Computational Statistics and Data Analysis 54 (2010) 2172–2186

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Semiparametric Bayes hierarchical models with mean and variance constraints

Mingan Yang [a,*], David B. Dunson [b], Donna Baird [c]

[a] School of Public Health, Saint Louis University, United States

[b] Department of Statistical Science, Duke University, United States

[c] Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, United States

ARTICLE INFO

*Article history:*
Received 17 May 2009
Received in revised form 8 January 2010
Accepted 26 March 2010
Available online 8 April 2010

*Keywords:*
Dirichlet process
Latent variables
Moment constraints
Nonparametric Bayes
Parameter expansion
Random effects

ABSTRACT

In parametric hierarchical models, it is standard practice to place mean and variance constraints on the latent variable distributions for the sake of identifiability and interpretability. Because incorporation of such constraints is challenging in semiparametric models that allow latent variable distributions to be unknown, previous methods either constrain the median or avoid constraints. In this article, we propose a centered stick-breaking process (CSBP), which induces mean and variance constraints on an unknown distribution in a hierarchical model. This is accomplished by viewing an unconstrained stick-breaking process as a parameter-expanded version of a CSBP. An efficient blocked Gibbs sampler is developed for approximate posterior computation. The methods are illustrated through a simulated example and an epidemiologic application.



## 1. Introduction

Hierarchical models that incorporate latent variables or random effects are very widely used. However, a common concern is the appropriateness of parametric assumptions on the latent variable distributions. This has motivated a rich literature on semiparametric approaches, which treat the latent variable distributions as unknown. For example, Bush and MacEachern (1996), Müller and Rosner (1997), Mukhopadhyay and Gelfand (1997), Kleinman and Ibrahim (1998) and Ishwaran and Takahara (2002) use Dirichlet process (DP) (Ferguson, 1973, 1974) mixture models (Escobar, 1994; Escobar and West, 1995) for modeling of unknown random effects distributions.

In many hierarchical models, it is important to constrain the latent variable distributions for the sake of interpretability and identifiability. For example, parametric latent factor models commonly constrain the latent variable distributions to have mean zero and variance one. In the semiparametric Bayes literature, several authors have proposed methods for constraining quantiles of an unknown distribution. Burr and Doss (2005) recently used mixtures of conditional Dirichlet processes (Doss, 1985) to model the random effects distribution in a meta-analysis application. Their formulation allows median constraints, as does the class of mixture models proposed by Kottas and Gelfand (2001). Hanson and Johnson (2002) instead proposed using mixtures of Pólya trees with median constrained to be zero. Dunson et al. (2003) used an alternative strategy for median regression relying on a substitution likelihood (Lavine, 1995). Li et al. (2007) proposed an approach to correct for bias in generalized linear mixed models with a DP prior on the random effects distribution. Their approach relies on post-processing of the samples from an MCMC algorithm.

* Corresponding author. Tel.: +1 919 541 4056; fax: +1 919 541 4311.
*E-mail addresses:* mingany@yahoo.com (M. Yang), dunson@stat.duke.edu (D.B. Dunson), baird@niehs.nih.gov (D. Baird).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.03.025

In contrast to the literature on semiparametric Bayes methods for median or quantile constraints, there has been essentially no work done (to our knowledge) on the problem of modeling of a random distribution subject to mean and variance constraints. A number of authors have proposed approaches for modeling of unknown symmetric densities having mean and mode at zero. For example, Brunner and Lo (1989) and Lavine and Mockus (1995) use DP mixtures of uniform distributions. Hoff (2003) proposed a general approach for defining probability measures in a convex set and applied it to construct measures with mean constraint. Hoff (2000) noted that mean-zero variance-one measures can be characterized using his theory, but difficulties arise in parameterizing the extreme points. Motivated by this problem and by the application to semiparametric latent factor regression, we develop a class of centered stick-breaking processes (CSBP).

In the Bayesian nonparametric literature, stick-breaking formulations of random probability measures have been considered by an increasing number of authors. In pioneering work, Sethuraman (1994) showed that the DP has a stick-breaking representation. In particular, let $G \sim \mathrm{DP}(\alpha G_0)$, where $G$ is a random probability measure, $\alpha$ a precision parameter, and $G_0$ a base probability measure,

$$G = \sum_{h=1}^{\infty} V_h \prod_{l<h}(1 - V_l)\delta_{\theta_h}, \quad V_h \overset{iid}{\sim} \mathrm{beta}(1, \alpha), \ \theta_h \overset{iid}{\sim} G_0, \tag{1}$$

with $\{V_h, h = 1, \ldots, \infty\}$ an infinite sequence of random stick-breaking probabilities, $\{\theta_h, h = 1, \ldots, \infty\}$ an infinite sequence of random atoms, and $\delta_\theta$ a probability measure concentrated at $\theta$. Ishwaran and James (2001) generalized the DP to a broad class of stick-breaking random measures by letting $V_h \sim \mathrm{beta}(a_h, b_h)$ in (1).

It is not straightforward to directly modify the components in (1) to constrain the mean and variance of $G$. Instead, we view the unconstrained stick-breaking random measure as a *parameter-expanded* formulation of a constrained stick-breaking random measure. Parameter expansion was initially proposed as an approach to accelerate convergence of the Gibbs sampler (Liu and Wu, 1999). However, recent work has also used parameter expansion to induce new families of prior distributions (Gelman, 2004, 2006). To our knowledge, this approach has not yet been considered in the context of nonparametric models.

Section 2 motivates the problem through an application to a semiparametric latent factor model, describing a standard Dirichlet process mixture model. Section 3 proposes the centered stick-breaking process (CSBP) and considers its properties. Section 4 develops an efficient parameter expansion blocked Gibbs sampler for posterior computation. Sections 5 and 6 provide simulations and real data analyses, and Section 7 discusses the results.

## 2. Semiparametric latent factor models

### 2.1. Motivation

As motivation, we focus initially on the latent factor model:

$$\begin{aligned} &y_{ij} = \tau_j + \boldsymbol{\lambda}_j'\boldsymbol{\eta}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma_j^2), \\ &\boldsymbol{\eta}_i \sim G, \quad i = 1, \ldots n, \end{aligned} \tag{2}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})' \in \Re^p$ is a vector of continuous measurements on subject $i$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)'$ is a mean vector, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)'$ is a $p \times r$ factor loadings matrix, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ir})'$ is a $r \times 1$ vector of latent factors, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})'$ is an $p \times 1$ vector of idiosyncratic measurement errors, and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is the residual covariance matrix. Model (2) and closely related models have been widely used in recent years due to flexibility in modeling of covariance structures in high-dimensional data (West, 2003). Although we focus initially on the case in which the measured variables are continuous for simplicity, the methods can be directly applied when the variables have mixed categorical and continuous measurement scales, as will be illustrated in Section 6.

Parametric analyses of model (2) typically assume that $G$ is the multivariate normal distribution $\mathrm{N}_r(\mathbf{0}, \mathbf{I})$. These constraints on the mean and variance, made for identifiability and interpretability, result in the marginal model: $\mathbf{y}_i \sim \mathrm{N}_p(\boldsymbol{\tau}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma})$. Further constraints are typically incorporated in the factor loadings matrix, $\boldsymbol{\Lambda}$, to ensure identifiability, as one can replace $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda}\mathbf{P}$, for any orthonormal matrix $\mathbf{P}$, without changing the likelihood (refer to Lopes and West, 2004).

Although the restrictions on the mean and variance of $G$ are clearly justified in order to set the scale and location of the latent variable distribution, the normality assumption is often called into question in applications. This has motivated a rich literature on frequentist semiparametric methods, which avoid a full likelihood specification (Pison et al., 2003; Pison and Van Aelst, 2004).

Our goal is to develop Bayesian semiparametric methods, which treat $G$ as an unknown distribution on $\Re^r$ with mean $\mathbf{0}$ and variance $\mathbf{I}$, with the dimension $r$ treated as known for ease in exposition. For a recent article on accommodating uncertainty in the number of factors in a normal linear factor model, refer to Bhattacharya and Dunson (2009); their approach is easily modified to accommodate factor selection in the semiparametric latent factor models we consider. The Bayesian approach has the distinct advantages of allowing inferences on the latent variable distributions, while also allowing estimation of posterior distributions for the latent variables.

### 2.2. Dirichlet process prior

Ignoring the problem of constraining the mean and variance, one could potentially allow the latent variable distribution, $G$, to be unknown by choosing a Dirichlet process (DP) prior: $G \sim \text{DP}(\alpha G_0)$. Relying on the stick-breaking representation (1), it is then straightforward to show that

$$\begin{aligned}\boldsymbol{\mu}_G = \text{E}(\boldsymbol{\eta}_i \mid G) &= \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)\boldsymbol{\theta}_h,\\ \boldsymbol{\Sigma}_G = \text{V}(\boldsymbol{\eta}_i \mid G) &= \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)(\boldsymbol{\theta}_h-\boldsymbol{\mu}_G)(\boldsymbol{\theta}_h-\boldsymbol{\mu}_G)',\end{aligned} \tag{3}$$

with $(\boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G) \neq (\mathbf{0}, \mathbf{I})$ almost surely. There is a rich literature focused on characterizing the exact distributions of functionals of a Dirichlet process, including the mean and variance (Regazzini et al., 2002; James, 2005, among others).

Conditionally on $G$, the marginal expectation and variance of $\mathbf{y}_i$ integrating over the latent variable distribution are:

$$\begin{aligned}\text{E}(\mathbf{y}_i \mid \boldsymbol{\tau}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, G) &= \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\mu}_G,\\ \text{V}(\mathbf{y}_i \mid \boldsymbol{\tau}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, G) &= \boldsymbol{\Lambda}\boldsymbol{\Sigma}_G\boldsymbol{\Lambda}' + \boldsymbol{\Sigma},\end{aligned} \tag{4}$$

so that $\boldsymbol{\tau}$ and $\boldsymbol{\Lambda}$ no longer have the same marginal interpretation as in the parametric analysis that chooses $G$ as $\text{N}_r(\mathbf{0}, \mathbf{I})$. Ignoring this issue can result in misleading inferences. Note that it is not sufficient to choose $G_0$ to correspond to $\text{N}_r(\mathbf{0}, \mathbf{I})$, as the resulting posterior distribution for $(\boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G)$ need not be concentrated around $(\mathbf{0}, \mathbf{I})$.

## 3. Centered Dirichlet process priors

### 3.1. Formulation

Let $G \sim \mathcal{P}$, where $G$ is a probability measure on $(\Re^r, \mathcal{B})$ and $\mathcal{P}$ is a probability measure on $(\Omega^r_{\mathbf{0},\mathbf{I}}, \mathcal{F})$, with $\Omega^r_{\mathbf{0},\mathbf{I}}$ the space of probability measures on $(\Re^r, \mathcal{B})$ corresponding to distributions with mean $\mathbf{0}$ and variance $\mathbf{I}$. Here, $\mathcal{B}$ and $\mathcal{F}$ are $\sigma$-algebras. Our focus is on the choice of $\mathcal{P}$. In particular, letting $\boldsymbol{\eta}_i \overset{iid}{\sim} G$, with $G \sim \mathcal{P}$, we choose

$$\begin{aligned}&G = \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)\delta_{\boldsymbol{\theta}_h},\\ &\boldsymbol{\theta}_h = \boldsymbol{\Sigma}_{G^*}^{-1/2}(\boldsymbol{\theta}^*_h - \boldsymbol{\mu}_{G^*}), \quad h = 1, \ldots, \infty,\\ &V_h \sim \text{beta}(a_h, b_h), \quad h = 1, \ldots, \infty,\\ &\boldsymbol{\theta}^*_h \overset{iid}{\sim} G_0, \quad h = 1, \ldots, \infty,\end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_{G^*}, \boldsymbol{\Sigma}_{G^*}$ are obtained from expression (3) substituting $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_h, h = 1, \ldots, \infty)$ for $\boldsymbol{\theta} = (\boldsymbol{\theta}_h, h = 1, \ldots, \infty)$. We refer to the choice of $\mathcal{P}$ implied by (5) as a centered stick-breaking process (CSBP). The centered Dirichlet Process (CDP) corresponds to the special case in which $a_h = 1, b_h = \alpha, h = 1, \ldots, \infty$.

**Lemma 1.** *Given specification (5), we have* $E(\boldsymbol{\eta}_i \mid G) = \mathbf{0}$ *and* $V(\boldsymbol{\eta}_i \mid G) = \mathbf{I}$.

The proof of Lemma 1 is straightforward. Note that Lemma 1 holds for any realization from the prior, $\mathcal{P}$, and hence $\mathcal{P}$ has support $\Omega^r_{\mathbf{0},\mathbf{I}}$ as required.

Expression (5) is identical to the class of stick-breaking random measures considered by Ishwaran and James (2001) except for the standardization of the atoms to constrain the random distribution to have mean $\mathbf{0}$ and covariance $\mathbf{I}$ (shown in line 2 of expression (5)).

### 3.2. Alternative formulation

In investigating properties and developing computational algorithms, it is useful to consider an alternative, but equivalent, specification to (5). In particular, note that $\boldsymbol{\eta}_i \sim G, i = 1, \ldots, n, G \sim \mathcal{P}$, with $\mathcal{P}$ a CSBP, is equivalent to the following:

$$\begin{aligned}&\boldsymbol{\eta}_i = \boldsymbol{\Sigma}_{G^*}^{-1/2}\left(\boldsymbol{\eta}^*_i - \boldsymbol{\mu}_{G^*}\right), \quad i = 1, \ldots, n,\\ &\boldsymbol{\eta}^*_i \sim G^*, \quad i = 1, \ldots, n,\\ &G^* = \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)\delta_{\theta^*_h},\end{aligned} \tag{6}$$

where $\boldsymbol{\mu}_{G^*}, \boldsymbol{\Sigma}_{G^*}, \mathbf{V} = (V_h, h = 1, \ldots, \infty)'$ and $\boldsymbol{\theta}^*$ are as defined in Section 2.1. Hence, the latent variables, $\boldsymbol{\eta}_i$, are treated as normalized transformations of latent variables, $\boldsymbol{\eta}^*_i$, having a distribution $G^* \sim \mathcal{P}^*$, with $\mathcal{P}^*$ an unconstrained stick-breaking prior.

Note that we are effectively using a form of parameter expansion, which is conceptually related to the approach proposed by Gelman (2006). Gelman (2006) induces a prior on the variance of a random effect in a parametric model by expressing the random effect as a transformation of a latent variable in an over-parameterized, or parameter-expanded (PX), model. His PX approach results in a prior with appealing properties, while also facilitating efficient posterior computation. In contrast, we induce a prior on a latent variable distribution with mean zero and identity covariance by expressing the latent variable as a transformation of a latent variable in an over-parameterized model that does not constrain the mean and variance. Similarly to Gelman (2006), we can use the PX form in (6) to construct efficient MCMC methods for posterior computation, modifying algorithms developed for unconstrained stick-breaking priors.

### 3.3. Truncations

For unconstrained stick-breaking priors, Ishwaran and James (2001) proposed a blocked Gibbs sampling algorithm for posterior computation, which relies on approximating the infinite-dimensional random measure by truncating the stick-breaking representation. In this section, we adapt their approach for the CSBP.

Let $\mathcal{P}^N$ denote the prior on $G$ resulting from the following specification, used as an approximation or alternative to (5):

$$\begin{aligned} G &= \sum_{h=1}^{N} V_h \prod_{l<h}(1-V_l)\delta_{\boldsymbol{\theta}_h}, \\ \boldsymbol{\theta}_h &= \boldsymbol{\Sigma}_{G_N^*}^{-1/2}(\boldsymbol{\theta}_h^* - \boldsymbol{\mu}_{G_N^*}), \quad h = 1, \ldots, N, \end{aligned} \tag{7}$$

where $V_h \sim \text{beta}(a_h, b_h)$, $h = 1, \ldots, N-1$, $V_N = 1$, $\boldsymbol{\theta}_h^* \overset{iid}{\sim} G_0$, $h = 1, \ldots, N$, and

$$\begin{aligned} \boldsymbol{\mu}_{G_N^*} &= \sum_{h=1}^{N} V_h \prod_{l<h}(1-V_l)\boldsymbol{\theta}_h^* \\ \boldsymbol{\Sigma}_{G_N^*} &= \sum_{h=1}^{N} V_h \prod_{l<h}(1-V_l)(\boldsymbol{\theta}_h^* - \boldsymbol{\mu}_{G_N^*})(\boldsymbol{\theta}_h^* - \boldsymbol{\mu}_{G_N^*})'. \end{aligned}$$

Letting $\boldsymbol{\eta}_i \sim G$, $i = 1, \ldots, n$, with $G \sim \mathcal{P}^N$, we can obtain the following equivalent specification:

$$\begin{aligned} \boldsymbol{\eta}_i &= \boldsymbol{\Sigma}_{G_N^*}^{-1/2}(\boldsymbol{\eta}_i^* - \boldsymbol{\mu}_{G_N^*}), \quad i = 1, \ldots, n, \\ \boldsymbol{\eta}_i^* &\sim G^*, \quad i = 1, \ldots, n, \\ G^* &= \sum_{h=1}^{N} V_h \prod_{l<h}(1-V_l)\delta_{\boldsymbol{\theta}_h^*}. \end{aligned} \tag{8}$$

Letting $\mathcal{P}_N^*$ denote the resulting prior on $G^*$, Theorem 2 of Ishwaran and James (2001) provides a bound on the $\mathcal{L}_1$ distance between $\mathcal{P}_N^*$ and $\mathcal{P}^*$. In the DP special case, this bound $\to 0$ at an exponential rate as $N$ increases, suggesting that a highly accurate approximation can be obtained for moderate sized $N$ in most cases. Because $\boldsymbol{\Sigma}_{G_N^*}^{-1/2}$ and $\boldsymbol{\mu}_{G_N^*}$ are functionals of $G^*$, this result also suggests that $\mathcal{P}_N$ should provide an accurate approximation to $\mathcal{P}$ for moderate $N$ when $V_h \sim \text{beta}(1, \alpha)$, with $\alpha$ small to moderate.

### 3.4. Centered stick-breaking mixtures

Assuming $\boldsymbol{\eta}_i \sim G$, $i = 1, \ldots, n$, with $G \sim \mathcal{P}$ and $\mathcal{P}$ a CSBP, $G$ is almost surely discrete. Hence, the $nr \times 1$ latent variable vectors for the different subjects will not be unique; instead, there will be $k \leq n$ unique values or clusters. The CSBP induces a prior on the set of partitions of the integers $\{1, \ldots, n\}$, which is identical to the prior under the uncentered stick-breaking process. This equivalence is a direct consequence of the fact that the centering modifies the locations of the atoms but not the stick-breaking weights.

Latent variable models that assume discrete distributions for the latent variables are typically referred to as latent class models (LCMs). The CSBP should be widely useful for constructing semiparametric Bayesian latent class models without the need to assume a known number of classes or induce parameter restrictions to identify the classes. In applications in which one wishes to cluster individuals it may be appealing to focus on a LCM.

However, in many settings, it is considered unrealistic to allow ties in the latent variables, as any two individuals are unlikely to be exactly the same. To allow unknown continuous latent trait distributions having zero mean and identity covariance, we propose a *centered stick-breaking mixture* (CSBM). In particular, starting with a parameter-expanded specification, we let

$$\begin{aligned} \boldsymbol{\eta}_i &= \left(\boldsymbol{\Sigma}_{G^*} + \mathbf{I}\right)^{-1/2}(\boldsymbol{\eta}_i^* - \boldsymbol{\mu}_{G^*}), \quad i = 1, \ldots, n, \\ \boldsymbol{\eta}_i^* &\sim \text{N}_r(\boldsymbol{\mu}_i^*, \mathbf{I}), \quad i = 1, \ldots, n, \\ \boldsymbol{\mu}_i^* &\sim G^*, \end{aligned} \tag{9}$$

where $G^*$ is assigned an uncentered stick-breaking prior and the other terms are as described above. Marginalizing out the latent variables $\{\boldsymbol{\eta}_i^*\}$, we obtain:

$$\begin{aligned}
&\boldsymbol{\eta}_i \sim \mathrm{N}_r\left(\boldsymbol{\mu}_i, (\boldsymbol{\Sigma}_{G^*} + \mathbf{I})^{-1}\right), \quad i = 1, \ldots, n,\\
&\boldsymbol{\mu}_i \sim G, \quad i = 1, \ldots, n,\\
&G = \sum_{h=1}^{\infty} V_h \prod_{l<h}(1 - V_l)\delta_{\boldsymbol{\theta}_h},\\
&\boldsymbol{\theta}_h = \left(\boldsymbol{\Sigma}_{G^*} + \mathbf{I}\right)^{-1/2}\left(\boldsymbol{\theta}_h^* - \boldsymbol{\mu}_{G^*}\right), \quad h = 1, \ldots, \infty.
\end{aligned} \tag{10}$$

Note that the implied prior for $G$ is identical to the CSBP of Section 2.1, with the exception that the pre-multiplier is $(\boldsymbol{\Sigma}_{G^*} + \mathbf{I})^{-1/2}$ instead of $\boldsymbol{\Sigma}_{G^*}^{-1/2}$. Hence, we obtain $\mathrm{V}(\boldsymbol{\mu}_i \mid G) = \boldsymbol{\Sigma}_{G^*}(\boldsymbol{\Sigma}_{G^*} + \mathbf{I})^{-1}$, so that

$$\mathrm{E}(\boldsymbol{\eta}_i \mid G) = \mathbf{0} \quad \text{and} \quad \mathrm{V}(\boldsymbol{\eta}_i \mid G) = \boldsymbol{\Sigma}_{G^*}(\boldsymbol{\Sigma}_{G^*} + \mathbf{I})^{-1} + (\boldsymbol{\Sigma}_{G^*} + \mathbf{I})^{-1} = \mathbf{I}.$$

Thus, the CSBM prior for $G$ has support on the space of absolutely continuous densities having mean $\mathbf{0}$ and covariance $\mathbf{I}$.

## 4. Parameter-expanded blocked Gibbs sampler

The latent factor model (2), with a CSBP or a CSBM prior for the latent variable distribution $G$, can be expressed in parameter-expanded form as a Dirichlet process mixture model relying on expression (6) or (9). In either case, computation proceeds under the working model:

$$y_{ij} = \tau_j^* + \lambda_j^* \eta_i^* + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2). \tag{11}$$

We complete a specification of the model with prior distributions for $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_p^*)'$, $\boldsymbol{\Lambda}^* = (\lambda_1^*, \ldots, \lambda_p^*)'$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_p^2)$. For convenience in computation, one can choose a normal prior for $\boldsymbol{\tau}^*$, normal or truncated normal priors for $\boldsymbol{\Lambda}^*$, and inverse-gamma priors for the diagonal elements of $\boldsymbol{\Sigma}$.

For the CSBP prior, we have $\eta_i^* \sim G^*$ with $G^* \sim$ CSBP, and approximate posterior computation can proceed through direct application of the blocked Gibbs sampling algorithm of Ishwaran and James (2001), which relies on truncating the stick-breaking from. After obtaining draws for the approximate parameter-expanded posterior, we transform back to the original hierarchical model using:

$$\tau_j = \tau_j^* + \lambda_j^{*'} \mu_{G_N^*}, \qquad \lambda_j = \lambda_j^{*'} \boldsymbol{\Sigma}_{G_N^*}^{1/2}, \qquad \eta_i = \boldsymbol{\Sigma}_{G_N^*}^{-1/2}(\eta_i - \mu_{G_N^*}). \tag{12}$$

Note that the convergence and mixing rates for the $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$ and $\eta$ parameters tends to be improved over that for the $\boldsymbol{\tau}^*$, $\boldsymbol{\Lambda}^*$, and $\eta^*$.

For the CSBM prior for $G$, a very similar approach can be used, with the transformations from the working to inferential parameterizations shown in (12) modified appropriately. Refer to the Appendix for the specific steps involved in posterior computation.

## 5. Fibroid tumor study

### 5.1. Scientific background and data description

We initially consider an application to data from an NIEHS study of uterine fibroids (Baird et al., 2003), a common reproductive tract tumor, which rarely becomes malignant, but leads to substantial morbidity. In cross-sectional analyses of data from this study, fibroid size was related to increased bleeding (Wegienka et al., 2003). The goal of the current study was to assess whether the current presence and size of uterine fibroids predict the future level of bleeding. In addition, investigators were interested in studying the distribution of bleeding intensity across women adjusting for fibroid size and African American ethnicity. This motivates a nonparametric approach, as the shape of the latent bleeding intensity density is not known in advance; indeed it was of substantial interest to assess whether there is multimodality, suggesting the presence of latent sub-populations and important unmeasured predictors of bleeding.

The uterine fibroid study was conducted by NIEHS in 1996 in collaboration with George Washington University Medical Center. Members aged 35–49 of an urban prepaid health plan in Washington D.C. were selected for the study, out of 1430 participants, 1245 were premenopausal. In the study, information on menstrual, medical and reproductive history as well as any previous fibroid diagnoses and treatment were collected by phone interview. Detailed information on fibroid location and size were collected by ultrasound examination during a clinic visit or from recent medical records if available. After 3–5 years, we attempted to re-contact the premenopausal women, 981 of whom were interviewed and asked about symptoms. If women had had a myomectomy, hysterectomy, or menopause prior to follow-up, they were asked about symptoms prior to those events. Generally, African American women have higher risk of uterine fibroids than other ethnic

groups (Baird et al., 2003). Our interest is in assessing how fibroid size at baseline and African American ethnicity relate to bleeding at the follow-up.

Size of the fibroid is categorized as 0, 1, 2 or 3, corresponding to none, small (<2 cm), medium (between 2 and 4 cm) or large (>4 cm). The following data are available on the intensity of bleeding at follow-up:

- Count data:
  - $Y_1$: number of days during menses of real blood flow.
  - $Y_2$: number of days of spotting.
  - $Y_3$: number of days each month in using more than 8 pads or tampons.
- Binary data:
  - $Y_4$: Is there intermenstrual spotting?
- Ordinal data (1–5 scale):
  - $Y_5$: How often do you have menstrual periods?
    1. Did not have any period.
    2. Too irregular to say.
    3. Less frequently than once a month (>34 days).
    4. About once a month (27–34 days).
    5. More frequently than once a month (<27 days).
  - $Y_6$: How often do you have gushing-type bleeding?
    1. Just once.
    2. During occasional periods.
    3. Most periods.
    4. Every period.
  - $Y_7$: How much did the menstrual bleeding limit social activities?
    1. Not at all.
    2. A little.
    3. Some.
    4. A lot.

Summary statistics for the bleeding symptom data are provided in Table 2. For flexibility in modeling and because most women had values close to 0, we treat the count data as ordinal data for our analysis.

Letting $\eta_i$ denote the latent bleeding intensity score for woman $i$, we used model

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta_i, \quad \delta_i \sim G \tag{13}$$

to relate fibroid size and African American ethnicity to bleeding intensity. The vector $\mathbf{x}_i$ is coded without an intercept and with indicators for $(x_{i1})$ small, $(x_{i2})$ medium and $(x_{i3})$ large fibroids as well as $(x_{i4})$ African American ethnicity. To relate the bleeding score $\eta_i$, to the ordered categorical symptom data, we used a continuation ratio measurement model:

$$P(y_{ij} = c | y_{ij} \geq c, \boldsymbol{\tau}, \lambda, \eta) = \Phi(\tau_{jc} - \lambda_j \eta_i), \quad c = 1, \ldots, C_j, \tag{14}$$

where $C_j$ is the number of categories for symptom type $j$, $\Phi(.)$ is the CDF of the standard normal distribution function, $\lambda_1, \ldots, \lambda_7$ are the loading factors for symptoms $Y_1 - Y_7$.

Note that (14) differs from the measurement model originally proposed in line 1 of model (2). We had initially focused on a linear Gaussian measurement model for simplicity, but now to illustrate the generality of the framework and accommodate the ordered categorical measurement scale of the bleeding symptoms, we modify the measurement model to correspond to a continuation ratio probit model. The same methodology can be applied for essentially any form of the measurement model describing the conditional distribution of the measured variables given the latent variables. For example, one can allow mixed discrete and continuous variables using measurement models proposed in Moustaki and Knott (2000) and Dunson (2000). For categorical measured variables, probit models provide a computationally convenient choice, as underlying variables can be introduced to allow conjugate updating. However, any generalized latent trait model (Moustaki and Knott, 2000) can be used with only a modest additional computational burden. In the second example, we use the logit link for the binary observed variable.

### 5.2. Simulation experiment

We assessed the performance of the approach through a simulation example designed to mimic the fibroid data described in Section 5.1. In this application, we are interested in inference under the latent factor regression model (13) with the same sample size and $\mathbf{x}_i$ values from the real data. For the simulation, we assume that the true parameter values are $\beta = (1, 1, 1, 1)'$, $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_7)' = (1, 1, 1, 1, 1, 1, 1)'$ and the latent variable density $\eta_i$ is the following mixture of four normals:

$$0.15N(-1.92, 0.24) + 0.05N(-0.95, 0.24) + 0.15N(0.024, 0.24) + 0.65N(0.51, 0.24)$$

which has mean 0 and variance 1.

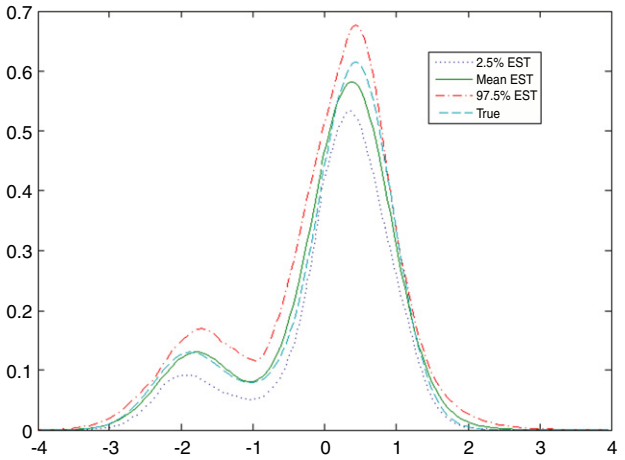


**Fig. 1.** True and estimated latent variable densities in simulation example.

The values of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $n$ are taken directly from the observed data. One of our goals was to assess whether the data contain sufficient information to reliably estimate the latent variable density.

We analyzed the simulation data using a CSBM prior for $G$, applying the algorithm of Section 4. The DP precision parameter, $\alpha$, was treated as unknown using a gamma(1, 1) hyperprior, while $G_0$ was assumed to correspond to N(0, 1). Conditionally conjugate priors were chosen for the remaining parameters as follows:

$$\pi(\boldsymbol{\mu}) = \mathrm{N}_p(\mathbf{0}, 10\mathbf{I}), \qquad \pi(\boldsymbol{\lambda}) = \prod_{j=1}^{p} \mathrm{N}_+(\lambda_j; 0, 10), \qquad \pi(\boldsymbol{\tau}) = \prod_{j=1}^{p} \mathrm{gamma}(\tau_j, 1, 1),$$

where $\mathrm{N}_+(m, v)$ refers to the $\mathrm{N}(m, v)$ distribution truncated to $(0, \infty)$, and $\boldsymbol{\tau} = (\sigma_1^{-2}, \ldots, \sigma_p^{-2})$. A blocked Gibbs sampler was implemented in each case, with the chain run for 100,000 iterations after a 20,000 iteration burn-in, we take every 20th sample resulting in a total of 4000 samples. To assess convergence, we ran several independent chains with widely dispersed starting values; for sensitivity to prior specification, we also tried with varied variances: priors with variance/2, priors with variance $\times 2$, priors with variance $\times 5$. With all these trials, we do not see much differences between the results.

Table 1 presents posterior summaries of the model parameters in each case, while Fig. 1 plots the estimated and true latent variable distributions. From these results, we can see that our approach can produce good results. The estimated latent variable density is very close to the true density, suggesting that the data are informative.

The centered Dirichlet process mixture (CDPM) model results are much more accurate than the results for the DPM model, as expected due to the non-identifiability problem. In general, the closer the latent variable distribution is to the base $G_0$, the better the performance of the DPM model. However, the performance of the DPM degrades in the presence of deviation from $G_0$, while the CDPM results are robust to the shape of the latent variable density.

### 5.3. Analysis of real data

We implemented the analysis as in the simulation example, and again found the results robust to the prior specification. Posterior summaries of the parameters are provided in Table 3. These results suggest a significant increase in bleeding intensity with increasing fibroid size and for African American women compared with other races. For small fibroids compared with no fibroids, the expected change in the latent bleeding intensity score is 0.05 and the 95% credible interval (CI) includes 0. Note that the latent variable regression coefficients have a clear interpretation due to the incorporation of the variance $= 1$ constraint. In particular, the coefficients for the indicators represent the number of standard deviations the mean bleeding intensity score shifts between the categories. Hence, a shift of 0.05 is clearly not a clinically significant change. However, the estimated shift of $\widehat{\beta}_1 + \widehat{\beta}_2 = 0.05 + 0.45 = 0.50$ between no fibroids and size category 2 is significant. The estimated shift between no fibroids and size category 3 is $\widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3 = 1.26$. Hence, fibroid size explains a sizable proportion of the variability in the latent bleeding score.

Interestingly, African American ethnicity is also a significant predictor of bleeding intensity, even adjusting for fibroid size. Although it is known that African Americans have a higher fibroid prevalence, so that it would not be surprising to see more fibroid related bleeding, the occurrence of higher bleeding rates adjusting for fibroid sizes is interesting. It may be that future development of fibroids between the screening examination and the measurement of bleeding symptoms at the follow-up time may explain this difference.

The estimated latent bleeding intensity residual density is plotted in Fig. 2. Interestingly, the density is quite similar to a normal density, though we have demonstrated power to detect non-normality in the simulation example.

**Table 1**
Parameter estimation of DPM & CDPM for simulation.

| Parameter | True value | DPM | | CDPM | |
|---|---|---|---|---|---|
| | | Estimate | 95% CI | Estimate | 95% CI |
| $\beta_1$ | 1.00 | 1.66 | (1.27, 2.09) | 0.88 | (0.68, 1.07) |
| $\beta_2$ | 1.00 | 1.61 | (1.26, 2.00) | 0.85 | (0.68, 1.01) |
| $\beta_3$ | 1.00 | 1.76 | (1.38, 2.24) | 0.93 | (0.74, 1.14) |
| $\beta_4$ | 1.00 | 1.73 | (1.44, 2.11) | 0.92 | (0.78, 1.05) |
| $\lambda_1$ | 1.00 | 0.54 | (0.44, 0.63) | 1.02 | (0.89, 1.15) |
| $\lambda_2$ | 1.00 | 0.56 | (0.46, 0.65) | 1.06 | (0.92, 1.20) |
| $\lambda_3$ | 1.00 | 0.56 | (0.45, 0.67) | 1.06 | (0.91, 1.21) |
| $\lambda_4$ | 1.00 | 0.58 | (0.48, 0.69) | 1.10 | (0.94, 1.29) |
| $\lambda_5$ | 1.00 | 0.62 | (0.48, 0.79) | 1.18 | (0.97, 1.42) |
| $\lambda_6$ | 1.00 | 0.61 | (0.49, 0.72) | 1.14 | (0.98, 1.32) |
| $\lambda_7$ | 1.00 | 0.58 | (0.47, 0.68) | 1.09 | (0.95, 1.24) |

**Table 2**
Empirical means within different fibroid size and ethnicity categories for the seven bleeding symptoms.

| Symptoms | Whites and other | | | | African American | | | |
|---|---|---|---|---|---|---|---|---|
| | Fibroid size | | | | Fibroid size | | | |
| | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| $Y_1$ | 4.05 | 3.80 | 4.92 | 4.81 | 3.94 | 3.88 | 4.63 | 5.88 |
| $Y_2$ | 1.97 | 2.29 | 2.63 | 2.15 | 1.81 | 1.59 | 1.76 | 2.06 |
| $Y_3$ | 0.51 | 0.25 | 0.98 | 1.41 | 0.75 | 1.07 | 1.73 | 2.45 |
| $Y_4$ | 0.92 | 0.87 | 0.87 | 0.98 | 0.94 | 0.97 | 0.91 | 0.90 |
| $Y_5$ | 2.86 | 2.65 | 2.79 | 2.75 | 2.80 | 2.84 | 2.91 | 2.91 |
| $Y_6$ | 1.57 | 1.41 | 2.00 | 2.00 | 1.79 | 2.15 | 2.39 | 2.80 |
| $Y_7$ | 1.27 | 1.27 | 1.40 | 1.75 | 1.29 | 1.54 | 1.75 | 1.93 |
| $n$ | 1453 | 496 | 601 | 383 | 826 | 550 | 1079 | 759 |

**Table 3**
Parameter estimation of DPM & CDPM for real data analysis.

| Parameter | DPM | | CDPM | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| $\beta_1$ | 0.065 | (−0.21, 0.35) | 0.05 | (−0.18, 0.28) |
| $\beta_2$ | 0.53 | (0.29, 0.91) | 0.45 | (0.25, 0.66) |
| $\beta_3$ | 0.91 | (0.60, 1.45) | 0.76 | (0.51, 1.01) |
| $\beta_4$ | 0.54 | (0.34, 0.87) | 0.46 | (0.28, 0.63) |
| $\lambda_1$ | 0.51 | (0.27, 0.71) | 0.60 | (0.43, 0.83) |
| $\lambda_2$ | 0.018 | (0.00, 0.06) | 0.02 | (0.00, 0.04) |
| $\lambda_3$ | 1.17 | (0.73, 1.55) | 1.37 | (1.04, 1.86) |
| $\lambda_4$ | 0.02 | (0.00, 0.07) | 0.02 | (0.00, 0.02) |
| $\lambda_5$ | 0.12 | (0.05, 0.19) | 0.14 | (0.06, 0.14) |
| $\lambda_6$ | 0.96 | (0.61, 1.23) | 1.13 | (0.88, 1.50) |
| $\lambda_7$ | 0.80 | (0.51, 1.01) | 0.93 | (0.73, 1.23) |

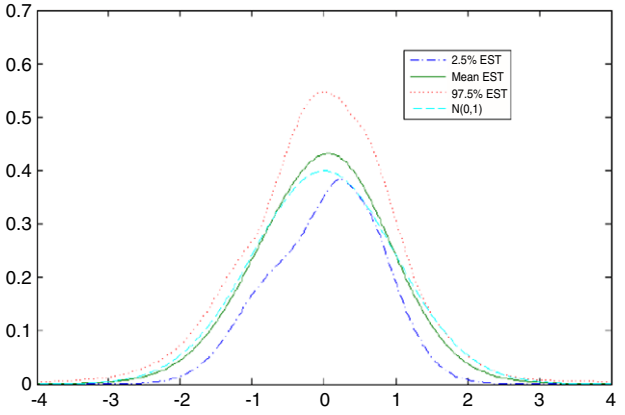


**Fig. 2.** Estimated density of the latent bleeding intensity score in the fibroid data application.

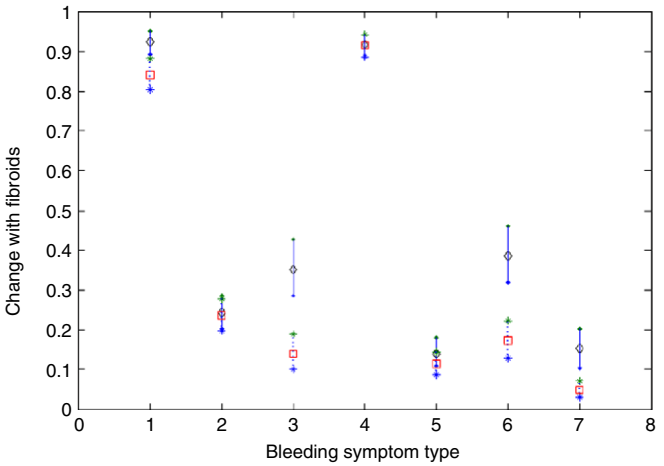


**Fig. 3.** Comparison of bleeding symptoms for black women with large fibroid size (solid line, diamond: estimated mean) vs. no fibroids (dashed line, square:estimated mean).

It is important to assess which symptoms provide the most information about the latent bleeding intensity score for a woman and hence are most sensitive to fibroid size. With this goal in mind, we plot the predicted mean symptom score in different fibroid size categories for African American women in Fig. 3. The plot for white women and other ethnicities shows a very similar pattern. For symptoms 2, 4 and 5, there are essentially no differences across the fibroid size categories and the factor loading parameters are low, suggesting that the bleeding intensity score has low correlation with these symptoms. Symptoms 2 and 4 relate to spotting, while symptom 5 relates to frequency of menstrual periods. In contrast, for symptom 1, there is a moderate shift across fibroid size categories, while for symptoms 3, 6 and 7, the shift is large, with non-overlapping 95% predictive intervals. These findings are quite plausible biologically, as symptoms 3 and 6 relate to frequency of severe bleeding, while symptom 7 measures bleeding that is sufficient to limit activities.

## 6. DDT and premature delivery study

### 6.1. Scientific background and data description

As a second example, we analyzed data from an epidemiology study investigating the relationship between DDT exposure and premature delivery (Longnecker et al., 2001). The study measured concentrations of $p, p'$-DDT and $p, p'$-DDE (1, 1-dichloro-2, 2-bis($p$-chlorophenyl) ethylene), a persistent metabolite of DDT, in 2613 third trimester maternal serum samples from the US Collaborative Perinatal Project (CPP). Although Longnecker et al. (2001) focused their analysis on serum concentration of $p, p'$-DDE ($x_{i1}$), data were also collected on lipid-adjusted $p, p'$-DDE ($x_{i2}$), serum concentration of $p, p'$-DDT ($x_{i3}$), and lipid-adjusted $p, p'$-DDT ($x_{i4}$). The $x_{i1}, x_{i2}, x_{i3}, x_{i4}$ variables are moderately to highly correlated, and it is not clear which should be used in assessing the relationship between DDT exposure and premature delivery.

For this reason, it is natural to consider a latent variable model of the form:

$$\begin{aligned} &\text{logit}\{\Pr(y_i = 1 \mid \eta_i, \mathbf{z}_i)\} = \mathbf{z}_i'\boldsymbol{\tau}_y + \lambda_y \eta_i \\ &x_{ij} = \tau_{x,j} + \lambda_{x,j}\eta_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2),\ j = 1, 2, 3, 4 \\ &\eta_i \sim G, \end{aligned} \tag{15}$$

where $y_i = 1$ if woman $i$ experiences a preterm birth and $y_i = 0$ otherwise, $\mathbf{z}_i$ is a vector of predictors with $\mathbf{z}_i = (1, \text{age}_i, \text{black}_i)'$, $\text{age}_i$ is standardized age for woman $i$, $\text{black}_i$ is an indicator of African American ethnicity, $\eta_i$ is a latent variable summarizing the level of DDT exposure, $\lambda_y$ is a coefficient characterizing the effect of $\eta_i$ on the log-odds of preterm birth, and $G$ is the distribution of $\eta_i$. Mean and variance constraints on $G$ are necessary to fix the scale of the latent variable, which is needed for identifiability and interpretability of the $\lambda_y$ coefficient. If $G$ has mean zero and variance one, then $\lambda_y$ is interpretable as the increase in log-odds of preterm birth attributable to a one standard deviation increase in exposure, and $\mathbf{z}'\boldsymbol{\tau}_y$ is the baseline log-odds of preterm birth among individuals with an average level of exposure having predictors $\mathbf{z}$.

To provide further motivation, DDT is a pesticide, which is currently banned in the United States and numerous countries. However, DDT is still in routine use in the developing world as an anti-malarial agent due to its effectiveness against mosquitoes. Decisions to continue the use of DDT must weigh this public health benefit against the increasing evidence of adverse human health effects.

Our interest focuses on investigating the impact on preterm delivery, which is a major public health concern worldwide, as babies born premature are at substantially increased risk of infant mortality as well as short and long term morbidity. In

**Table 4**
Summaries of the posterior distributions under the normal latent factor model and the centered Dirichlet process mixture latent factor model for data simulated to mimic the DDT and premature delivery data.

| Parameter | True value | Normal | | CDPM | |
|---|---|---|---|---|---|
| | | Estimate | 95% CI | Estimate | 95% CI |
| $\tau_{y,1}$ | −2.2 | −2.09 | (−2.43, −1.85) | −2.19 | (−2.71, −2.03) |
| $\tau_{y,2}$ | 0.10 | 0.11 | (0.08, 0.11) | 0.10 | (0.09, 0.12) |
| $\tau_{y,3}$ | 1.00 | 1.12 | (0.98, 1.24) | 1.03 | (0.95, 1.20) |
| $\lambda_{y,1}$ | 0.20 | 0.15 | (0.09, 0.21) | 0.20 | (0.13, 0.26) |
| $\lambda_{x,1}$ | 1.00 | 0.99 | (0.94, 1.04) | 1.01 | (0.97, 1.06) |
| $\lambda_{x,2}$ | 2.00 | 2.01 | (1.94, 2.08) | 1.99 | (1.92, 2.06) |
| $\lambda_{x,3}$ | 3.00 | 3.01 | (2.92, 3.11) | 3.00 | (2.93, 3.10) |
| $\lambda_{x,4}$ | 4.00 | 4.01 | (3.88, 4.13) | 3.99 | (3.89, 4.11) |
| $\sigma_{x,1}$ | 1.00 | 1.01 | (0.98, 1.04) | 0.98 | (0.95, 1.01) |
| $\sigma_{x,2}$ | 1.00 | 0.98 | (0.96, 0.99) | 1.01 | (0.99, 1.04) |
| $\sigma_{x,3}$ | 1.00 | 0.96 | (0.92, 1.02) | 0.97 | (0.93, 1.01) |
| $\sigma_{x,4}$ | 1.00 | 0.99 | (0.94, 1.07) | 1.02 | (0.98, 1.09) |
| $\tau_{x,1}$ | 0.00 | 0.00 | (−0.06, 0.06) | 0.00 | (−0.06, 0.05) |
| $\tau_{x,2}$ | 0.00 | −0.01 | (−0.08, 0.08) | −0.04 | (−0.13, 0.04) |
| $\tau_{x,3}$ | 0.00 | −0.04 | (−0.17, 0.08) | −0.05 | (−0.16, 0.08) |
| $\tau_{x,4}$ | 0.00 | −0.01 | (−0.17, 0.14) | −0.06 | (−0.16, 0.05) |

addition, extended hospital stays and care associated with prematurity is extremely expensive. In assessing public health impact and conveying this impact to clinicians and policy makers, it is important to have a simple and interpretable summary of association between DDT exposure and risk of prematurity.

It is not possible to accurately measure the level of external exposure to DDT for different women in a large, prospective epidemiologic study, such as the US Collaborative Perinatal Project (CPP). Instead, Longnecker et al. (2001) relied on assaying maternal serum samples collected in the third trimester of pregnancy. Blood levels of $p, p'$-DDT and the persistent metabolite, $p, p'$-DDE, provide surrogates of the external exposure to DDT, with the health impact of the external exposure being the primary interest from a public health perspective. Because $p, p'$-DDT and $p, p'$-DDE are lipid-soluble, there has been some controversy in the literature over whether one should basis the analysis on serum levels in $\mu$g/L or on lipid-adjusted values. As there are valid arguments on both sides, our preference is to include both types of measurements through the latent variable model (15).

An additional argument in favor of the latent variable model (15) is interpretability. As an overall measure of the association between $p, p'$-DDE and preterm birth, one can present an estimated logistic regression coefficient, say $\widehat{\theta}$. The coefficient $\theta$ is interpretable as the increase in the log-odds of preterm birth attributable to a one $\mu$g/L increase in serum level of $p, p'$-DDE. Unfortunately, the value of $\widehat{\theta}$ provides no insight into public health impact in the absence of careful thinking about the population distribution of $p, p'$-DDE. In contrast, the coefficient $\lambda_y$ in model (15) is interpretable as the increase in log-odds of exposure due to a one standard deviation increase in the summary measure of exposure to DDT.

### 6.2. Simulation experiment

For comparison, as in Section 5.2, we repeated the analysis under the assumption that the latent variable had a $N(0, 1)$ density instead of an unknown density. Fig. 5 presents the estimated and true latent variable densities in comparison with the standard normal. It is clear that the data contain substantial information about the latent variable density in that we obtained a very accurate estimate of the density in this simulation, which was chosen to have the same sample size and data structure as the DDT and preterm birth application presented in Section 6. Table 4 provides posterior summaries of the parameters in the CDPM and normal latent variable analysis. Most of the parameter estimates seemed robust to misspecification of the latent variable density in that the normal analysis produced posterior means close to the true values for the factor loadings, residual variances, and intercepts in the predictor model. However, as shown in Fig. 6 there was evidence of bias in estimation of the exposure effect coefficient, $\lambda_y$, which would be the primary interest in most applications. For the CDPM analysis, the posterior density of $\lambda_y$ was centered on the true value of 0.2, while in the normal analysis the true value was in the right tail of the posterior. Although the results are not presented here, the uncentered DPM produced poor results in this and other simulation cases we considered.

### 6.3. Analysis and results

The Longnecker et al. (2001) study had serum measurements for 2613 women selected from the CPP. Our analysis focused on 2380 women, obtained after removing 233 women who had missing covariate values. Data for a single pregnancy was available for each woman, with $361/2380 = 15\%$ of the pregnancies resulting in a preterm birth. The average age for women in the sample was 24, with an interquartile range of 20–28.

Fig. 4 shows histograms for $x_{i1} - x_{i4}$, while Table 6 shows the correlations in the different DDT surrogates. It is clear from Fig. 4 that the surrogates are not normally distributed and there is a tendency towards a large right skew in the distributions.

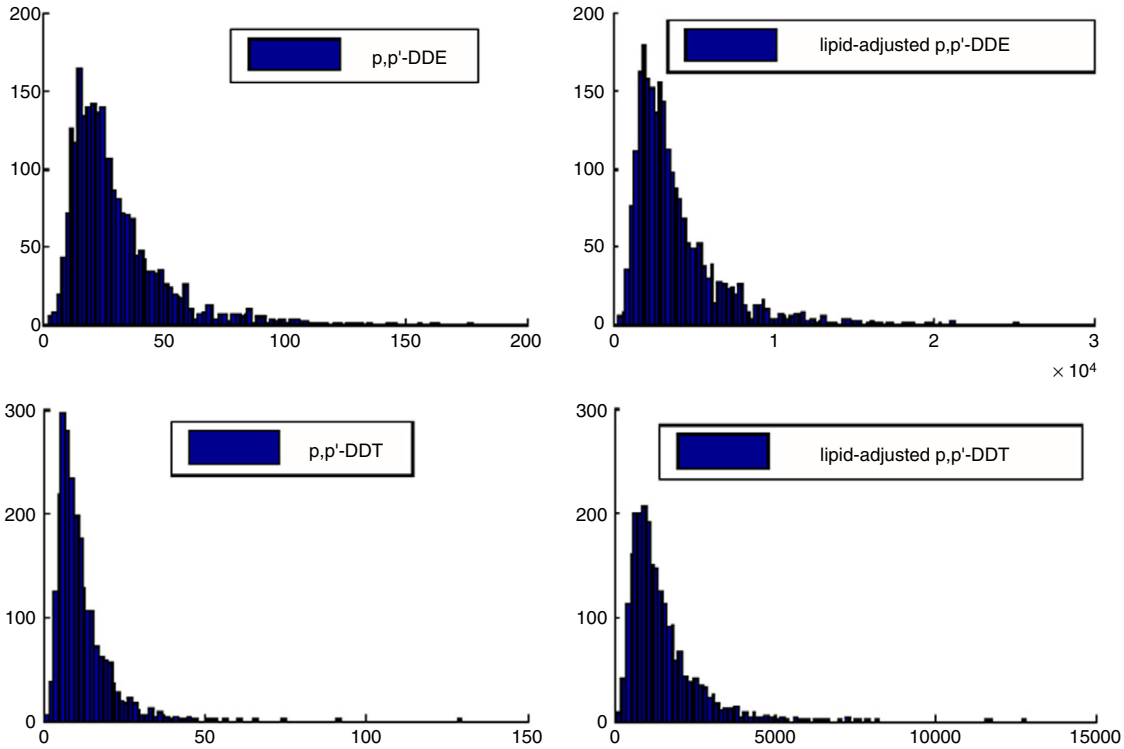


**Fig. 4.** Histograms of the DDT data.

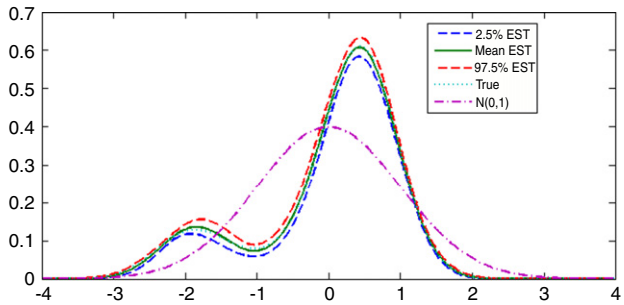


**Fig. 5.** True and estimated latent variable densities for the simulation example. The dotted line is the posterior mean estimate, the solid line is the true density, and the dashed lines are 95% pointwise credible intervals.

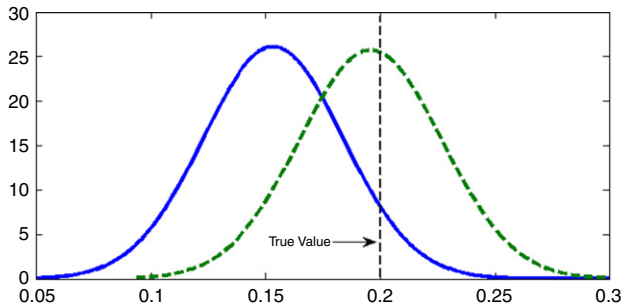


**Fig. 6.** Estimated posterior densities for $\lambda_y$ in the simulation example. The dotted line is the density in the normal analysis, and the solid line is the density in the CDPM analysis. The vertical line shows the true value.

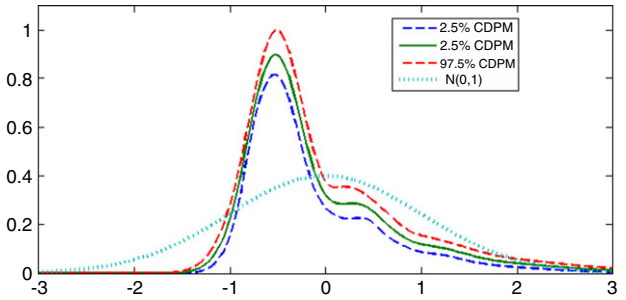


**Fig. 7.** Estimated density of the latent DDT exposure variable in the DDT and premature birth application. The solid line is the posterior mean, the dashed lines are 95% pointwise credible intervals, and the dotted line is the standard normal density.

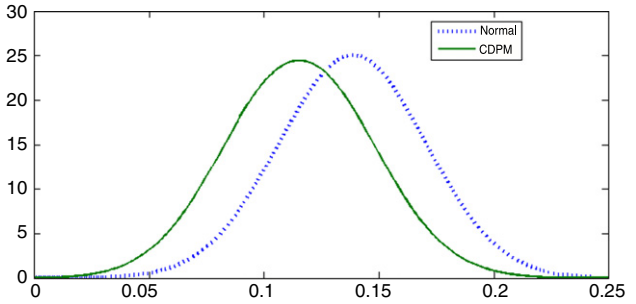


**Fig. 8.** Estimated posterior densities for $\lambda_y$ in the DDT and preterm birth application. The solid line is the estimate in the CDPM analysis, while the dotted line is the estimate for the normal latent variable model.

**Table 5**
Posterior summaries under the normal latent factor model and the centered DPM latent factor model for the DDT and preterm birth data.

| Parameter | Normal | | CDPM | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| $\tau_{y,1}$ | −1.11 | (−1.39, −0.84) | −1.14 | (−1.42, −0.87) |
| $\tau_{y,2}$ | 0.00 | (−0.01, 0.00) | 0.00 | (−0.01, 0.01) |
| $\tau_{y,3}$ | 0.27 | (0.13, 0.41) | 0.26 | (0.12, 0.40) |
| $\lambda_y$ | 0.14 | (0.07, 0.20) | 0.11 | (0.05, 0.18) |
| $\lambda_{x,1}$ | 0.95 | (0.92, 0.98) | 0.71 | (0.65, 0.79) |
| $\lambda_{x,2}$ | 0.95 | (0.92, 0.99) | 0.74 | (0.67, 0.81) |
| $\lambda_{x,3}$ | 0.72 | (0.69, 0.76) | 0.96 | (0.89, 1.05) |
| $\lambda_{x,4}$ | 0.74 | (0.70, 0.78) | 0.98 | (0.90, 1.07) |
| $\sigma_{x,1}$ | 0.33 | (0.30, 0.34) | 0.70 | (0.68, 0.73) |
| $\sigma_{x,2}$ | 0.30 | (0.28, 0.32) | 0.68 | (0.66, 0.70) |
| $\sigma_{x,3}$ | 0.70 | (0.68, 0.72) | 0.34 | (0.32, 0.36) |
| $\sigma_{x,4}$ | 0.69 | (0.66, 0.71) | 0.28 | (0.26, 0.31) |
| $\tau_{x,1}$ | 0.00 | (−0.04, 0.04) | 0.00 | (−0.04, 0.04) |
| $\tau_{x,2}$ | 0.00 | (−0.04, 0.04) | 0.00 | (−0.04, 0.04) |
| $\tau_{x,3}$ | 0.00 | (−0.04, 0.05) | 0.00 | (−0.03, 0.05) |
| $\tau_{x,4}$ | 0.00 | (−0.04, 0.05) | 0.00 | (−0.03, 0.05) |

**Table 6**
DDT surrogates correlation table.

| | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ | $x_{i4}$ |
|---|---|---|---|---|
| $x_{i1}$ | 1.00 | 0.91 | 0.69 | 0.61 |
| $x_{i2}$ | 0.91 | 1.00 | 0.61 | 0.70 |
| $x_{i3}$ | 0.69 | 0.61 | 1.00 | 0.91 |
| $x_{i4}$ | 0.61 | 0.70 | 0.91 | 1.00 |

In addition, the surrogates are moderately to highly correlated, providing support for the incorporation of a single summary latent variable $\eta_i$ in model (15). The women in the right tail of the distribution for one surrogate tended to be in the right tail for the other surrogates.

To minimize the influence of subjective hyperparameter selection, which was necessary to avoid the possibility of an improper posterior, we standardized the surrogates prior to analysis. A repeat analysis without standardization did not change the results. We implemented the analysis for the CDPM and normal latent variable models using the same prior and computational implementation as that described in Section 5. Table 5 presents posterior summaries for each of the parameters. In both the normal and CDPM analysis, there was no association between preterm birth and age, likely due to the fact that women in the sample were primarily young. In addition, the analyses estimated a 0.26–0.27 increase in the log-odds of preterm birth among African Americans. It is well known in the literature that African American's tend to have consistently higher rates of preterm birth.

The results of the parametric and CDPM analysis were much less consistent for the parameters characterizing the relationship between the latent variable, $\eta_i$, and the measured variables, $y_i$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$. As shown in Fig. 8, there is a non-negligible shift in the posterior distribution for $\lambda_y$, with the normal analysis producing an estimate of $\widehat{\lambda}_y = 0.14$ (95% CI $= [0.07, 0.20]$) and the CDPM analysis resulting in $\widehat{\lambda}_y = 0.11$ (95% CI $= [0.05, 0.18]$). In both cases, there is a clear increase in the risk of preterm birth as level of exposure increases, and 95% credible intervals have similar widths, suggesting that the nonparametric approach does not in general result in a reduction in efficiency. In fact, we have observed in other cases that the credible intervals are commonly narrower for the nonparametric analysis in cases in which the parametric model provides a poor approximation.

Fig. 7 shows the estimated latent variable density along with pointwise 95% credible intervals. Clearly, there are substantial departures from normality, with the latent variable density having a large right skew consistent with the surrogate distributions shown in Fig. 4 and with exploratory analyses of the data. It appears that many individuals have a low level of exposure to DDT, but there are a few individuals with very high levels of exposure. This is certainly a common scenario in epidemiology. We note that log transforming the surrogate data results in a latent variable density that is closer to normal, but with clear departures. Post hoc choices of transformations can result in an underestimation of uncertainty and biased inferences, and we find it unappealing to obscure the large differences in exposure level through a transformation that reduces the right tail. In addition to obtaining a very different estimate of the latent variable density under our semiparametric Bayes analysis, we also note that the variance component estimates are substantially different than those obtained in the normal latent variable analysis. This is apparent in examining Table 3.

We repeated the convergence assessments and sensitivity analyses reported in Section 5, and obtained very similar results. In particular, rates of convergence and mixing were quite good based on examination of trace plots and standard diagnostics, and inferences were not sensitive to local changes in the hyperparameter choice.

## 7. Discussion

In this article, we propose a centered stick-breaking process that constrains the mean and variance for latent variable distributions in a hierarchical model. We accomplish this method with the use of parameter expansion, that is, by viewing the uncentered stick-breaking process as a parameter-expanded version of the centered stick-breaking process. This is a simple but useful idea that has a clear impact on the results, reducing bias and improving interpretability over uncentered methods. An appealing feature is that approximate posterior computation can proceed as in the uncentered case with a very simple post-processing algorithm applied to the MCMC draws. This bypasses the need to implement computation directly for the constrained model, which is very challenging.

## Acknowledgements

The authors thank Lianming Wang and Shannon Laughlin for their critical reading of the manuscript.

## Appendix

Here, we present the conditional posterior distribution used in implementing the Gibbs sampler for the following models:

$$y_{ij} = \tau_j + \lambda_j \eta_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2)$$
$$\eta_i = \mathbf{x}_i\boldsymbol{\beta} + \mu_i + \delta_i, \quad \delta_i \sim N(0, 1), \ \mu_i \sim G.$$

- With prior $\tau_j \sim N(\mu_\tau, \sigma_\tau^2)$, the posterior is $N(\mu_\tau^*, \sigma_\tau^{*2})$.

$$\sigma_\tau^{*-2} = \sigma_\tau^{-2} + n_j \sigma_j^{-2}$$
$$\mu_\tau^* = \sigma_\tau^{*2} \left\{ \mu_\tau \sigma_\tau^{-2} + \sum_{i=1}^{n} \sigma_\tau^{-2} (y_{ij} - \lambda_j \eta_i) \right\}.$$

- For prior $\lambda_j \sim N_+(\mu_\lambda, \sigma_\lambda^2)$, the posterior is truncated positive normal with parameters

$$\sigma_\lambda^{*-2} = \sigma_\lambda^{-2} + \sum_{i=1}^{n_j} \eta_i^2/\sigma_j^2$$

$$\mu_\lambda^* = \sigma_\lambda^{*2}\left\{\mu_\lambda\sigma_\lambda^{-2} + \sum_{i=1}^{n_j}\eta_i\sigma_j^{-2}(y_{ij}-\tau_j)\right\}.$$

- For prior $\sigma_j^2 \sim \mathrm{IG}(d_1, d_2)$, the posterior is also inverse-gamma distribution with parameters

$$d_1^* = d_1 + n_j/2$$

$$d_2^* = d_2 + \sum_{i=1}^{n_j}(y_{ij}-\tau_j-\lambda_j\eta_i)^2/2.$$

- Sample $\eta_i$ from the posterior normal distribution with parameters

$$\sigma_{\eta_i}^{*-2} = 1.0 + \sum_{j=1}^{p}\lambda_j^2/\sigma_j^2$$

$$\mu_{\eta_i}^* = \sigma_{\eta_i}^{*2}\left\{\mathbf{x}_i'\boldsymbol{\beta} + \mu_i + \sum_{j=1}^{p}\lambda_j\sigma_j^{-2}(y_{ij}-\tau_j)\right\}.$$

- With prior of $N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0})$, sample $\boldsymbol{\beta}$ from the posterior normal distribution with parameters

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{*-1} = \mathbf{x}\mathbf{x}' + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}^{-1}$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}}^* = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^*\left\{\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\mathbf{x}_i(\eta_i-\mu_i)\right\}.$$

- The unique values of $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_N^*)$, which corresponds to $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, $n \geq N$. With prior $N(g1, g2)$, we sample the posterior $\mu_i^*$ from normal distribution with parameters

$$g_2^* = g_2^{-1} + \sum_{i:\mu_i=\mu_j^*}\mu_j^{*2}$$

$$g_1^* = g_2^*\left\{g_1/g_2 + \sum_{i:\mu_i=\mu_j^*}(\eta_i - \mathbf{x}_i\boldsymbol{\beta})\right\}.$$

- Allocate individuals to latent classes for $\mu_i^*$ by sampling $K_i^*$ from

$$K_i^* \sim \sum_{k=1}^{N} p_{k,i}\delta_k(.)$$

$$p_{k,i} \propto p_k N(\eta_i|\mathbf{x}_i\boldsymbol{\beta} + \mu_i, 1.0).$$

- Sample $V_j^*$ from posterior Beta distribution $V_j^* \sim \mathrm{Beta}(1.0+q_j, c+\sum_{l=j+1}^{N} q_l)$, $j = 1, \ldots, N-1$, where $q_j = \sum_{i=1}^{n} I(\mu_i = \mu_j^*)$. Set $p_1 = V_1^*, \ldots, p_j = (1-V_1^*)(1-V_2^*)\cdots(1-V_{j-1}^*)V_j^*$ for $j = 2, \ldots, N-1$.
- With prior gamma$(g, h)$, sample $c$ from posterior gamma with parameters

$$g^* = g + N$$

$$h^* = h - \sum_{i=1}^{N}\log(1-V_i^*).$$

## References


Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D., Schectman, J.M., 2003. High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. American Journal of Obstetrics and Gynecology 188, 100–107.

Bhattacharya, A., Dunson, D.B., 2009. Sparse Bayesian infinite factor models. Discussion Paper. 23. Department of Statistical Science, Duke University, Durham, NC.

Brunner, L., Lo, A., 1989. Bayes methods for a symmetric unimodal density and its mode. The Annals of Statistics 17, 1550–1566.

Burr, D., Doss, H., 2005. A Bayesian semiparametric model for random-effects meta-analysis. Journal of the American Statistical Association 100, 242–251.

Bush, C.A., MacEachern, S.N., 1996. A semiparametric Bayesian model for randomized block designs. Biometrika 83, 275–285.

Doss, H., 1985. Bayesian nonparametric estimation of the median: part 1: computation of the estimates. The Annals of Statistics 13, 1432–1444.
Dunson, D.B., 2000. Bayesian latent variable models for clustered mixed outcomes. Journal of the Royal Statistical Society B 62, 355–366.
Dunson, D.B., Watson, M., Taylor, J.A., 2003. Bayesian latent variable models for median regression on multiple outcomes. Biometrics 59, 296–304.
Escobar, M., 1994. Estimating normal means with a Dirichlet process prior. Journal of the American Statistical Association 89, 268–277.
Escobar, M., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.
Ferguson, T.S., 1974. Prior distributions on spaces of probability measures. The Annals of Statistics 2, 615–629.
Gelman, A., 2004. Parameterization and Bayesian modeling. Journal of the American Statistical Association 99, 537–545.
Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 515–534.
Hanson, T., Johnson, W.O., 2002. Modeling regression error with a mixture of polya trees. Journal of the American Statistical Association 97, 1020–1033.
Hoff, P.D., 2000. Constrained nonparametric estimation via mixtures. Doctoral Dissertation. Department of Statistics, University of Wisconsin.
Hoff, P.D., 2003. Nonparametric estimation of convex models via mixtures. The Annals of Statistics 31, 174–200.
Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96, 161–173.
Ishwaran, H., Takahara, G., 2002. Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. Journal of the American Statistical Association 97, 1154–1166.
James, L.F., 2005. Functionals of Dirichlet processes, the Cifarelli–Regazzini identity and beta–gamma processes. The Annals of Statistics 33, 647–660.
Kleinman, K.P., Ibrahim, J.G., 1998. A semiparametric Bayesian approach to the random effects model. Biometrics 54, 921–938.
Kottas, A., Gelfand, A.E., 2001. Bayesian semiparametric median regression modeling. Journal of the American Statistical Association 96, 1458–1468.
Lavine, M., 1995. On an approximate likelihood for quantiles. Biometrika 82, 220–222.
Lavine, M., Mockus, A., 1995. A nonparametric Bayes method for isotonic regression. Journal of Statistical Planning and Inference 46, 235–248.
Li, Y., Muller, P., Lin, X, Bias-Corrected Inference in Semiparametric Bayesian Mixed Models Technical Report. 2007.
Liu, J.S., Wu, Y.N., 1999. Parameter expansion for data augmentation. Journal of the American Statistical Association 94, 1264–1274.
Lopes, H.F., West, M., 2004. Bayesian model assessment in factor analysis. Statistica Sinica 14, 41–67.
Longnecker, M.P., Klebanoff, M.A., Zhou, H.B., Brock, J.W., 2001. Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. Lancet 358, 110–114.
Moustaki, I., Knott,, 2000. Generalized latent trait models. Psychometrika 65, 391–441.
Mukhopadhyay, S., Gelfand, A.E., 1997. Dirichlet process mixed generalized linear models. Journal of the American Statistical Association 92, 633–639.
Müller, P., Rosner, G., 1997. A Bayesian population model with hierarchical mixture priors applied to blood count data. Journal of the American Statistical Association 92, 1279–1292.
Pison, G., Rousseeuw, P.J., Flizmoser, P., Croux, C., 2003. Robust factor analysis. Journal of Multivariate Analysis 84, 145–172.
Pison, G., Van Aelst, S., 2004. Diagnostic plots for robust multivariate methods. Journal of Computational and Graphical Statistics 13, 310–329.
Regazzini, E., Guglielmi, A., Di Nunno, G., 2002. Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. The Annals of Statistics 30, 1376–1411.
Sethuraman, J., 1994. A constructive definition of the Dirichlet process. Statistica Sinica 4, 639–650.
Wegienka, G., Baird, D., Hertz-Picciotto, I., Harlow, S., Steege, J., Hill, M., Schectman, M., Hartmann, K., 2003. Self-reported heavy bleeding associated with uterine leiomyomata. American Journal of Obstetrics and Gynecology 3, 431–437.
West, M., 2003. Bayesian factor regression models in the large $p$, small $n$ paradigm. Bayesian Statistics 7, 723–732.