\frac{\partial \log \mathcal{P}(\zeta, X|\hat{\alpha}_0, \omega)}{\partial \omega}\right|_{\omega_0} + \left.\frac{\partial \log \mathcal{P}(\zeta, X|\alpha, \omega_0)}{\partial \alpha}\right|_{\hat{\alpha}_0} \left.\frac{d\hat{\alpha}(\omega)}{d\omega}\right|_{\omega_0}\right) \\
&\text{where} \quad \left.\frac{d\hat{\alpha}(\omega)}{d\omega}\right|_{\omega_0} = -\left(\left.\frac{\partial G(\alpha, m(\alpha, \omega_0))}{\partial \alpha}\right|_{\hat{\alpha}_0}\right)^{-1}\left(\left.\frac{\partial G(\hat{\alpha}_0, m)}{\partial m}\right|_{m_0} \left.\frac{\partial m(\hat{\alpha}_0, \omega)}{\partial \omega}\right|_{\omega_0}\right), \\
&\left.\frac{\partial G(\alpha, m(\alpha, \omega_0))}{\partial \alpha}\right|_{\hat{\alpha}_0} = \left.\frac{\partial G(\alpha, m_0)}{\partial \alpha}\right|_{\hat{\alpha}_0} + \left.\frac{\partial G(\hat{\alpha}_0, m)}{\partial m}\right|_{m_0} \left.\frac{\partial m(\alpha, \omega_0)}{\partial \alpha}\right|_{\hat{\alpha}_0}, \\
&\left.\frac{\partial m(\alpha, \omega_0)}{\partial \alpha}\right|_{\hat{\alpha}_0} = \mathop{\mathrm{Cov}}_{\mathcal{P}(\zeta|X,\hat{\alpha}_0,\omega_0)}\left(F(\zeta), \left.\frac{\partial \log \mathcal{P}(\zeta, X|\alpha, \omega_0)}{\partial \alpha}\right|_{\hat{\alpha}_0}\right),
\end{aligned}$$

[5] For compactness, we have suppressed some evaluation notation in this display; for example, we write $\hat{\alpha}_0$ in place of $\alpha = \hat{\alpha}_0$. The evaluation is always done for the parameter with respect to which we are differentiating.

$$\text{and} \quad \left.\frac{\partial m(\alpha_0,\omega)}{\partial\omega}\right|_{\omega_0} = \underset{\mathcal{P}(\zeta|X,\hat{\alpha}_0,\omega_0)}{\text{Cov}}\left(F(\zeta), \left.\frac{\partial \log \mathcal{P}(\zeta, X|\hat{\alpha}_0,\omega)}{\partial\omega}\right|_{\omega_0}\right). \tag{6}$$

One might ask whether eq. (6) can be applied to the EB procedure used by McAuliffe et al. (2006) for the base measure. Unfortunately, since McAuliffe et al. (2006) estimate the base measure nonparametrically, the space of possible base measures is infinite-dimensional, and so one cannot apply eq. (1) directly. Versions of eq. (1) for infinite-dimensional parameters exist (see, e.g., Chapter 4 of Zeidler, 1986), though applying them in practice seems to be challenging and beyond the scope of this short rejoinder. Alternatively, one could represent the base measure using a large but finite basis and apply eq. (6).

Loss functions **MacEachern and Lee** ask whether we can compute sensitivity to the loss function in a Bayesian analysis. Formally, for a posterior $\mathcal{P}(\zeta|X)$ and loss function L, under common regularity conditions we have

$$\hat{\theta}_{\text{loss}} := \underset{\theta}{\text{argmin}} \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}[L(\zeta,\theta)] \quad \Leftrightarrow \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}\left[\left.\frac{\partial L(\zeta,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_{\text{loss}}}\right] = 0. \tag{7}$$

We will consider the common situation described in eq. (7). However, we note that, by exchanging the order of local robustness derivatives and posterior expectations, this approach could be naturally extended to estimators of the form given in Lee and MacEachern (2014), i.e., $\underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}\left[\text{argmin}_\theta \int L(y,\theta)\zeta(dy)\right]$ for a distribution-valued ζ.

Equation (7) defines an estimating equation for $\hat{\theta}_{\text{loss}}$. We can parameterize a path to a different loss function using $L(\zeta,\theta,\omega) = L(\zeta,\theta) + \omega\Delta(\zeta,\theta)$ for some $\Delta(\zeta,\theta)$. Then, apply eq. (1) to eq. (7), and interchange differentiation and integration to get

$$\left.\frac{d\hat{\theta}_{\text{loss}}(\omega)}{d\omega}\right|_{\omega=0} = -\left(\underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}\left[\left.\frac{\partial^2 L(\zeta,\theta)}{\partial\theta\partial\theta}\right|_{\theta=\hat{\theta}_{\text{loss}}}\right]\right)^{-1} \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}\left[\left.\frac{\partial\Delta(\zeta,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_{\text{loss}}}\right]. \tag{8}$$

For example, to estimate the effect of replacing the mean with the median, we could take $L(\zeta,\theta) = \frac{1}{2}(\zeta-\theta)^2$, $\Delta = |\zeta-\theta| - L(\zeta,\theta)$, and

$$\begin{aligned}\text{Median}(\mathcal{P}(\zeta|X)) - \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}[\zeta] &= \hat{\theta}_{\text{loss}}(1) - \hat{\theta}_{\text{loss}}(0) \approx \left.\frac{d\hat{\theta}_{\text{loss}}(\omega)}{d\omega}\right|_{\omega=0}(1-0) \\ &= \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}\left[\mathbb{I}\left(\zeta > \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}[\zeta]\right)\right] - \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}\left[\mathbb{I}\left(\zeta < \underset{\mathcal{P}(\zeta|X)}{\mathbb{E}}[\zeta]\right)\right].\end{aligned} \tag{9}$$

For example, this approximation reasonably asserts that the median will exceed the mean when the posterior is asymmetric, with a greater mass to the right of the mean than to the left. (But we will discuss some of its limitations in Section 3 below.)

Since Δ could be any function satisfying basic regularity conditions, one could in principle use eq. (8) to explore the space of loss functions—if one can believe that

the approximation provided by $\Delta \mapsto \left.\frac{d\hat{\theta}_{\text{loss}}(\omega)}{d\omega}\right|_{\omega=0}$ is a good one uniformly over the candidate set of perturbations Δ. However, again, we do not necessarily recommend this for this particular path through the space of loss functions. On the contrary, we will use this example in Section 3 below as an example of a derivative that may not serve its intended purpose very well.

3 What makes a derivative useful?

In Section 2.1 above, we derived local robustness measures for several settings requested by our respondents, and we expect that readers can readily derive most of the rest for themselves. Have we solved all their problems? Certainly not! To the contrary, we will argue that the computation of derivatives is straightforward, but showing their utility is harder.

In our view, a useful derivative should (at least) satisfy a few related "usefulness desiderata": (1) be readily computable to the desired accuracy, (2) be easily interpretable, and, most importantly, (3) extrapolate well so as to provide a reasonable approximation to the "global robustness" problem. We have endeavored to show that certain derivatives are at least plausibly useful, according to these criteria, for DP priors in VB approximations. In addition to considering Fréchet differentiability—which is, arguably, a rather low bar for a derivative to pass—we primarily demonstrated our local robustness measure's ability to extrapolate through careful experiments and comparison with refitting. In certain situations, such as case influence in large datasets, one can sometimes prove good extrapolation by bounding the second derivative under readily interpretable conditions (as in Giordano et al., 2019b).

Evaluating the usefulness desiderata for the derivatives given in Section 2.1 above is worth doing, and it is the work which constitutes most of the effort. We do not believe that all the results of Section 2.1 will pass the test. Let us focus on the loss function example, though many of these potential problems apply to the other settings as well.

Computability. The expected Hessian inside the inverse in eq. (8) will have Monte Carlo error if estimated with MCMC, which will bias the inverse. Furthermore, if the difference between L and $L + \Delta$ occurs mostly in the tails, the expectation of the derivative of Δ may suffer from high MCMC noise.

Interpretability. The loss function derivative in eq. (8) will behave pathologically as an approximation to losses that are pointwise close to the original loss function, but have very large derivatives near $\hat{\theta}_{\text{loss}}$. Without a judiciously constrained search space to exclude such alternative loss functions, eq. (8) will provide poor guidance when exploring the space of alternative loss functions. Unfortunately, a more complex search space comes at a cost, which is the computational difficulty of optimizing a linear form (i.e., the derivative in eq. (8), viewed as a functional of Δ) over this space.

Extrapolation. The example of the mean and median shows that the derivative eq. (8) may not always extrapolate well in common use cases. Though we know that the mean and median can be arbitrarily different in general, the approximation to the difference in eq. (9) cannot be larger in magnitude than one.

Attempting to investigate and repair these deficiencies—e.g., by improved MCMC sampling, alternative paths through the space of loss functions, and the selection of search sets in the space of loss functions—is an interesting and valuable project, and one that may require considerably more effort than derivation of the local robustness approximation.

The usefulness desiderata sufficed for our objective, which was primarily to provide a tool for quickly exploring the space of stick-breaking priors in a way that is not too computationally or technically burdensome, for a particular quantity of interest. On the other hand, we do not attempt to detect sensitivity of the entire model (e.g., with a whole-distribution divergence measure), we do not assert that a large worst-case derivative implied non-robustness (e.g., the corresponding prior may have looked subjectively unreasonable), and we do not assert that our good results on extrapolation will necessarily hold in very different settings (we primarily showed good extrapolation via experiment). In this sense, our objectives are somewhat different than much of the classical robustness literature, which we see as attempting to provide more universal notions of "robustness." For example, the foundational works of Ruggeri and Wasserman (1993), Basu et al. (1996), Gustafson (1996), to which we are quite indebted, appear to take as their task the *definition* of a single number which can be interrogated, relatively free of context, to ascertain whether a posterior is "robust" or not. This goal is reflected in how their techniques are used, for example, in Basu (2000). A similar goal of finding universal metrics of "data importance" motivates much of the case deletion literature; it is perhaps for this reason that many authors focus on various forms of whole-model KL divergence or likelihood ratios (see, e.g., Johnson and Geisser, 1983; Cook, 1986; Carlin and Polson, 1991). The production of a universally valid local robustness metric requires even stricter conditions on the derivative than our usefulness desiderata; e.g., it must extrapolate well in all directions, its worst-case perturbation must lead to a subjectively reasonable prior, the researcher must actually care about whole-model sensitivity and not just a particular posterior quantity, and so on.

Our context-specific approach and a more universalist approach need not be at odds. On the contrary, the intuition and best practices arising from routine and systematic assessment of prior assumptions might lead, in the end, to better and more universally applicable metrics of robustness. Similarly, asymptotic analysis of the sort advocated by **Ascolani, Catalano, and Prünster** can inform and be informed by the robustness of particular finite-data settings.

Where local approximations fail to satisfy the usefulness desiderata, there is ample room for creativity. The analysis of **Griffin and Kalli**, which both clearly demonstrates a failure of a linear approximation to extrapolate and suggests a solution, seems exemplary to us. Their idea of performing sensitivity analysis separately for a number of local modes seems promising; to their suggestion we might add forming a (second order) approximation to the value of the ELBO at these modes as well, in order to assess when the relative ordering of the modes changes. Local approximations might also augment other schemes, such as suggesting quadratic transforms for the importance sampling techniques of MacEachern and Peruggia (2000).

We hope that researchers will feel empowered by this work to creatively explore the space of model perturbations, relatively unencumbered by the difficulty of deriving local

robustness measures, but attentive to their ability to answer useful questions in their own modeling contexts.

References

Averbukh, V. and Smolyanov, O. (1967). "The theory of differentiation in linear topological spaces." *Russian Mathematical Surveys*, 22(6): 201–258. MR0223886. 356

Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. (2022). "If Influence Functions are the Answer, Then What is the Question?" In *Advances in Neural Information Processing Systems*. 356

Basu, S. (2000). *Bayesian Robustness and Bayesian Nonparametrics*, 223–240. New York, NY: Springer New York. MR1795218. doi: https://doi.org/10.1007/978-1-4612-1306-2_12. 363

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). "Local posterior robustness with parametric priors: Maximum and average sensitivity." In *Maximum Entropy and Bayesian Methods*, 97–106. Springer. 363

Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons. MR0534323. 356

Carlin, B. and Polson, N. (1991). "An expected utility approach to influence diagnostics." *Journal of the American Statistical Association*, 86(416): 1013–1021. 358, 363

Cook, D. (1977). "Detection of influential observation in linear regression." *Technometrics*, 19(1): 15–18. MR0436478. doi: https://doi.org/10.2307/1268249. 357

Cook, R. (1986). "Assessment of local influence." *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2): 133–155. MR0867994. 363

Diaconis, P. and Freedman, D. (1986). "On the consistency of Bayes estimates." *The Annals of Statistics*, 1–26. MR0829555. doi: https://doi.org/10.1214/aos/1176349830. 356

Fleming, W. (2012). *Functions of Several Variables*. Springer Science & Business Media. MR0422527. 356

Gil-Leyva, M. and Mena, R. (2021). "Stick-breaking processes with exchangeable length variables." *Journal of the American Statistical Association*, 1–14. 357, 358, 359

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). "A higher-order swiss army infinitesimal jackknife." *arXiv preprint arXiv:1907.12116*. 356

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). "A Swiss army infinitesimal jackknife." In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1139–1147. PMLR. 357, 362

Gustafson, P. (1996). "Local sensitivity of posterior expectations." *The Annals of Statistics*, 24(1): 174–195. MR1389886. doi: https://doi.org/10.1214/aos/1033066205. 356, 363

Hampel, F. (1974). "The influence curve and its role in robust estimation." *Journal of the American Statistical Association*, 69(346): 383–393. MR0362657. 356, 357

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley-Interscience; New York. MR0829458. 358

Hattori, S. and Kato, M. (2009). "Approximate subject-deletion influence diagnostics for Inverse Probability of Censoring Weighted (IPCW) method." *Statistics and Probability Letters*, 79(17): 1833–1838. MR2749935. doi: https://doi.org/10.1016/j.spl.2009.05.013. 356

Huber, P. J. (1964). "Robust estimation of a location parameter." *The Annals of Mathematical Statistics*, 35(1): 73–101. URL http://www.jstor.org/stable/2238020 MR0161415. doi: https://doi.org/10.1214/aoms/1177703732. 357

Jacobi, L., Joshi, M., and Zhu, D. (2018). "Automated sensitivity analysis for Bayesian inference via Markov chain Monte Carlo: Applications to Gibbs sampling." *Available at SSRN 2984054*. 356

Johnson, W. and Geisser, S. (1983). "A predictive view of the detection and characterization of influential observations in regression analysis." *Journal of the American Statistical Association*, 78(381): 137–144. MR0696858. 358, 363

Koh, P. and Liang, P. (2017). "Understanding black-box predictions via influence functions." In *International Conference on Machine Learning (ICML)*. 357

Krantz, S. and Parks, H. (2012). *The Implicit Function Theorem: History, Theory, and Applications*. Springer Science & Business Media. MR2977424. doi: https://doi.org/10.1007/978-1-4614-5981-1. 356

Lee, J., James, L., and Choi, S. (2016). "Finite-dimensional BFRY priors and variational Bayesian inference for power law models." *Advances in Neural Information Processing Systems*. 357

Lee, J. and MacEachern, S. (2014). "Inference functions in high dimensional Bayesian inference." *Statistics and its Interface*, 7(4): 477–486. MR3302376. doi: https://doi.org/10.4310/SII.2014.v7.n4.a5. 361

MacEachern, S. and Peruggia, M. (2000). "Importance link function estimation for Markov chain Monte Carlo methods." *Journal of Computational and Graphical Statistics*, 9(1): 99–121. MR1819867. doi: https://doi.org/10.2307/1390615. 363

Maclaurin, D. (2016). "Modeling, Inference and Optimization With Composable Differentiable Procedures." MR3706076. 356

McAuliffe, J., Blei, D., and Jordan, M. I. (2006). "Nonparametric empirical Bayes for the Dirichlet process mixture model." *Statistics and Computing*, 16: 5–14. MR2224185. doi: https://doi.org/10.1007/s11222-006-5196-2. 357, 360, 361

McCulloch, R. (1989). "Local model influence." *Journal of the American Statistical Association*, 84(406): 473–478. 358

McInerney, A. (2013). *First Steps in Differential Geometry*. Springer. MR3098248. doi: https://doi.org/10.1007/978-1-4614-7732-7. 356

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2020). "Monte Carlo gradient estimation in machine learning." *Journal of Machine Learning Research*, 21(132): 1–62. MR4138116. 356

Murray, M. and Rice, J. (1993). *Differential Geometry and Statistics*, volume 48. CRC Press. MR1293124. doi: https://doi.org/10.1007/978-1-4899-3306-5. 356

Ruggeri, F. and Wasserman, L. (1993). "Infinitesimal sensitivity of posterior distributions." *Canadian Journal of Statistics*, 21(2): 195–203. MR1234761. doi: https://doi.org/10.2307/3315811. 356, 363

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons. MR0595165. 357

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer Science & Business Media. MR1351010. doi: https://doi.org/10.1007/978-1-4612-0795-5. 357

Shi, L., Lu, J., Zhao, J., and Chen, G. (2016). "Case deletion diagnostics for GMM estimation." *Computational Statistics & Data Analysis*, 95: 176–191. MR3425947. doi: https://doi.org/10.1016/j.csda.2015.10.003. 356

Stigler, S. (2010). "The changing history of robustness." *The American Statistician*, 64(4): 277–281. MR2758558. doi: https://doi.org/10.1198/tast.2010.10159. 357

Thomas, W. and Cook, D. (1989). "Assessing influence on regression coefficients in generalized linear models." *Biometrika*, 76(4): 741–749. MR1041419. doi: https://doi.org/10.1093/biomet/76.4.741. 356

Thomas, Z., MacEachern, S., and Peruggia, M. (2018). "Reconciling curvature and importance sampling based procedures for summarizing case influence in Bayesian models." *Journal of the American Statistical Association*, 113(524): 1669–1683. MR3902237. doi: https://doi.org/10.1080/01621459.2017.1360777. 358

van der Vaart, A. and Wellner, J. (1996). *Empirical Processes and Weak Convergence*. Springer, New York. MR1385671. doi: https://doi.org/10.1007/978-1-4757-2545-2. 357

von Mises, R. (1947). "On the asymptotic distribution of differentiable statistical functions." *The Annals of Mathematical Statistics*, 18(3): 309–348. MR0022330. doi: https://doi.org/10.1214/aoms/1177730385. 357

Walker, S. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics — Simulation and Computation*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 359

Zeidler, E. (1986). *Nonlinear Functional Analysis and its Applications I: Fixed-point Theorems*. Springer-Verlag. MR0816732. doi: https://doi.org/10.1007/978-1-4612-4838-5. 356, 361