

Semiparametric Bayes hierarchical models with mean and variance constraints

David B. Dunson¹, Mingan Yang¹ and Donna Baird²

¹*Biostatistics Branch, MD A3-03*

²*Epidemiology Branch*

National Institute of Environmental Health Sciences

Research Triangle Park, NC 27709

dunson1@niehs.nih.gov

In parametric hierarchical models, it is standard practice to place mean and variance constraints on the latent variable distributions for 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 posterior computation. The methods are illustrated through a simulated example and an epidemiologic application.

Key Words: Dirichlet process; Latent variables; Moment constraints; Nonparametric Bayes; Parameter expansion; Random effects.

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 sake of interpretability and identifiability. For example, parametric latent factor models commonly constrain the latent variables 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, Watson and Taylor (2003) used an alternative strategy for median regression relying on a substitution likelihood (Lavine, 1996). 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.

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 use DP mixtures of uniforms. Hoff (2003) proposed a general approach for defining probability measures in a convex set, applying this approach 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, letting $G \sim DP(\alpha G_0)$, with G a random probability measure, α a precision parameter, and G_0 a base probability measure,

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

with $\{V_h, h = 1, \dots, \infty\}$ an infinite sequence of random stick-breaking probabilities, $\{\theta_h, h = 1, \dots, \infty\}$ an infinite sequence of random atoms, and δ_{θ} a probability measure concentrated at θ . Ishwaran and James (2001) generalized the DP to a broad class of stick-breaking random measures by letting $V_h \sim \text{beta}(a_h, b_h)$ 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 properties. Section 4 develops an efficient parameter expansion blocked Gibbs sampling algorithm for posterior computation. Section 5 applies the approach to simulation data examples, Section 6 considers an epidemiologic application, and Section 7 discusses the results.

2. Semiparametric Latent Factor Models

2.1 Motivation

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

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

where $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})'$ is a vector of measurements on subject i , $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p)'$ is a mean vector, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_p)'$ is a $p \times r$ factor loadings matrix, $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{ir})'$ is a $r \times 1$ vector of latent factors, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{ip})'$ are idiosyncratic measurement errors, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ 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).

Parametric analyses of model (2) typically assume that G corresponds to $N_r(\mathbf{0}, \mathbf{I})$, the multivariate normal distribution with zero mean and identity covariance. These constraints on the mean and variance, made for identifiability and interpretability, result in the marginal model: $\mathbf{y}_i \sim N_p(\boldsymbol{\tau}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma})$. 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 \mathfrak{R}^r with mean $\mathbf{0}$ and variance \mathbf{I} , with the dimension r treated as known for ease in exposition. 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: $G \sim DP(\alpha G_0)$. Relying on the stick-breaking representation (1), it is then straightforward to show that

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

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, Guglielmi and Di Nunno, 2002; James, 2005; among others).

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

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

so that τ and $\boldsymbol{\Lambda}$ no longer have the same marginal interpretation as in the parametric analysis that chooses G as $N_r(\mathbf{0}, \mathbf{I})$. Ignoring this issue can result in misleading inferences.

Note that it is not sufficient to choose G_0 to correspond to $N_r(\mathbf{0}, \mathbf{I})$, 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 $(\mathbb{R}^r, \mathcal{B})$ and \mathcal{P} is a probability measure on $(\Omega_{\mathbf{0}, \mathbf{I}}^r, \mathcal{F})$, with $\Omega_{\mathbf{0}, \mathbf{I}}^r$ the space of probability measures on $(\mathbb{R}^r, \mathcal{B})$ corresponding to distributions with mean $\mathbf{0}$ and variance \mathbf{I} . Here, \mathcal{B} and \mathcal{F} are σ -algebras. Our focus is on the choice of \mathcal{P} . In particular, letting $\boldsymbol{\eta}_i \stackrel{iid}{\sim} G$, with $G \sim \mathcal{P}$, we choose

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

where $\boldsymbol{\mu}_{G^*}, \boldsymbol{\Sigma}_{G^*}$ are obtained from expression (3) substituting $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_h^*, h = 1, \dots, \infty)$ for $\boldsymbol{\theta} = (\boldsymbol{\theta}_h, h = 1, \dots, \infty)$. 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 $a_h = 1, b_h = \alpha, h = 1, \dots, \infty$.

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

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 $\Omega_{\mathbf{0}, \mathbf{I}}^r$ 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 $\boldsymbol{\eta}_i \sim G$, $i = 1, \dots, n$, $G \sim \mathcal{P}$, with \mathcal{P} a CSBP, is equivalent to the following:

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

where $\boldsymbol{\mu}_{G^*}$, $\boldsymbol{\Sigma}_{G^*}$, $\mathbf{V} = (V_h, h = 1, \dots, \infty)'$ 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 (PE), model. His PE 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 PE 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, while providing some theoretical justification.

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

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

where $V_h \sim \text{beta}(a_h, b_h)$, $h = 1, \dots, N-1$, $V_N = 1$, $\boldsymbol{\theta}_h^* \stackrel{iid}{\sim} G_0$, $h = 1, \dots, N$, 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 $\boldsymbol{\eta}_i \sim G$, $i = 1, \dots, n$, with $G \sim \mathcal{P}^N$, we can obtain the following equivalent specification:

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

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 $\rightarrow 0$ at an exponential rate as N increases, suggesting that a highly accurate approximation can be obtained for moderate sized N in most cases.

In order to argue that \mathcal{P}_N is also close to \mathcal{P} in \mathcal{L}_1 for moderate sized N , we rely on the Ishwaran and James (2001) result, while also examining the accuracy in approximating the random mean and covariance in Theorem 1.

Theorem 1. Assuming G_0 is the probability measure corresponding to $N_q(\mathbf{0}, \mathbf{I})$,

$a_h = 1, b_h = \alpha, h = 1, \dots, \infty$, we have

$$\mathbb{E}(\boldsymbol{\mu}_G - \boldsymbol{\mu}_{G_N}) = \mathbf{0} \quad \text{and} \quad \mathbb{V}(\boldsymbol{\mu}_G - \boldsymbol{\mu}_{G_N}) = \mathbb{E}(\boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_{G_N}) = \left(\frac{\alpha}{\alpha + 2} \right)^{N-1} \left(\frac{\alpha}{\alpha + 1} \right) \mathbf{I}.$$

3.4 Centered Stick-Breaking Mixtures

Assuming $\boldsymbol{\eta}_i \sim G$, $i = 1, \dots, n$, with $G \sim \mathcal{P}$ and \mathcal{P} a CSBP, G is almost surely discrete. Hence, the n $r \times 1$ 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 $\{1, \dots, n\}$, 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

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

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:

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

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 $V(\boldsymbol{\mu}_i | G) = \boldsymbol{\Sigma}_{G^*}(\boldsymbol{\Sigma}_{G^*} + \mathbf{I})^{-1}$, so that

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

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 $\boldsymbol{\tau}^* = (\tau_1^*, \dots, \tau_p^*)'$, $\boldsymbol{\Lambda}^* = (\lambda_1^*, \dots, \lambda_p^*)'$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$. 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 $G^* \sim \text{CSBP}$, and 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 parameter-expanded posterior, we transform back to the original hierarchical model using:

$$\tau_j = \tau_j^* + \lambda_j^{*'} \mu_{G_N^*}, \quad \lambda_j = \lambda_j^{*'} \Sigma_{G_N^*}^{1/2}, \quad \eta_i = \Sigma_{G_N^*}^{-1/2} (\eta_i - \mu_{G_N^*}). \quad (12)$$

Note that the convergence and mixing rates for the τ , Λ and η parameters tends to be improved over that for the τ^* , Λ^* , and η^* .

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.

5. Simulation Examples

We assessed the performance of the approach through a simulation example designed to mimic the fibroid data application in section 6. In this application, we are interested in inference under a latent factor regression model:

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

with x_i a 4×1 predictor vector, β a 4×1 vector of regression coefficients, and G a unknown latent variable residual density having mean 0 and variance 1. For the simulation, we assume that the true parameter values are $\beta = (1, 1, 1, 1)'$, $\Lambda = (1, 1, 1, 1, 1, 1, 1)'$ and the latent variable density η_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.

The measurement model relating η_i to the observed y_i is described in section 6. The values of $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ 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, α , 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:

$$\pi(\boldsymbol{\mu}) = N_p(\mathbf{0}, 10\mathbf{I}), \quad \pi(\boldsymbol{\lambda}) = \prod_{j=1}^p N_+(\lambda_j; 0, 10), \quad \pi(\boldsymbol{\tau}) = \prod_{j=1}^p G(\tau_j, 1, 1),$$

where $N_+(m, v)$ refers to the $N(m, v)$ distribution truncated to $(0, \infty)$, and $\boldsymbol{\tau} = (\sigma_1^{-2}, \dots, \sigma_p^{-2})$.

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 Figure 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.

6. Epidemiology Application

6.1 *Scientific Background and Data Description*

We illustrate the method using 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.

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 followup, 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 ($< 2\text{cm}$), medium (between 2 and 4cm) or large ($> 4\text{cm}$). 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.

6.2 Model and Prior Specification

Letting η_i denote the latent bleeding intensity score for woman i , we used model (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 η_i , to the ordered categorical symptom data, we used a continuation ratio measurement model:

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

where C_j is the number of categories for symptom type j , $\lambda_1, \dots, \lambda_7$ are the loading factors for symptoms $Y_1 - Y_7$. Prior specification was as described in section 5 for the simulation example.

6.3 Analysis and Results

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 Figure 2. Interestingly, the density is quite similar to a normal density, though we have demonstrated power to detect non-normality in the simulation example.

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 Figure 3. The plot for white women and other ethnicities shows a very similar pattern. For symptoms 2, 4 and 5, there are essentially no difference across the fibroid size categories and the factor loadings 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 symptoms 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.

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 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.

Acknowledgments

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

APPENDIX: PROOF OF THEOREM 1

Let $\boldsymbol{\mu}_G = \sum_{h=1}^{\infty} V_h \prod_{l<h} \bar{V}_l \boldsymbol{\theta}_h$, $\boldsymbol{\mu}_G^N = \sum_{h=1}^{N-1} V_h \prod_{l<h} \bar{V}_l \boldsymbol{\theta}_h + \prod_{l<N} \bar{V}_l \boldsymbol{\theta}_N$ and $\boldsymbol{\Delta}_N = \boldsymbol{\mu}_G^N - \boldsymbol{\mu}_G$.

$$\begin{aligned} \boldsymbol{\Delta}_N &= \left(\prod_{l=1}^N \bar{V}_l \right) \left(\boldsymbol{\theta}_N - \boldsymbol{\theta}_{N+1} V_{N+1} - \boldsymbol{\theta}_{N+2} V_{N+2} \bar{V}_{N+1} - \boldsymbol{\theta}_{N+3} V_{N+3} \bar{V}_{N+2} \bar{V}_{N+1} - \dots \right) \\ &= \left(\prod_{l=1}^N \bar{V}_l \right) \sum_{h=1}^{\infty} V_h^* \prod_{l<h} \bar{V}_l^* \boldsymbol{\Theta}_h, \end{aligned}$$

where $\bar{V}_h = 1 - V_h$, $V_h^* = V_{N+h}$, $\boldsymbol{\Theta}_h = \boldsymbol{\theta}_N - \boldsymbol{\theta}_{N+h}$, $h = 1, \dots, \infty$. Assuming G_0 corresponds to $N_q(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\Theta}_h \sim N_q(\boldsymbol{\theta}_N, \mathbf{I})$, for $h = 1, \dots, \infty$. Letting $\boldsymbol{\Delta}_N = \lim_{T \rightarrow \infty} \boldsymbol{\Delta}_N^T$ with $\boldsymbol{\Delta}_N^T$ setting terms $h = T + 1, \dots, \infty$ to 0, we have

$$(\boldsymbol{\Delta}_N^T | V_1, \dots, V_{N+T}) \sim N_q \left(\mathbf{0}, \left(\prod_{l=1}^N \bar{V}_l^2 \right) \left(1 + \sum_{h=1}^T V_h^{*2} \prod_{l<h} \bar{V}_l^{*2} \right) \mathbf{I} \right).$$

It follows directly that $\mathbb{E}(\boldsymbol{\Delta}_N^T) = \mathbf{0}$ and

$$\begin{aligned} \mathbb{V}(\boldsymbol{\Delta}_N^T) &= \mathbb{E} \left\{ \left(\prod_{l=1}^N \bar{V}_l^2 \right) \left(1 + \sum_{h=1}^T V_h^{*2} \prod_{l<h} \bar{V}_l^{*2} \right) \mathbf{I} \right\}, \\ &= \left(\frac{\alpha}{\alpha + 2} \right)^N \left\{ 1 + \frac{2}{(\alpha + 1)(\alpha + 2)} \sum_{h=0}^T \left(\frac{\alpha}{\alpha + 2} \right)^h \right\}. \end{aligned}$$

Taking the limit as $T \rightarrow \infty$, we then have

$$\mathbb{E}(\boldsymbol{\Delta}_N) = \mathbf{0} \quad \text{and} \quad \mathbb{V}(\boldsymbol{\Delta}_N) = \left(\frac{\alpha}{\alpha + 2} \right)^{N-1} \left(\frac{\alpha}{\alpha + 1} \right) \mathbf{I}.$$

Letting $\Psi_N = \Sigma_G - \Sigma_{G_N}$ and noting $E(\boldsymbol{\theta}_h \boldsymbol{\theta}'_h) = \mathbf{I}$, we have

$$\begin{aligned} E(\Psi_N) &= E\left\{ \sum_{h=1}^{\infty} V_h \prod_{l<h} \bar{V}_l \boldsymbol{\theta}_h \boldsymbol{\theta}'_h \right\} - E\left\{ \sum_{h=1}^{N-1} V_h \prod_{l<h} \bar{V}_l \boldsymbol{\theta}_h \boldsymbol{\theta}'_h + \prod_{l<N} \bar{V}_l \boldsymbol{\theta}_N \boldsymbol{\theta}'_N \right\} \\ &\quad - E(\boldsymbol{\mu}_G \boldsymbol{\mu}'_G) + E(\boldsymbol{\mu}_{G_N} \boldsymbol{\mu}'_{G_N}) \mathbf{I}. \end{aligned}$$

The first line of this expression reduces to

$$\left(\frac{\alpha}{\alpha+1} \right)^{N-1} \left\{ \frac{1}{\alpha+1} - 1 + \left(\frac{\alpha}{\alpha+1} \right) \left(\frac{1}{\alpha+1} \right) \sum_{h=0}^{\infty} \left(\frac{\alpha}{\alpha+1} \right)^h \right\} \mathbf{I} = \mathbf{0}.$$

Hence, $E(\Psi_N)$ equals the second line, which reduces to

$$\begin{aligned} &\left\{ -E(V^2) \prod_{l=1}^{N-1} E(\bar{V}^2) - \prod_{l=1}^N E(\bar{V}^2) \sum_{h=1}^{\infty} E(V^2) \prod_{l=1}^{h-1} E(\bar{V}^2) + \prod_{l=1}^{N-1} E(\bar{V}^2) \right\} E(\boldsymbol{\theta} \boldsymbol{\theta}') \\ &= \left(\frac{\alpha}{\alpha+2} \right)^{N-1} \left\{ -\frac{2}{(\alpha+1)(\alpha+2)} + 1 - \frac{\alpha}{\alpha+2} \frac{2}{(\alpha+1)(\alpha+2)} \frac{\alpha+2}{2} \right\} \mathbf{I} \\ &= \left(\frac{\alpha}{\alpha+2} \right)^{N-1} \left(\frac{\alpha}{\alpha+1} \right) \mathbf{I}, \end{aligned}$$

dropping the subscripts on $V_h, \boldsymbol{\theta}_h$ in taking expectations.

References

- Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D. and 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.
- Brunner, L. and Lo, A. (1989), "Bayes Methods for a Symmetric Unimodal Density and its Mode," *The Annals of Statistics*, 17, 1550-1566.
- Burr, D. and Doss, H. (2005), "A Bayesian Semiparametric Model for Random-Effects Meta-Analysis," *Journal of the American Statistical Association*.
- Bush, C.A. and 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., Watson, M. and 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.
- Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209-230.
- Ferguson, T.S. (1974), "Prior Distributions on Spaces of Probability Measures," *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. and 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," *Annals of Statistics*, 31, 174-200.
- Ishwaran, H. and James, L.F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161-173.

- Ishwaran, H. and Takahara, G. (2002), “Independent and Identically Distributed Monte Carlo Algorithms for Semiparametric Linear Mixed Models,” *Journal of the American Statistical Association*, 97, 1154-1166.
- Ishwaran, H. and Zarepour, M. (2000), “Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models,” *Biometrika*, 87, 371-390.
- 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. and Ibrahim, J.G. (1998), “A Semiparametric Bayesian Approach to the Random Effects Model,” *Biometrics*, 54, 921-938.
- Kottas, A. and 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. and Mockus, A. (1995), “A Nonparametric Bayes Method for Isotonic Regression,” *Journal of Statistical Planning and Inference*, 46, 235-248.
- Li, Y., Muller, P., and Lin, X. “Bias-Corrected Inference in Semiparametric Bayesian Mixed Models” *Technical Report*.
- Liu, J.S. and Wu, Y.N. (1999), “Parameter Expansion for Data Augmentation,” *Journal of the American Statistical Association*, 94, 1264-1274.
- Lopes, H.F. and West, M. (2004), “Bayesian Model Assessment in Factor Analysis,” *Statistica Sinica*, 14, 41-67.
- Mukhopadhyay, S. and Gelfand, A.E. (1997), “Dirichlet Process Mixed Generalized Linear Models,” *Journal of the American Statistical Association*, 92, 633-639.

- Pison, G., Rousseeuw, P.J., Flizmoser, P. and Croux, C. (2003), "Robust Factor Analysis," *Journal of Multivariate Analysis*, 84, 145-172.
- Pison, G. and Van Aelst, S. (2004), "Diagnostic Plots for Robust Multivariate Methods," *Journal of Computational and Graphical Statistics*, 13, 310-329.
- Regazzini, E., Guglielmi, A. and Di Nunno, G. (2002), "Theory and Numerical Analysis for Exact Distributions of Functionals of a Dirichlet Process," *The Annals of Statistics*, 30, 1376-1411.
- Wegienka, G. Baird, D., Hertz-Picciotto, I., Harlow, S., Steege, J., Hill, M., Schectman, M. and 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.

Table 1: Parameter estimation of DPM & CDPM for simulation

Parameter	True value	DPM		CDPM	
		Estimate	95 % CI	Estimate	95 % CI
β_1	1.00	1.66	(1.27,2.09)	0.88	(0.68,1.07)
β_2	1.00	1.61	(1.26,2.00)	0.85	(0.68,1.01)
β_3	1.00	1.76	(1.38,2.24)	0.93	(0.74,1.14)
β_4	1.00	1.73	(1.44,2.11)	0.92	(0.78,1.05)
λ_1	1.00	0.54	(0.44,0.63)	1.02	(0.89,1.15)
λ_2	1.00	0.56	(0.46,0.65)	1.06	(0.92,1.20)
λ_3	1.00	0.56	(0.45,0.67)	1.06	(0.91,1.21)
λ_4	1.00	0.58	(0.48,0.69)	1.10	(0.94,1.29)
λ_5	1.00	0.62	(0.48,0.79)	1.18	(0.97,1.42)
λ_6	1.00	0.61	(0.49,0.72)	1.14	(0.98,1.32)
λ_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
β_1	0.065	(-0.21, 0.35)	0.05	(-0.18,0.28)
β_2	0.53	(0.29, 0.91)	0.45	(0.25,0.66)
β_3	0.91	(0.60, 1.45)	0.76	(0.51,1.01)
β_4	0.54	(0.34,0.87)	0.46	(0.28,0.63)
λ_1	0.51	(0.27,0.71)	0.60	(0.43, 0.83)
λ_2	0.018	(0.00,0.06)	0.02	(1.04,1.86)
λ_3	1.17	(0.73,1.55)	1.37	(1.04, 1.86)
λ_4	0.02	(0.00,0.07)	0.024	(0.00,0.02)
λ_5	0.12	(0.05,0.19)	0.14	(0.06, 0.14)
λ_6	0.96	(0.61,1.23)	1.13	(0.88, 1.50)
λ_7	0.80	(0.51,1.01)	0.93	(0.73, 1.23)

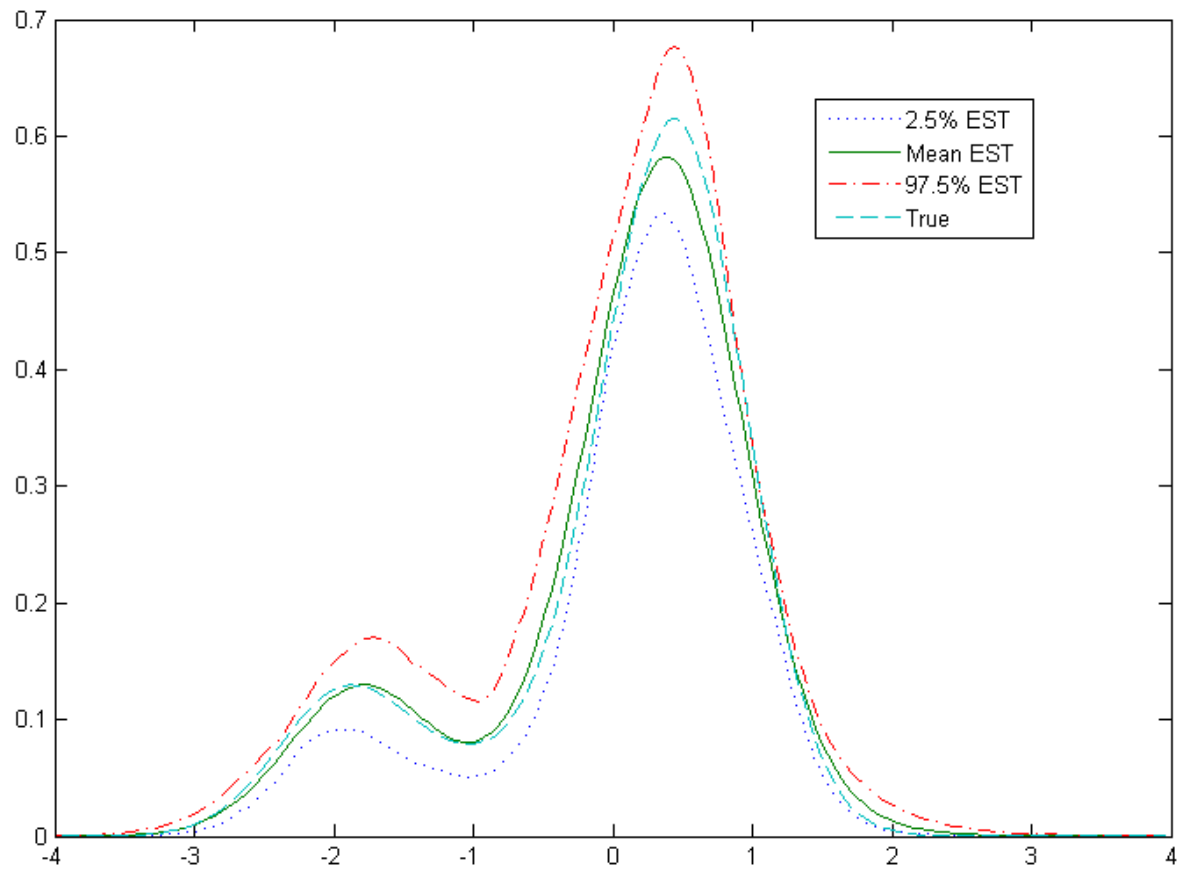


Figure 1: True and estimated latent variable densities in simulation example.

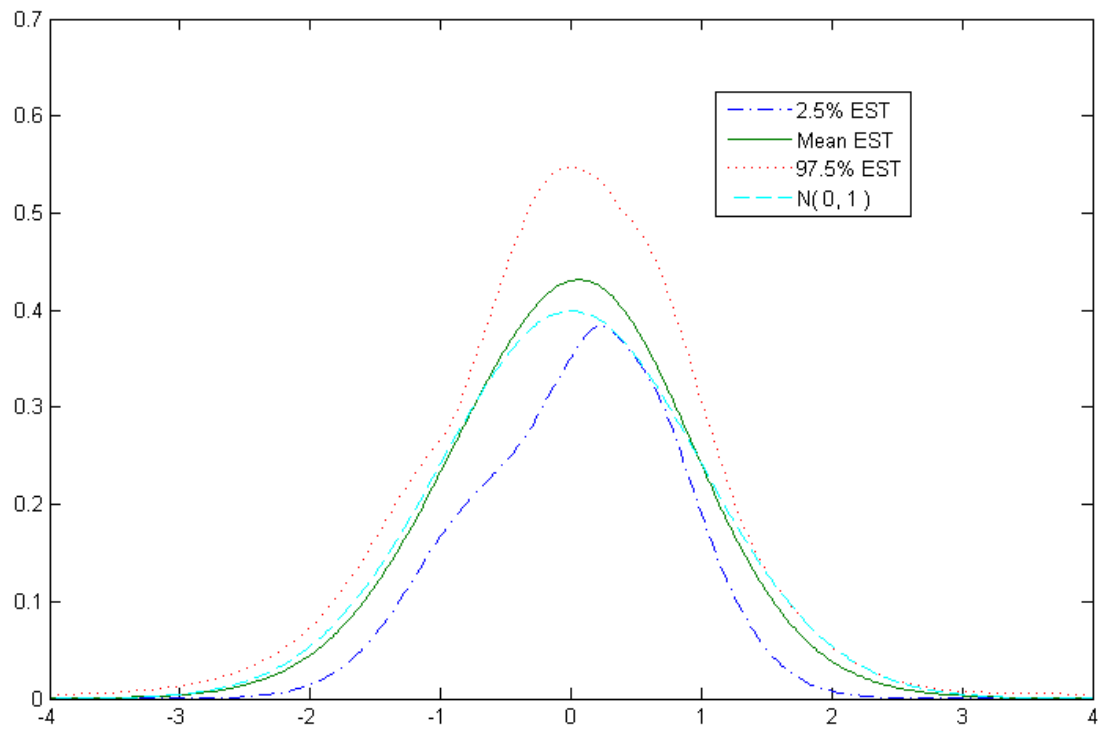


Figure 2: Estimated density of the latent bleeding intensity score in the fibroid data application

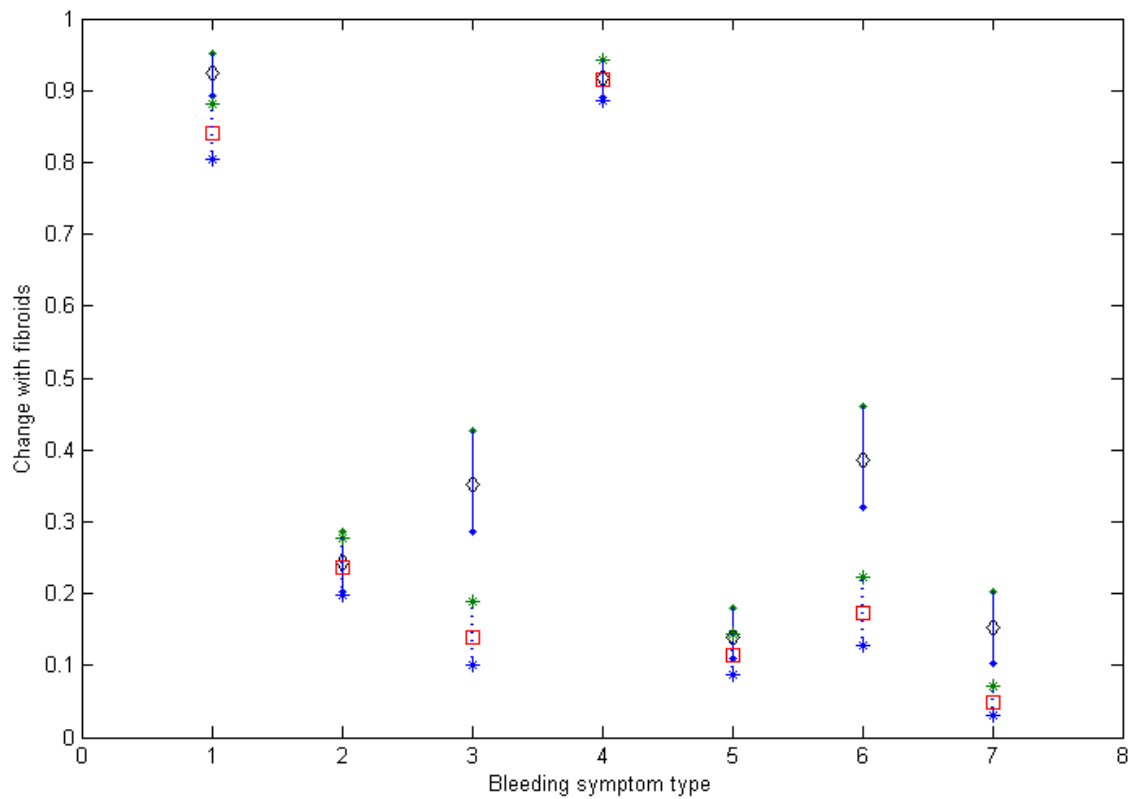


Figure 3: Comparison of bleeding symptoms for black women with large fibroid size (solid line, diamond: estimated mean) vs. no fibroids (dashed line, square:estimated mean)