

Variable Selection in Nonparametric Random Effects Models

Bo Cai and David B. Dunson

Biostatistics Branch, MD A3-03

National Institute of Environmental Health Sciences

P.O. Box 12233

Research Triangle Park, NC 27709, U.S.A.

In analyzing longitudinal or clustered data with a mixed effects model (Laird and Ware, 1982), one may be concerned about violations of normality. Such violations can potentially impact subset selection for the fixed and random effects components of the model, inferences on the heterogeneity structure, and the accuracy of predictions. This article focuses on Bayesian methods for subset selection in nonparametric random effects models in which one is uncertain about the predictors to be included and the distribution of their random effects. We characterize the unknown distribution of the individual-specific regression coefficients using a weighted sum of Dirichlet process (DP)-distributed latent variables. By using carefully-chosen mixture priors for coefficients in the base distributions of the component DPs, we allow fixed and random effects to be effectively dropped out of the model. A stochastic search Gibbs sampler is developed for posterior computation, and the methods are illustrated using simulated data and real data from a multi-laboratory bioassay study.

Key Words: Dirichlet process; Gibbs sampler; Latent Variables; Nonparametric Bayes; Random effects; Stochastic search; Subset selection.

1. INTRODUCTION

Linear mixed effects models (Laird and Ware, 1982), which assume normally distributed random effects, are widely used in analyzing longitudinal and clustered data. Concern about sensitivity to the normality assumption (Agresti, Caffo and Ohman-Strickland, 2004) has motivated the development of nonparametric methods, which allow the random effects distribution to be estimated. Some recent frequentist papers on this topic include Zhang and Davidian (2001), Chen, Zhang and Davidian (2002), Lai and Shih (2003), and Ghidry, Lesaffre and Eilers (2004). In addition, there is a rich literature on Bayesian methods using Dirichlet process (DP) priors (Ferguson, 1973), DP mixtures (DPM) (Antoniak, 1974), and other specifications to allow unknown random effects distributions (Bush and MacEachern, 1996; Kleinman and Ibrahim, 1998; Ishwaran and Takahara, 2002; Lopes, Müller and Rosner, 2003; Müller et al., 2005; among others). All of these methods do not accommodate uncertainty in the predictors to be included in the fixed and random effects components of the model.

The focus of this article is on developing Bayesian methods for accommodating such uncertainty, while allowing the joint distribution of the random effects included in the model to be unknown. Such methods are also useful for comparing models with and without random effects, a problem addressed by Lin (1997) using a global score test for the null hypothesis that all variance components are zero. Lin's approach does not require specification of a parametric form for the random effects density. Later authors have considered alternative score tests (Hall and Praestgaard, 2001; Verbeke and Molenberghs, 2003; Zhu and Fung, 2004) and generalized likelihood ratio tests (Craniniceanu and Ruppert, 2004). These methods are not useful for the general subset selection problem. Bayesian methods for the model comparison in normal variance component models have been proposed by Pauler, Wakefield and Kass (1999), Sinharay and Stern (2001) and Chen and Dunson (2003).

A Bayesian solution to the simplified problem of comparing a normal linear mixed model with a random intercept to the corresponding model without the random intercept has been considered by Albert and Chib (1997). They proposed using a mixture prior with point mass at zero for the

random effects variance. Potentially, the normal random effect assumption could be relaxed by using a DPM for the unknown random effects distribution. One could then compare the resulting semiparametric Bayesian model with the fully parametric linear model that excludes the random effect using the approach of Basu and Chib (2003), which was developed to calculate marginal likelihoods and Bayes factors for DPMs. Such an approach is potentially useful when the number of competing random effects models is small. However, in subset selection problems, the number of possible models is typically very large, and it is necessary to develop an automated search procedure that does not require fitting of all the models in the list (Geweke, 1996; George and McCulloch, 1997). For a review of Bayesian approaches to this problem in parametric models, refer to Clyde and George (2004).

Chen and Dunson (2003) proposed a Bayesian approach for random effects selection in the normal linear mixed model based on using variable selection priors for the components in a special decomposition of the random effects covariance. In particular, they proposed using mixture priors with point mass at zero for key parameters in the decomposition to allow random effects to effectively drop out of the model. The decomposition and priors chosen led to convenient computation via a stochastic search variable selection (SSVS) algorithm. Unfortunately, it is not straightforward to modify this procedure to allow unknown random effects distributions due to difficulties in incorporating moment constraints.

To avoid the need for such moment constraints, and to improve mixing through the use of a centered parameterization, we reparameterize the linear mixed model using latent variables relating to the fixed and random effects components. Our specification places independent DPs on these latent variables, resulting in a weighted convolution of DPs for the random effects. This differs from the usual approach, which places a DP directly on the random effects distributions as in Kleinman and Ibrahim (1998). By using carefully-tailored mixture priors for parameters in the base measures of the component DPs, our specification allows fixed and random effects selection.

In Section 2, we present our reparameterization of the linear mixed model, nonparametric specification, and variable selection priors. In Section 3, we outline the posterior computational

strategy. In Section 4, we illustrate the approach by a simulation study. In Section 5, we apply the method to data from a multi-laboratory bioassay study. We provide the discussion in Section 6.

2. Priors for Nonparametric Random Effects Selection

2.1 Nonparametric Linear Mixed Model

For observation j ($j = 1, \dots, n_i$) from subject i ($i = 1, \dots, n$), let y_{ij} denote the response variable, let \mathbf{x}_{ij} denote a $p \times 1$ vector of candidate predictors, and let \mathbf{z}_{ij} denote a $q \times 1$ vector of candidate predictors. The normal linear mixed effects model can be expressed as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\zeta}_i + \boldsymbol{\epsilon}_i, \quad (1)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})'$, $\mathbf{Z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})'$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effect regression coefficients (referred to as fixed since the coefficients are constant for all subjects), $\boldsymbol{\zeta}_i = (\zeta_{i1}, \dots, \zeta_{iq})' \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$ is a $q \times 1$ vector of subject-specific random effects with covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\epsilon}_i$ is a residual vector, typically assumed to be $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

Our focus is on selecting the predictors to be included in the fixed and random effects components of the model. For the fixed effects component, there are p candidate predictors, while for the random effects component, there are q candidate predictors. In implementing subset selection and Bayesian model averaging, we also want to avoid the assumption of normality of the random effects. In the simplified case in which all of the candidate predictors are included, we could let $\boldsymbol{\zeta}_i \sim G$, where G is the unknown random effect distribution. Following a Bayesian approach, as in Kleinman and Ibrahim (1998), we could then choose a prior distribution for G with support on the space of random probability measures. A natural choice would be the Dirichlet process prior, which could be specified as $G \sim DP(\alpha G_0)$, where α is a precision parameter and G_0 is the base distribution of the Dirichlet process.

Under this specification, for any partition $\mathbf{B} = (B_1, \dots, B_k)'$ of \mathfrak{R} , we have

$$\{G(B_1), \dots, G(B_k)\} \sim D(\alpha G_0(B_1), \dots, \alpha G_0(B_k)),$$

where $D(\cdot)$ denotes the finite Dirichlet density. In addition, we have $E(G) = G_0$, with a natural choice of G_0 being the $N_q(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, so that the prior is centered on the normal linear

mixed model of expression (1). Under this specification, the expected value of \mathbf{y}_i conditional on \mathbf{X}_i and \mathbf{Z}_i , but integrating out the random effects ζ_i , is $E(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_iE(\zeta_i)$. Under the stick-breaking representation of Sethuraman (1994), we have

$$G = \sum_{t=1}^{\infty} p_t \delta_{\boldsymbol{\xi}_t}, \quad p_t / \prod_{l=1}^{t-1} (1 - p_l) \stackrel{iid}{\sim} \text{beta}(1, \alpha), \quad \boldsymbol{\xi}_t \stackrel{iid}{\sim} G_0, \quad (2)$$

with δ_{ξ} denoting the degenerate distribution with all its mass at ξ . Hence, the random distribution G can be represented as a infinite set of point masses at locations generated independently from the base distribution. This implies that $E(\zeta_i)$, for $\zeta_i \sim G$, is equal to

$$E(\zeta_i) = \sum_{t=1}^{\infty} p_t E(\boldsymbol{\xi}_t) = \mathbf{0},$$

so that $E(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{X}_i\boldsymbol{\beta}$, as required in order for $\boldsymbol{\beta}$ to be interpretable as fixed effects.

A potential problem with the parameterization shown in expression (1) from a Bayesian perspective is the *uncentered* form. In particular, it is well known that computational efficiency of Gibbs sampling algorithms for posterior computation in linear mixed models (and other hierarchical models) tends to depend strongly on the parameterization (Gelfand, Sahu and Carlin, 1995). For greater efficiency, one can focus on the centered parameterization: $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$, with $\boldsymbol{\beta}_i \sim G$, $G \sim DP(\alpha G_0)$, and $G_0 = N_p(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, assuming $\mathbf{X}_i = \mathbf{Z}_i$ so $p = q$. To allow for uncertainty in $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, normal and inverse-Wishart priors can be chosen. Unfortunately, this approach assumes that the same predictors are included in the fixed and random components, and these predictors are known.

2.2 Reparameterization and Nonparametric Priors

Following Geweke (1996) and others, subset selection for the fixed effect predictors can potentially proceed by choosing mixture priors for the regression coefficients $\boldsymbol{\beta}$. In particular, because $\beta_l = 0$ corresponds to the l th candidate predictor being effectively excluded from the fixed effect component, a prior that assigns positive probability to both $H_{0l} : \beta_l = 0$ and $H_{1l} : \beta_l \neq 0$, for $l = 1, \dots, p$, allows uncertainty in the subset of predictors to be included. In linear regression, many choices of mixture priors have been proposed, and a variety of algorithms are available for posterior computation.

Subset selection for the random effects component is more challenging. For the normal linear mixed model, one can effectively remove the l th random effect by setting the elements in the l th row and column of Σ equal to 0. This implies zero variance for the l th random effect, so that we have $\zeta_i \equiv \mathbf{0}$ for all i , and hence there is no heterogeneity among subjects in the effect of the l th candidate predictor. Potentially, a predictor can have a fixed effect without a random effect and vice versa. Due to the non-negative definite constraint, it is difficult to specify a prior directly for Σ that allows zero rows and columns. Certainly, the commonly used inverse-Wishart prior is not sufficiently flexible. For this reason, Chen and Dunson (2003) proposed a reparameterization of model (1) based on a decomposition of Σ that facilitated variable selection (see also Cai and Dunson, 2005).

The Chen and Dunson (2003) approach relies on the incorporation of standard normal latent variables underlying the random effects. Therefore, the approach does not generalize naturally to the nonparametric case in which the distributions of the random effects to be included are unknown. Although one could conceptually replace the standard normal assumption with the assumption that the latent variables have an unknown distribution with mean 0 and variance 1, such moment constraints are not straightforward to incorporate in existing nonparametric priors. In addition, the Chen and Dunson (2003) parameterization is uncentered, leading to potential problems with slow mixing.

Instead, letting \mathbf{X}_i include the candidate predictors for both the fixed and random effects components, we propose a new approach based on the centered parameterization:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad (3)$$

where $\boldsymbol{\beta}_i$ follows an unknown distribution with

$$\beta_{ih} = \beta_{ih}^* + \sum_{l=1}^{h-1} \gamma_{hl} (\beta_{il}^* - \beta_l), \quad \text{for } h = 1, \dots, p, \quad (4)$$

where $\boldsymbol{\beta}_i^* = (\beta_{i1}^*, \dots, \beta_{ip}^*)'$ is a vector of latent variables underlying $\boldsymbol{\beta}_i$, and γ_{hl} ($h = 1, \dots, p$, $l = 1, \dots, h-1$) is the (h, l) th element of the lower triangular matrix $\boldsymbol{\Gamma}$. Let $\beta_{ih}^* \stackrel{iid}{\sim} G_h^*$, independently for $h = 1, \dots, p$, with $G_h^* \sim DP(\alpha_h G_{0h})$, where G_{0h} is the normal distribution with mean β_h and variance ψ_h .

Let $\xi_{ht}, t = 1, \dots, \infty$, denote the atoms in G_h^* . From the stick-breaking formulation (2), it follows that

$$\begin{aligned} \mathbb{E}(\beta_{ih}) &= \mathbb{E}(\beta_{ih}^*) + \sum_{l=1}^{h-1} \gamma_{hl} \{\mathbb{E}(\beta_{il}^*) - \beta_l\} \\ &= \mathbb{E} \left\{ \sum_{t=1}^{\infty} p_{ht} \mathbb{E}(\xi_{ht}) \right\} + \sum_{l=1}^{h-1} \gamma_{hl} \left[\mathbb{E} \left\{ \sum_{s=1}^{\infty} p_{ls} \mathbb{E}(\xi_{ls}) \right\} - \beta_l \right] = \beta_h, \end{aligned} \quad (5)$$

with $\{p_{ht}, t = 1, \dots, \infty\}$ denoting the random probability mass for G_h^* . Hence, $\mathbb{E}(\beta_i) = \beta$, so that β are interpretable as fixed effect regression coefficients, with the subset of predictors having $\beta_l = 0$ effectively excluded from the fixed effect component. In addition, the random effects variance is

$$\begin{aligned} \mathbb{V}(\beta_{ih}) &= \mathbb{V}(\beta_{ih}^*) + \sum_{l=1}^{h-1} \gamma_{hl}^2 \mathbb{V}(\beta_{il}^*) = \mathbb{E} \left[\sum_{t=1}^{\infty} p_t \mathbb{E}\{(\xi_{ht} - \beta_h)^2\} \right] + \sum_{l=1}^{h-1} \gamma_{hl}^2 \mathbb{E} \left[\sum_{s=1}^{\infty} p_{ls} \mathbb{E}\{(\xi_{ls} - \beta_l)^2\} \right] \\ &= \psi_h + \sum_{l=1}^{h-1} \gamma_{hl}^2 \psi_l, \end{aligned} \quad (6)$$

and the random effects covariance ($h < m$) is

$$\begin{aligned} \text{Cov}(\beta_{im}, \beta_{ih}) &= \mathbb{E}(\beta_{im}\beta_{ih}) - \mathbb{E}(\beta_{im})\mathbb{E}(\beta_{ih}) \\ &= \mathbb{E} \left\{ \left[\beta_{im}^* + \sum_{s=1}^{m-1} \gamma_{ms}(\beta_{is}^* - \beta_s) \right] \left[\beta_{ih}^* + \sum_{l=1}^{h-1} \gamma_{hl}(\beta_{il}^* - \beta_l) \right] \right\} - \beta_m\beta_h \\ &= \mathbb{E} \left[\gamma_{mh}\beta_{ih}^*(\beta_{ih}^* - \beta_h) + \sum_{s=1}^{m-1} \gamma_{ms}(\beta_{is}^* - \beta_s) \cdot \sum_{l=1}^{h-1} \gamma_{hl}(\beta_{il}^* - \beta_l) \right] \\ &= \gamma_{mh}\mathbb{V}(\beta_{ih}^*) + \sum_{l=1}^{h-1} \gamma_{hl}\gamma_{ml}\mathbb{E}(\beta_{il}^* - \beta_l)^2 \\ &= \sum_{l=1}^h \gamma_{ml}\gamma_{hl}\psi_l, \end{aligned} \quad (7)$$

with $\mathbb{V}(\beta_i) = \mathbf{\Gamma}\mathbf{\Psi}\mathbf{\Psi}\mathbf{\Gamma}'$ (the standard Cholesky decomposition of the covariance matrix) and $\mathbf{\Psi} = \text{diag}(\psi_1^{\frac{1}{2}}, \dots, \psi_p^{\frac{1}{2}})$. Hence, the l th random effect can be excluded by setting $\psi_l = 0$, and further restricting $\gamma_{hl} = 0$ if $\psi_h = 0$ or $\psi_l = 0$.

Note that as $\psi_h \rightarrow 0$, all the atoms in G_h^* are effectively generated from a point mass at β_h , so that there is no heterogeneity in the β_{ih} coefficients among subjects, regardless of the value of α_h . However, if $\psi_h > 0$ and the h th random effect is included, then the precision parameter α_h plays an important role in determining the degree of shrinkage of G_h^* towards G_{0h} and the

degree of clustering. The clustering property will be discussed further in Section 3. From the standpoint of selection of fixed and random effects, a crucial property of our specification is that we can drop predictors out of the fixed and random effects components by choosing mixture priors for the parameters β and Ψ characterizing the normal G_0 . Hence, the nonparametric characterization does not complicate the selection process, as we illustrate in Section 3.

Our interest is in the collection of random distributions $\mathbf{G} = \{G_1, \dots, G_p\}$ of the random effects. Due to the structure of (4), the distributions within \mathbf{G} will be dependent due to dependence on the shared DP-distributed basis distributions G_1^*, \dots, G_p^* . Theorem 1 describes basic properties of the distributions within \mathbf{G} . We expect that related weighted convolution approaches for defining priors for dependent, unknown distributions should be broadly useful. For simplicity, we let $\otimes_{l=1}^h A_l = A_1 \otimes A_2 \otimes \dots \otimes A_h$ and $\beta_{il}^{*h} = \gamma_{hl}(\beta_{il}^* - \beta_l)$, for $h = 1, \dots, p$ and $l = 1, \dots, h$. The proof is given in Appendix A.

Theorem 1 *Suppose that the elements of $\beta_h = (\beta_{1h}, \dots, \beta_{nh})'$, for $h = 1, \dots, p$, in the linear mixed model (3) satisfy the expression (4), with the latent variables $\beta_{ih}^* \stackrel{iid}{\sim} G_h^*$, for $i = 1, \dots, n$, with $G_h^* \sim DP(\alpha_h G_{0h})$. Let G_{0l}^{*h} denote the base measure of the resulting $DP(\alpha_l G_{0l}^{*h})$ for β_{il}^{*h} , for $l = 1, \dots, h$. Let $\mathcal{B} \subseteq \mathcal{R}$ denote a Borel set, and $f \otimes g$ denote the convolution of f and g . Then we have*

1. $E(G_h(\mathcal{B})) = \otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B});$
2. $Var(G_h(\mathcal{B})) = \frac{\otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B})}{\prod_{l=1}^h (1 + \alpha_l)} \left[1 + \sum_{l=1}^h \alpha_l G_{0l}^{*h}(\mathcal{B}) + \sum_{i \neq j}^h \alpha_i \alpha_j G_{0i}^{*h} \otimes G_{0j}^{*h}(\mathcal{B}) + \dots - \left(\prod_{l=1}^h (1 + \alpha_l) - \prod_{l=1}^h \alpha_l \right) \otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B}) \right];$
3. $Cov(G_h(\mathcal{B}), G_m(\mathcal{B})) = \frac{\otimes_{l=1}^m G_{0l}^{*m}(\mathcal{B})}{\prod_{l=1}^h (1 + \alpha_l)} \left[1 + \sum_{l=1}^h \alpha_l G_{0l}^{*m}(\mathcal{B}) + \sum_{i \neq j}^h \alpha_i \alpha_j G_{0i}^{*m} \otimes G_{0j}^{*m}(\mathcal{B}) + \dots - \left(\prod_{l=1}^h (1 + \alpha_l) - \prod_{l=1}^h \alpha_l \right) \otimes_{l=1}^h G_{0l}^{*m}(\mathcal{B}) \right], \text{ for } h < m.$

In Theorem 1, each random distribution within \mathbf{G} is centered at the convolutions of the corresponding base measures of the implied DPs for β_{il}^{*h} 's. The variance of the random distribution G_h within \mathbf{G} is a function of the base measures and the precision parameters of the DPs for β_{il}^{*h} 's. It is clear that when $\alpha_l \rightarrow \infty$, for $l = 1, \dots, h$, the random distribution G_l^* tends to G_{0l} . In the extreme

case of all α 's tending to infinity, which means that all latent variables are generated directly from their corresponding base distributions (i.e. the parametric case), the variance and covariance of the random distributions, G_h 's, are zeroes, implying that the random distribution G_h becomes fixed. In terms of the precision parameter, the dependence of two random distributions, G_h and G_m , relies only on the first h α 's ($h < m$). This is reasonable because, for example, when the first h α 's tend to infinity, G_h becomes fixed so that there is no relationship between G_h and G_m . In contrast, in the limit as each of the α 's tends to zero, which corresponds to the degenerate case in which there is a single atom in each component distribution, the variance of the random distribution G_h reduces to $\otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B}) [1 - \otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B})]$.

In the important special case in which the base measure G_{0h} corresponds to the normal distribution with mean β_h and variance ψ_h , we can obtain more explicit expressions after some algebra, which provide clearer properties:

1. $E(G_h(\mathcal{B})) = N(\beta_{ih} \in \mathcal{B}; \beta_h, \sum_{l=1}^h \gamma_{hl}^2 \psi_l)$,
2. $\text{Var}(G_h(\mathcal{B})) = \frac{N(\beta_{ih} \in \mathcal{B}; \beta_h, \sum_{l=1}^h \gamma_{hl}^2 \psi_l)}{\prod_{l=1}^h (1 + \alpha_l)} \left[1 + \sum_{l=1}^h \alpha_l N(\beta_{il}^{*h} \in \mathcal{B}_{hl}; 0, \gamma_{hl}^2 \psi_l) + \sum_{j \neq k}^h \alpha_j \alpha_k N(\beta_{ik}^{*h} \in \mathcal{B}_{hk}; 0, \gamma_{hj}^2 \psi_j + \gamma_{hk}^2 \psi_k) + \dots - \left(\prod_{l=1}^h (1 + \alpha_l) - \prod_{l=1}^h \alpha_l \right) N(\beta_{ih} \in \mathcal{B}; \beta_h, \sum_{l=1}^h \gamma_{hl}^2 \psi_l) \right]$,
3. $\text{Cov}(G_h(\mathcal{B}), G_m(\mathcal{B})) = \frac{N(\beta_{im} \in \mathcal{B}; \beta_m, \sum_{l=1}^m \gamma_{ml}^2 \psi_l)}{\prod_{l=1}^h (1 + \alpha_l)} \left[1 + \sum_{l=1}^h \alpha_l N(\beta_{il}^{*m} \in \mathcal{B}_{ml}; 0, \gamma_{ml}^2 \psi_l) + \sum_{j \neq k}^h \alpha_j \alpha_k N(\beta_{ik}^{*m} \in \mathcal{B}_{mk}; 0, \gamma_{mj}^2 \psi_j + \gamma_{mk}^2 \psi_k) + \dots - \left(\prod_{l=1}^h (1 + \alpha_l) - \prod_{l=1}^h \alpha_l \right) N(\beta_{im} \in \mathcal{B}; \beta_m, \sum_{l=1}^h \gamma_{ml}^2 \psi_l) \right]$, for $h < m$,

where \mathcal{B}_{ml} denotes the subset corresponding to β_{il}^{*m} with the atoms $\xi_{r,l}^{*m}$ generated from G_{0l}^{*m} , for $r = 1, \dots$ and $l = 1, \dots, h$. More precisely, if $\beta_{im} \in \mathcal{B}$, due to (4), the subset \mathcal{B}_{ml} of β_{il}^{*m} can be defined as the rescaled and shifted subset \mathcal{B} according to γ_{ml} and β_l . One attractive property in this special case is that the random distribution for the elements of β_h is centered at the normal distribution with mean β_h and variance $\sum_{l=1}^h \gamma_{hl}^2 \psi_l$, which are confirmed by (5) and (6). When $\gamma_{hl} = 0$, for $h = 1, \dots, p$ and $l = 1, \dots, h - 1$, which implies that the random effects are uncorrelated and $\beta_{ih} = \beta_{ih}^*$, it is easy to show that $\text{Var}(G_h(\mathcal{B}))$ reduces to $\text{Var}(G_h^*(\mathcal{B}))$. In addition, it is straightforward that $\text{Cov}(G_h(\mathcal{B}), G_m(\mathcal{B})) = 0$ when $\gamma_{ml} = 0$, for $h < m$, $m = 1, \dots, p$ and

$l = 1, \dots, h$.

2.3 Prior Specifications

Following standard convention, we choose a conjugate gamma prior for the residual precision of the model, i.e. $\pi(\sigma^{-2}) \propto \mathcal{G}(c, d)$. To allow β_h to effectively drop out of the model, we choose a mixture prior consisting of a point mass at zero (with probability $\lambda_{1,0h}$) and a normal density:

$$\pi(\beta_h) = \lambda_{1,0h}\delta_0(\beta_h) + (1 - \lambda_{1,0h})\mathcal{N}(\beta_h; \beta_{0h}, \sigma_{0h}^2), \quad (8)$$

where $\delta_0(z)$ denotes the point mass at zero, and $\lambda_{1,0h}$, β_{0h} and σ_{0h}^2 are hyperparameters specified by the investigators. We refer to prior (8) as a zero-inflated normal density, $\mathcal{N}_{\delta_0}(\beta_{0h}, \sigma_{0h}^2)$. The prior probability that the h th predictor of the p candidate predictors is excluded from the fixed effect component is then $\lambda_{1,0h} = \Pr(\beta_h = 0)$. Note that we can accommodate cases in which one or more of the p candidate predictors are only candidates for the random effects component by setting the λ 's for these predictors equal to one.

In the special case as $\psi_h \rightarrow 0$, all the β_{ih}^* are generated from the point mass at β_h and so we have $\beta_{ih}^* = \beta_h$, for $i = 1, \dots, n$. As a convention in the case in which ψ_h is exactly equal to zero, instead of including the individual β_{ih}^* 's in the model, we simply replace these individual-specific parameters with β_h . This convention has no effect on the likelihood, but does impact posterior computation, as will become apparent in Section 3. As a prior for ψ_h , we choose a zero-inflated inverse gamma distribution:

$$\pi(\psi_h) = \lambda_{2,0h}\delta_0(\psi_h) + (1 - \lambda_{2,0h})\mathcal{IG}(\psi_h; c_{0h}, d_{0h}). \quad (9)$$

with $\psi_h = 0$ indexing the model in which we replace the β_{ih}^* s with β_h . We refer to prior (9) as $\mathcal{IG}_{\delta_0}(c_{0h}, d_{0h})$. Thus, the overall prior probability of excluding all the random effects from the model is $\prod_{h=1}^p \lambda_{2,0h}$.

An important point in (4) is that γ_{hl} is only defined when both the h th and the l th random effects are included in the model, implying that $\boldsymbol{\gamma} = (\gamma_{hl} : h = l+1, \dots, p; l = 1, \dots, p-1)'$ depends on $\boldsymbol{\psi} = \text{diag}(\psi_1, \dots, \psi_p)$. If $\psi_h = 0$, we exclude γ_{hl} and γ_{mh} from the model, for $l = 1, \dots, h-1$,

$m = h + 1, \dots, p$, and $h = 1, \dots, p$. Letting γ_{ψ} denote the remaining free elements of γ included in the model, we choose a $N(\gamma_{\psi}; E\gamma_{\psi}, V\gamma_{\psi})$ prior. Note that in cases in which interest focuses on the covariance structure between random effects in the model, this prior can be adapted to include mass at zero

One of the most useful properties of the DP prior is the Pólya urn characterization (Blackwell and MacQueen, 1973), which was exploited by Escobar (1994), MacEachern (1994), and West et al. (1994). Let $\mathbf{S}_h = (S_{1h}, \dots, S_{nh})'$ denote a configuration of $\beta_h^* = (\beta_{1h}, \dots, \beta_{nh})'$ into $k_h \leq n$ distinct values $\theta_h = (\theta_{1h}, \dots, \theta_{k_h h})'$, with $S_{ih} = s$ if $\beta_{ih}^* = \theta_{sh}$ denoting that subject i belongs to cluster s for the h th random effect, for $h = 1, \dots, p$ and $s = 1, \dots, k_h$. Then the conditional Pólya urn prior of β_{ih}^* , $h = 1, \dots, p$, is derived by

$$\pi(\beta_{ih}^* | \theta_h^{(i)}, k_h^{(i)}, \mathbf{S}_h^{(i)}) \propto \frac{\alpha_h}{\alpha_h + n - 1} G_{0h} + \sum_{s=1}^{k_{-i,h}} \frac{r_{s,h}^{(i)}}{\alpha_h + n - 1} \delta_{\theta_{s,h}^{(i)}}(\cdot), \quad i = 1, \dots, n, \quad (10)$$

where $\mathbf{S}_h^{(i)}$ denotes the configuration of $\beta_h^{*(i)} = (\beta_{1h}^*, \dots, \beta_{i-1,h}^*, \beta_{i+1,h}^*, \dots, \beta_{nh}^*)'$ with $k_h^{(i)}$ distinct values $\theta_h^{(i)} = \theta_h \setminus \beta_{ih}^*$, $r_{s,h}^{(i)}$ denotes the frequency of the unique value $\theta_{s,h}^{(i)}$ occurring among $\beta_h^{*(i)}$, and $\delta_{\theta}(\cdot)$ denotes the distribution degenerate at θ . Let $\{q_{i,s,h}(\alpha_h)\}_{s=0}^{k_h^{(i)}}$ denote the set of probabilities of the different components in (10) which sum to one. Obviously, $k_h^{(i)}$ equals k_h if $r_{s,h}^{(i)} \geq 1$, and $k_h - 1$ otherwise.

The treatment of the concentration parameter α_h ($h = 1, \dots, p$) of the underlying Dirichlet process is important, since this hyperparameter can impact inference. Escobar and West (1995) used a gamma prior while McAuliffe, Blei and Jordan (2004) proposed an empirical Bayes method to estimate α . In this article, we adopt a gamma prior $\mathcal{G}(a, b)$ for $\pi(\alpha_h)$.

For an additional subject $n + 1$, the predictive distribution of $\beta_{n+1,h}^*$, which is the posterior mean $E(G_h^* | \theta_h, k_h, \mathbf{S}_h)$, can be expressed as

$$\pi(\beta_{n+1,h}^* | \theta_h, k_h, \mathbf{S}_h) \propto \frac{\alpha_h}{\alpha_h + n} G_{0h} + \frac{1}{\alpha_h + n} \sum_{s=1}^{k_h} r_{sh} \delta_{\theta_{sh}}(\cdot). \quad (11)$$

There is no closed form for the predictive density of $\beta_{n+1,h}$. However, the predicted $\beta_{n+1,h}$ can be calculated based on β_{n+1}^* and γ . More explicitly, when G_{0l} denotes the normal distribution, each

element of the second term in (4) follows the distribution

$$\frac{\alpha_l}{\alpha_l + n} \mathbf{N}(\cdot; 0, \gamma_{hl}^2 \psi_l) + \frac{1}{\alpha_l + n} \sum_{s=1}^{k_l} r_{sl} \delta_{\gamma_{hl}(\theta_{sl} - \beta_l)}(\cdot).$$

3. POSTERIOR COMPUTATIONS

We outline a Gibbs sampler for posterior computation, providing details in Appendix B. After specifying initial values for the parameters and latent variables, our proposed MCMC algorithm proceeds as follows:

1. Update β_{ih}^* , for $i = 1, \dots, n$ and $h = 1, \dots, p$, from the full conditional posterior distribution of β_{ih}^* derived using the Pólya urn scheme (Escobar, 1994; MacEachern, 1994; West et al., 1994) given the data and current values of $\boldsymbol{\theta}_h^{(i)}$, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\mathbf{S}_h^{(i)}$ and $\boldsymbol{\sigma}$.
2. Update α_h , for $h = 1, \dots, p$, from the full conditional posterior distribution by updating latent parameter ϕ_h based on the most recently updated α_h followed by updating α_h based on the most recent k_h and the just updated ϕ_h .
3. Update $\boldsymbol{\gamma}_\psi$ from the full conditional posterior distribution given the data and the current values of β_i^* and $\boldsymbol{\beta}$.
4. Update β_h , for $h = 1, \dots, p$, from the full conditional posterior distribution proportional to

$$\hat{\lambda}_{1,h} \delta_0(\beta_h) + (1 - \hat{\lambda}_{1,h}) \mathbf{N}(\beta_h; \hat{E}_h, \hat{V}_h) \quad (12)$$

given the data and the current values of \mathbf{S}_h , $\boldsymbol{\theta}_h$, $\boldsymbol{\beta}_{-h}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\sigma}$. With probability $\hat{\lambda}_{1,h}$, $\beta_h = 0$, and with probability $1 - \hat{\lambda}_{1,h}$, β_h is drawn from $\mathbf{N}(\beta_h; \hat{E}_h, \hat{V}_h)$.

5. Update ψ_h , for $h = 1, \dots, p$, from the full conditional posterior distribution proportional to

$$\hat{\lambda}_{2,h} \delta_0(\psi_h) + (1 - \hat{\lambda}_{2,h}) \mathcal{IG}(\psi_h; \hat{c}_h, \hat{d}_h) \quad (13)$$

given the data and the current values of \mathbf{S}_h , $\boldsymbol{\theta}_h$, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\sigma}$. With probability $\hat{\lambda}_{2,h}$, $\psi_h = 0$, and with probability $1 - \hat{\lambda}_{2,h}$, ψ_h is drawn from $\mathcal{IG}(\psi_h; \hat{c}_h, \hat{d}_h)$.

6. Update σ^{-2} straightforwardly from the full conditional posterior distribution given the data and the current values of β_i^* and β .

This algorithm allows the candidate models to adaptively move around the model space. After apparent convergence of the samples for the parameters and latent variables, the posterior densities of the parameters and posterior probabilities for each of the different submodels can be straightforwardly calculated.

REMARK In the Gibbs sampling under DP with uncertainty, if the h th random effect is excluded from the model, i.e. $\psi_h = 0$, β_{ih}^* will be drawn from point masses at β_h . In this case, there is no need to update the clusters, the corresponding configuration, and distinct values through Pólya urn scheme. Thus, we only update the set of values \mathbf{S}_h , k_h and $\boldsymbol{\theta}_h$ when the h th random effect is included in the model. Instead of sampling β_{ih}^* directly, we update as follows:

1. Conditional on current $\boldsymbol{\theta}_h$ and \mathbf{S}_h , a new configuration can be generated by sequentially sampling from the posterior probabilities:

$$Pr(S_{ih} = s | \boldsymbol{\theta}_h^{(i)}, \mathbf{S}_h^{(i)}, k_h^{(i)}) = q_{i,s,h}^*, \quad s = 0, 1, \dots, k_h^{(i)}.$$

When $S_{ih} = 0$, a new β_{ih}^* is generated from $G_{i,0,h}$. A new cluster is constructed because the new value is different from the values in existing clusters.

2. Conditional on \mathbf{S}_h and k_h , a new set of parameters $\boldsymbol{\theta}_h$ can be generated by sampling each new θ_{sh} from the full conditional posterior distribution:

$$\prod_{i \in I_{sh}} \pi(\mathbf{y}_i | \theta_{sh}, \mathbf{X}_i) G_{0h}(\theta_{sh}),$$

where $I_{sh} = \{i : S_{ih} = s\}$. By updating the grouped β_{ih}^* , i.e. θ_{sh} , we can speed up the convergence for the Gibbs sampler over the entire parameter space (Bush and MacEachern, 1996).

4. SIMULATION STUDY

In order to evaluate the computational performance of our proposed Gibbs sampling algorithm and assess the results, we conducted a simulation study in which we generated 200 subjects, each with 5 repeated measurements. There are four covariates contained in $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ij4})'$, where x_{ij1} is fixed as one, and the rest are generated from a uniform distribution. We chose $\beta_{i1}^* \sim N(0, 4)$, $\beta_{i2}^* = 2$, $\beta_{i3}^* \sim 0.4N(1.5, 1) + 0.6N(4, 1)$, $\beta_{i4}^* \sim N(6, 1)$, which implies $\beta = (0, 2, 3, 6)'$. We chose

$$\Sigma = \begin{pmatrix} 4 & 0 & 2.40 & 6 \\ 0 & 0 & 0 & 0 \\ 2.40 & 0 & 3.94 & 5.85 \\ 6 & 0 & 5.85 & 12 \end{pmatrix},$$

which implies $\gamma = (0, 0.6, 0, 1.5, 0, 0.9)'$, the diagonal elements of the covariance matrix of random effects are $(d_{11}, d_{22}, d_{33}, d_{44}) = (4, 0, 3.94, 12)$, and the random effect correlations are $(\rho_{21}, \rho_{31}, \rho_{32}, \rho_{41}, \rho_{42}, \rho_{43}) = (0, 0.60, 0, 0.86, 0, 0.85)$. σ^{-2} is chosen as 1. The response variable y_{ij} is generated from (3) and (4).

We chose the prior distribution for σ^{-2} as $\mathcal{G}(0.05, 0.05)$. The prior distributions for the elements of β are chosen as $N_{\delta_0}(0, 10)$. To study the effect of the prior probability that $\beta_h = 0$, we chose $\lambda_{1,0h} = 0.2, 0.5$ and 0.8 , respectively. The prior distributions for the elements of $\gamma\psi$ are independent $N(0, 2)$. We also chose the mixture prior distributions for the elements of ψ as independent $\mathcal{IG}_{\delta_0}(0.1, 0.1)$. Similarly, we chose three different values for $\lambda_{2,0h}$. The prior distribution for α_h is chosen as $\mathcal{G}(1, 1)$, essentially identical results were obtained for $\mathcal{G}(2, 4)$.

We ran the Gibbs sampling algorithm described in Section 3 for 30,000 iterations after an 8,000 burn-in. Diagnostic tests were carried out by using Geweke (1992) and Raftery and Lewis (1992), which showed rapid convergence and efficient mixing. A sample of size 6000 was obtained by thinning the MCMC chain by a factor of 5. We calculated posterior probabilities for the possible submodels, estimated posterior means, and 95% credible intervals for each of the parameters. To obtain proper samples of α_h , we ran 20 sub-iterations within each iteration (Escobar and West, 1995).

To compare the results of our nonparametric mixed effect analysis (NME) with other approaches, we also fit a frequentist linear mixed effects model (LME) and applied Chen and Dunson's method (CDM) (2003) (in this article, we modify Chen and Dunson by adding fixed effect

selection). Table 1 shows the true and estimated values of the parameters for the three different methods. On average, the NME estimates were slightly closer to the truth than the other estimates. Table 2 presents the posterior probabilities of all submodels selected by our nonparametric mixed effect method and Chen and Dunson’s method. We also show the corresponding deviance information criterion (DIC) (Spiegelhalter et al, 2002), obtained by running separate linear mixed effects model analyses for each model in the list. Although each method chooses the true model as the best model, our NME approach assigns higher posterior probability to the true model than CDM, while the DIC is not calibrated in terms of posterior probability. The DIC approach also requires a separate MCMC analysis for each model, which may not be feasible in higher dimensions. Interestingly, mixing and computational efficiency was substantially better for NME than for CDM.

Figure 1 shows the posterior densities of the latent parameters β_i^* based on our nonparametric mixed effect model, and the corresponding true densities. Figure 2 presents the posterior densities of β_i and their corresponding true densities. It appears that our nonparametric mixed effect model successfully captured the right densities of β_i^* and β_i .

Sensitivity of the results to the prior specification was assessed by repeating the analyses with the following different hyperparameters: (a) priors with variance $/2$; (b) priors with variance $\times 2$; (c) priors with moderately different means. Although we do not show details, inferences for all models are robust to the prior specification. The ranges in the tables illustrate this robustness.

5. APPLICATION

We illustrate our approach through analysis of data from an international validation study of the rat uterotrophic bioassay, a new animal model designed to detect *in vivo* estrogenic responses to test chemicals (Kanno et al., 2001). The data were collected from 19 participating laboratories in 8 nations and consisted of 2681 female rats. One or more out of four protocols were chosen by each laboratory. Two of the protocols used the immature female rat model, with administration of doses by oral gavage and subcutaneous injection for 3 days at 24-hr intervals followed by humane killing 24 hours after the last administration, respectively. The other two protocols used the adult

ovariectomized (OVX) rat model, both with administration by subcutaneous injection but for 3 and 7 days, respectively. Each of the protocols contained 11 groups of 6 animals within which, except for an untreated control group and a vehicle control group, a series of seven doses of ethinyl estradiol (EE), a known estrogen agonist, in half-log steps from 0.01 to 10 $\mu\text{g}/\text{kg}/\text{day}$ was used for 7 groups, while the rest two groups contained two ZM (an estrogen antagonist) doses combined with a fixed EE dose. The response of interest was blotted uterus weight. Using simple parametric models, Kanno et al. (2001) reported that all laboratories and all protocols had significant dose response trends in uterus weight with EE exposure. They were unable to formally assess heterogeneity among laboratories using their approach.

Our analysis focuses on identifying predictors of uterus weight. Predictors included in the fixed effect component have an average effect on mean weight, while predictors included in the random effect component vary in their effects across laboratories. It is particularly interesting to assess whether laboratories vary in the slope of the dose response for EE, suggesting differential sensitivity of the assay. In the presence of heterogeneity, the distribution of the random slope is of substantial interest, and we wish to investigate how labs vary. The preliminary analyses showed the goodness of fit of the linearity assumption. Let $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6})'$ denote the vector of candidate predictors with $x_{ij1} = 1$, x_{ij2} , x_{ij3} and x_{ij4} denoting 0/1 indicators of the protocols for the immature rat model, the adult rat model with 3 days and 7 days, respectively, x_{ij5} denoting dose of EE, and x_{ij6} denoting dose of ZM for the j th rat in laboratory i . The response variable, y_{ij} , is the log-transformed blotted uterus weight.

The prior distributions for the elements of $\boldsymbol{\beta}$ are chosen as $N_{\delta_0}(0, 10)$ where $\lambda_{1,0h} = 0.2, 0.5$ and 0.8 , respectively. The prior distributions for the free elements of $\boldsymbol{\gamma}$ are independent $N(0, 1)$. The mixture prior distributions of the elements of $\boldsymbol{\psi}$ are chosen as independent $\mathcal{IG}_{\delta_0}(0.1, 0.1)$ with $\lambda_{2,0h} = 0.5$. We also chose $\mathcal{G}(1, 1)$ and $\mathcal{G}(0.05, 0.05)$ as the prior distribution for α_h and σ^{-2} , respectively.

We ran the Gibbs sampling algorithm described in Section 3 for 100,000 iterations after a 10,000 burn-in. The diagnostic tests showed rapid convergence and efficient mixing. A sample of

size 5,000 was obtained by thinning the MCMC chain by a factor of 20. Posterior probabilities for the possible submodels, estimated posterior means, and 95% credible intervals for each of the parameters are calculated thereafter.

Sensitivity of the results to the prior specification was assessed by repeating the analyses with the following different hyperparameters: (a) priors with variance $/2$; (b) priors with variance $\times 2$; (c) priors with moderately different means; (d) $\lambda_{1,0h}, \lambda_{2,0h} = 0.2, 0.5, 0.8$. Not surprisingly, the choice of prior for ψ impacts the rates of convergence and mixing of the Gibbs sampler, with mixing slower for more diffuse priors. This is consistent with behavior for normal linear mixed models with inverse-gamma priors on the variance components. We recommend choosing small to moderate variance, motivated by substantive consideration.

Table 3 presents the posterior probabilities of top ten submodels selected by our method. The model including all fixed effects predictors except the protocol for the juvenile rat model, and the predictors for the random intercept and the two protocols for the adult rat model is the highest posterior-probability model. The posterior means for fixed effects $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and β_6 are 3.52, 0.05, 1.25, 1.34, 0.14 and -0.49 , respectively. These results agree with previous analyses of positive relationship between uterine weights and EE doses used, and of negative relationship between uterus weights and ZM doses used. The corresponding 95% credible intervals for $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and β_6 are (3.11, 4.05), (0.01, 0.10), (1.11, 1.33), (1.15, 1.56), (0.13, 0.16) and $(-0.60, -0.39)$, respectively. The means and 95% confidence intervals for $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and β_6 from the frequentist linear mixed effect model are $3.62_{(3.52, 3.72)}$, $0.07_{(0.02, 0.12)}$, $1.16_{(1.03, 1.29)}$, $1.18_{(1.04, 1.32)}$, $0.14_{(0.13, 0.15)}$ and $-0.47_{(-0.57, -0.37)}$, respectively. In addition, our nonparametric random effect method suggests to include the random intercept and random slopes for the protocols for the adult rat, implying heterogeneity across the laboratories. Interestingly, the model with the highest posterior probability suggests that the EE slope and the ZM slope did not vary across laboratories. In contrast, the frequentist linear mixed effect model suggests that the random effects for the juvenile protocol (variance is $2e-3$) and EE doses (variance is $1e-4$) were excluded, while random intercept effects (variance is 0.04), random effects for two adult protocols (variances are 0.05 and 0.06, respectively)

and ZM doses (variance is 0.03) weakly exist.

Figure 3 shows the posterior densities of the intercept and slopes for protocols, EE doses and ZM doses. There is an evidence of the random intercept following a multi-modal distribution. This result provides not only the existence of heterogeneity among the laboratories but also that there might be multiple groups or clusters across the laboratories. Interestingly, in terms of the random intercept, the participating laboratories seem to be grouped into three clusters with the highest probability (0.268). Furthermore, based on the configuration, the members of the three clusters can be predicted. The 9 laboratories from France, Germany, Netherlands and UK have the highest probability (0.438) of falling in cluster 1, the 4 laboratories from Korea and USA fall in cluster 2 with the highest probability (0.353), while the 6 laboratories from Japan belong to cluster 3 with the highest probability (0.396).

6. DISCUSSION

In this article, we propose a Bayesian approach to the problem of nonparametric random effects models where both the predictors to be included and distributions of their random effects are unknown. Due to a carefully-developed computational algorithm, which incorporates centering parameterization and utilizes conjugacy, the approach is very efficient and straightforward to implement. In fact, we found the performance to be much better than for Chen and Dunson (2003) parametric approach, which is subject to slow-mixing. Using data augmentation, generalization to allow categorical outcomes are straightforward.

The proposed approach characterizes the unknown random effects distributions using a weighted convolution of independent DP-distributions, with a mixture structure chosen for the base measure. Although motivated by the random effects selection problem, the proposed approach provides a general strategy for modeling of dependency in related unknown distributions. We provide some initial theoretical results for weighted convolutions of DPs, but detailed consideration of general properties provides an interesting area of future research.

ACKNOWLEDGEMENTS

The authors thank Beth Gladen and Grace Kissling for valuable comments and suggestions. This research was supported by the Intramural Research Program of the NIH, and NIEHS.

REFERENCES

- Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, 47, 639-653.
- Albert, J. and Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, 92, 916-925.
- Antoniak, C.E. (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2, 1152-1174.
- Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, 98, 224-235.
- Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1, 353-355.
- Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83, 275-285.
- Cai, B. and Dunson, D.B. (2005). Bayesian covariance selection in generalized linear mixed models. *ISDS Discussion Paper*, Institute of Statistics and Decision Sciences, Duke University.
- Chen, J.L., Zhang, D.W. and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distributions. *Biostatistics*, 3, 347-360.
- Chen, Z. and Dunson, D.B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59, 762-769.
- Clyde, M. and George, E.I. (2004). Model uncertainty. *Statistical Science*, 19, 81-94.

- Crainiceanu, C.M. and Ruppert, D. (2004). Restricted likelihood ratio tests in nonparametric longitudinal models. *Statistica Sinica*, 14, 713-729.
- Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 268-277.
- Escobar, M.D. and 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 non-parametric problems. *The Annals of Statistics*, 1, 209-230.
- Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995). Efficient parameterizations for normal linear mixed models. *Biometrika*, 82, 479-488.
- George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339-373.
- Geweke, J. (1992). Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments. *Bayesian Statistics 4*, (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds), Oxford University Press, Oxford. 169-193.
- Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics 5*, (J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith, eds), Oxford University Press, Oxford. 609-620.
- Ghidey, W., Lesaffre, E. and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60, 945-953.
- Hall, D.B. and Praestgaard, J.T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, 88, 739-751.
- 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.
- Kanno, J., Onyon, L., Haseman, J., Fenner-Crisp, P., Ashby, J., and Owens, W. (2001). The OECD program to validate the rat uterotrophic bioassay to screen compounds for *in vivo* estrogenic responses: phase 1. *Environmental Health Perspectives*, 109(8), 785-794.
- Kleinman, K.P. and Ibrahim, J.G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54, 921-938.
- Lai, T.L. and Shih, M.C. (2003). Nonparametric estimation in nonlinear mixed effects models. *Biometrika*, 90, 1-13.
- Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Lin, X.H. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84, 309-326.
- Lopes, H.F., Müller, P. and Rosner, G.L. (2003). Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics*, 59, 66-75.
- MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23, 723-741.
- McAuliffe, J.D., Blei, D.M., and Jordan, M.I. (2004). Nonparametric empirical Bayes for the Dirichlet process mixture model. *UC Berkeley Statistics technical report 675*.
- Müller, P., Rosner, G.L., De Iorio, M. and MacEachern, S. (2005). A nonparametric Bayesian model for inference in related longitudinal studies. *Applied Statistics*, 54, 611-626.
- Pauler, D.K., Wakefield, J.C., and Kass, R.E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association*, 94, 1242-1253.

- Raftery, A.E. and Lewis, S. (1992). How Many Iterations in the Gibbs Sampler? *Bayesian Statistics 4* (Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., eds), Oxford: Oxford University Press, 763-773.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.
- Sinharay, S. and Stern, H.S. (2001). Bayes factors for variance component testing in generalized linear mixed models. *Bayesian Methods with Applications to Science, Policy and Official Statistics (ISBA 2000 Proceedings)*, 507-516.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Linde, A.V.D. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of Royal Statistical Society B*, 64, 1-34.
- Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59, 254-262.
- West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D.V.Lindley*, Ed. Smith, A.F.M. and Freeman, P., 363-386. New York: Wiley.
- Zhang, D.W. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57, 795-802.
- Zhu, Z.Y. and Fung, W.K. (2004). Variance component testing in semiparametric mixed models. *Journal of Multivariate Analysis*, 91, 107-118.

APPENDIX A

Proof of Theorem 1:

Since $\beta_{i1} = \beta_{i1}^*$, the random distribution of β_{i1} , G_1 , can be expressed as a stick-breaking formulation,

$$G_1 = G_1^* = \sum_{s=1}^{\infty} p_{s,1} \delta_{\xi_{s,1}}, \quad p_{s,1} = V_{s,1} \prod_{l=1}^{s-1} (1 - V_{l,1}) \quad \text{with} \quad V_{s,1} \stackrel{iid}{\sim} \text{beta}(1, \alpha_1), \quad \xi_{s,1} \stackrel{iid}{\sim} G_{01}.$$

Similarly, G_2^* can be expressed as

$$G_2^* = \sum_{t=1}^{\infty} p_{t,2} \delta_{\xi_{t,2}}, \quad p_{t,2} = V_{t,2} \prod_{l=1}^{t-1} (1 - V_{l,2}) \quad \text{with} \quad V_{t,2} \stackrel{iid}{\sim} \text{beta}(1, \alpha_2), \quad \xi_{t,2} \stackrel{iid}{\sim} G_{02}.$$

From (4), we have $\beta_{i2} = \beta_{i2}^* + \gamma_{21}(\beta_{i1}^* - \beta_1)$. Clearly, the random distribution of $\gamma_{21}(\beta_{i1}^* - \beta_1)$ is still from the DP with the same α_1 due to its independence. We define the random distribution for $\gamma_{21}(\beta_{i1}^* - \beta_1)$ as $G_1^{**} = \sum_{s=1}^{\infty} p_{s,1} \delta_{\xi_{s,1}^{*2}}$, $\xi_{s,1}^{*2} \stackrel{iid}{\sim} G_{01}^{*2}$, where G_{01}^{*2} is the base measure of the $\text{DP}(\alpha_1 G_{01}^{*2})$. Since β_{i1}^* and β_{i2}^* are independent, the distribution of β_{i2} , G_2 , can be expressed as the convolution of G_1^{**} and G_2^* , i.e. $G_1^{**} \otimes G_2^*$. For any Borel set $\mathcal{B} \subseteq \mathcal{R}$, by mathematical induction, we have

$$G_h(\mathcal{B}) = G_1^{**} \otimes \cdots \otimes G_{h-1}^{**} \otimes G_h^*(\mathcal{B}) = \otimes_{l=1}^{h-1} G_l^{**} \otimes G_h^*(\mathcal{B}), \quad h = 1, \dots, p,$$

where $G_l^{**} = \sum_{s=1}^{\infty} p_{s,l} \delta_{\xi_{s,l}^{*h}}$, $\xi_{s,l}^{*h} \stackrel{iid}{\sim} G_{0l}^{*h}$, where G_{0l}^{*h} is the base measure of the $\text{DP}(\alpha_l G_{0l}^{*h})$ for $\gamma_{hl}(\beta_{il}^* - \beta_l)$, for $l = 1, \dots, h-1$.

It is obvious that $\text{E}(G_h^*(\mathcal{B})) = G_{0h}(\mathcal{B})$, and for $h = 1, \dots, p$,

$$\begin{aligned} \text{Var}(G_h^*(\mathcal{B})) &= \text{E}\left(\sum_{s=1}^{\infty} p_{s,h}^2 \text{Var}(\delta_{\xi_{s,h}}(\mathcal{B}))\right) \\ &= \text{E}\left(\sum_{s=1}^{\infty} p_{s,h}^2\right) (G_{0h}(\mathcal{B}) - G_{0h}^2(\mathcal{B})) \\ &= (G_{0h}(\mathcal{B}) - G_{0h}^2(\mathcal{B})) \sum_{s=1}^{\infty} \text{E}(V_{s,h}^2) \prod_{l=1}^{s-1} (1 - V_{l,h})^2 \\ &= (G_{0h}(\mathcal{B}) - G_{0h}^2(\mathcal{B})) \sum_{s=1}^{\infty} \text{E}(V_{s,h}^2) \prod_{l=1}^{s-1} (1 - 2\text{E}(V_{l,h}) + 2\text{E}(V_{l,h}^2)) \\ &= (G_{0h}(\mathcal{B}) - G_{0h}^2(\mathcal{B})) \sum_{s=1}^{\infty} \frac{2}{(1 + \alpha_h)(2 + \alpha_h)} \left(\frac{\alpha_h}{2 + \alpha_h}\right)^{s-1} \\ &= \frac{G_{0h}(\mathcal{B})(1 - G_{0h}(\mathcal{B}))}{1 + \alpha_h}, \end{aligned}$$

where $p_{s,h} = V_{s,h} \prod_{l=1}^{s-1} (1 - V_{l,h})$ with $V_{s,h} \stackrel{iid}{\sim} \text{beta}(1, \alpha_h)$.

When $h = 2$, we have

$$\begin{aligned} \text{E}(G_2(\mathcal{B})) &= \text{E}\left(\int G_1^{**}(\beta_{i2} - \beta_{i2}^*) G_2^*(\beta_{i2}^*) d\beta_{i2}^*(\mathcal{B})\right) \\ &= \text{E}\left(\int \sum_{s=1}^{\infty} p_{s,1} \delta_{\xi_{s,1}^{*2}} \sum_{t=1}^{\infty} p_{t,2} \delta_{\xi_{t,2}} d\beta_{i2}^*(\mathcal{B})\right) \\ &= \text{E}\left(\sum_{s=1}^{\infty} \sum_{t=1}^{\infty} p_{s,1} p_{t,2}\right) \text{E}\left(\int \delta_{\xi_{s,1}^{*2}} \delta_{\xi_{t,2}} d\beta_{i2}^*(\mathcal{B})\right) \end{aligned}$$

$$\begin{aligned}
&= G_{01}^{*2} \otimes G_{02}(\mathcal{B}) \mathbb{E} \left(\sum_{s=1}^{\infty} \sum_{t=1}^{\infty} p_{s,1} p_{t,2} \right) \\
&= G_{01}^{*2} \otimes G_{02}(\mathcal{B}).
\end{aligned}$$

By mathematical induction, for $h = 1, \dots, p$, we have

$$\mathbb{E}(G_h(\mathcal{B})) = \otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B}).$$

The variance of the random distribution of β_{i2} can be expressed as

$$\begin{aligned}
\text{Var}(G_2(\mathcal{B})) &= \mathbb{E}(G_2^2(\mathcal{B})) - \mathbb{E}^2(G_2(\mathcal{B})) \\
&= \mathbb{E} \left(\left(\int \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} p_{s,1} p_{t,2} \delta_{\xi_{s,1}^{*2}} \delta_{\xi_{t,2}} d\beta_{i2}^*(\mathcal{B}) \right)^2 \right) - (G_{01}^{*2} \otimes G_{02}(\mathcal{B}))^2 \\
&= \mathbb{E} \left(\left(\sum_{s=1}^{\infty} \sum_{t=1}^{\infty} p_{s,1} p_{t,2} \int \delta_{\xi_{s,1}^{*2}} \delta_{\xi_{t,2}} d\beta_{i2}^*(\mathcal{B}) \right)^2 \right) - (G_{01}^{*2} \otimes G_{02}(\mathcal{B}))^2 \\
&= \mathbb{E} \left(\sum_{s=1}^{\infty} \sum_{t=1}^{\infty} p_{s,1}^2 p_{t,2}^2 \right) G_{01}^{*2} \otimes G_{02}(\mathcal{B}) + \left(1 - \mathbb{E} \left(\sum_{s=1}^{\infty} p_{s,1}^2 \right) \right) \mathbb{E} \left(\sum_{t=1}^{\infty} p_{t,2}^2 \right) G_{01}^{*2}(\mathcal{B}) G_{01}^{*2} \otimes G_{02}(\mathcal{B}) \\
&\quad + \mathbb{E} \left(\sum_{s=1}^{\infty} p_{s,1}^2 \right) \left(1 - \mathbb{E} \left(\sum_{t=1}^{\infty} p_{t,2}^2 \right) \right) G_{02}(\mathcal{B}) G_{01}^{*2} \otimes G_{02}(\mathcal{B}) \\
&\quad + \left(1 - \mathbb{E} \left(\sum_{s=1}^{\infty} p_{s,1}^2 \right) \right) \left(1 - \mathbb{E} \left(\sum_{t=1}^{\infty} p_{t,2}^2 \right) \right) (G_{01}^{*2} \otimes G_{02}(\mathcal{B}))^2 - (G_{01}^{*2} \otimes G_{02}(\mathcal{B}))^2 \\
&= \frac{G_{01}^{*2} \otimes G_{02}(\mathcal{B})}{\prod_{l=1}^2 (1 + \alpha_l)} \left(1 + \alpha_1 G_{01}^{*2}(\mathcal{B}) + \alpha_2 G_{02}(\mathcal{B}) - (1 + \alpha_1 + \alpha_2) G_{01}^{*2} \otimes G_{02}(\mathcal{B}) \right)
\end{aligned}$$

Again, by mathematical induction, we have

$$\begin{aligned}
\text{Var}(G_h(\mathcal{B})) &= \frac{\otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B})}{\prod_{l=1}^h (1 + \alpha_l)} \left[1 + \sum_{l=1}^h \alpha_l G_{0l}^{*h}(\mathcal{B}) + \sum_{i \neq j}^h \alpha_i \alpha_j G_{0i}^{*h} \otimes G_{0j}^{*h}(\mathcal{B}) + \dots \right. \\
&\quad \left. - \left(\prod_{l=1}^h (1 + \alpha_l) - \prod_{l=1}^h \alpha_l \right) \otimes_{l=1}^{h-1} G_{0l}^{*h} \otimes G_{0h}(\mathcal{B}) \right].
\end{aligned}$$

To calculate the covariance of two random distributions, G_h and G_m , $h < m$, we rewrite expression

(4) as

$$\beta_{ih} - \beta_h = \sum_{l=1}^h \gamma_{hl} (\beta_{il}^* - \beta_l).$$

It is clear that $\text{Cov}(G_h(\mathcal{B}), G_m(\mathcal{B})) = \text{Cov}(G_h^{\Delta}(\mathcal{B}), G_m^{\Delta}(\mathcal{B}))$, where G_h^{Δ} denotes the random distribution for $\beta_{ih} - \beta_h$. Let $\mathcal{B}_l^{hm} = \mathcal{B}_{hl} \cap \mathcal{B}_{ml}$. We note that

$$\delta_{\xi_{r,l}^{*h}} \delta_{\xi_{s,l}^{*m}} = \begin{cases} 1 & \text{if } \xi_{r,l}^{*h}, \xi_{s,l}^{*m} \in \mathcal{B}_l^{hm} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathbb{E}(\delta_{\xi_{r,l}^{*h}}(\mathcal{B}_l^{hm})\delta_{\xi_{s,l}^{*m}}(\mathcal{B}_l^{hm})) = \begin{cases} G_{0l}^{*m}(\mathcal{B}_l^{hm}) & \text{if } r = s \\ G_{0l}^{*h}(\mathcal{B}_l^{hm})G_{0l}^{*m}(\mathcal{B}_l^{hm}) & \text{otherwise.} \end{cases}$$

We first consider $\text{Cov}(G_1^\Delta(\mathcal{B}), G_2^\Delta(\mathcal{B})) = \mathbb{E}(G_1^\Delta(\mathcal{B})G_2^\Delta(\mathcal{B})) - \mathbb{E}(G_1^\Delta(\mathcal{B}))\mathbb{E}(G_2^\Delta(\mathcal{B}))$, where

$$\begin{aligned} \mathbb{E}(G_1^\Delta(\mathcal{B})G_2^\Delta(\mathcal{B})) &= \mathbb{E}\left(\sum_{r=1}^{\infty} p_{r,1}\delta_{\xi_{r,1}^{*1}} \int \sum_{s=1}^{\infty} p_{s,1}\delta_{\xi_{s,1}^{*2}} \sum_{t=1}^{\infty} p_{t,2}\delta_{\xi_{t,2}^{*2}} d\beta_{i2}^*(\mathcal{B})\right) \\ &= \mathbb{E}\left(\int \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} p_{r,1}p_{s,1}\delta_{\xi_{r,1}^{*1}} \delta_{\xi_{s,1}^{*2}} \sum_{t=1}^{\infty} p_{t,2}\delta_{\xi_{t,2}^{*2}} d\beta_{i2}^*(\mathcal{B})\right) \\ &= \mathbb{E}\left(\sum_{r=1}^{\infty} \sum_{t=1}^{\infty} p_{r,1}^2 p_{t,2}\right) \mathbb{E}\left(\int \delta_{\xi_{r,1}^{*1}} \delta_{\xi_{r,1}^{*2}} \delta_{\xi_{t,2}^{*2}} d\beta_{i2}^*(\mathcal{B})\right) \\ &\quad + \mathbb{E}\left(\sum_{r \neq s}^{\infty} \sum_{t=1}^{\infty} p_{r,1}p_{s,1}p_{t,2}\right) \mathbb{E}\left(\int \delta_{\xi_{r,1}^{*1}} \delta_{\xi_{s,1}^{*2}} \delta_{\xi_{t,2}^{*2}} d\beta_{i2}^*(\mathcal{B})\right) \\ &= \frac{1}{1+\alpha_1} G_{01}^{*2} \otimes G_{02}^{*2}(\mathcal{B}) + \left(1 - \frac{1}{1+\alpha_1}\right) G_{01}^{*2}(\mathcal{B})G_{01}^{*2} \otimes G_{02}^{*2}(\mathcal{B}). \end{aligned}$$

Thus,

$$\text{Cov}(G_1(\mathcal{B}), G_2(\mathcal{B})) = \frac{1}{1+\alpha_1} (1 - G_{01}^{*2}(\mathcal{B}))G_{01}^{*2} \otimes G_{02}^{*2}(\mathcal{B}).$$

By mathematical induction, we have

$$\begin{aligned} \text{Cov}(G_h(\mathcal{B}), G_m(\mathcal{B})) &= \frac{\otimes_{l=1}^m G_{0l}^{*m}(\mathcal{B})}{\prod_{l=1}^h (1+\alpha_l)} \left[1 + \sum_{l=1}^h \alpha_l G_{0l}^{*m}(\mathcal{B}) + \sum_{i \neq j}^h \alpha_i \alpha_j G_{0i}^{*m} \otimes G_{0j}^{*m}(\mathcal{B}) + \dots \right. \\ &\quad \left. - \left(\prod_{l=1}^h (1+\alpha_l) - \prod_{l=1}^h \alpha_l \right) \otimes_{l=1}^h G_{0l}^{*m}(\mathcal{B}) \right]. \end{aligned}$$

APPENDIX B

Full conditional distributions in Section 3

- Let $\mathbf{y}_i^{(1)} = \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\Gamma}^* \boldsymbol{\beta} - \mathbf{X}_i \boldsymbol{\Gamma}_{-h} \boldsymbol{\beta}_{i,-h}^*$, $\boldsymbol{\Gamma}_h$ denote the h th column of $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_{-h}$ denote the submatrix of $\boldsymbol{\Gamma}$ excluding the h th column, and $\boldsymbol{\beta}_{i,-h}^*$ denote the subvector of $\boldsymbol{\beta}_i^*$ with β_{ih} excluded. With probability proportional to

$$r_{s,h}^{(i)} \sigma^{-n_i} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y}_i^{(1)} - \mathbf{X}_i \boldsymbol{\Gamma}_h \boldsymbol{\theta}_{s,h}^{(i)})' (\mathbf{y}_i^{(1)} - \mathbf{X}_i \boldsymbol{\Gamma}_{(h)} \boldsymbol{\theta}_{s,h}^{(i)})\right),$$

we choose β_{ih}^* from degenerate distribution $\delta_{\theta_{s,h}^{(i)}}$, which means that we have $\beta_{ih}^* = \theta_{s,h}^{(i)}$. In addition, with probability proportional to

$$\alpha_h \Sigma_{ih}^{\frac{1}{2}} \psi_h^{-\frac{1}{2}} \sigma^{-n_i} \exp\left(-\frac{1}{2}(\psi_h^{-1} \beta_h^2 + \sigma^{-2} \mathbf{y}_i^{(1)'} \mathbf{y}_i^{(1)} - \Sigma_{ih}^{-1} \mu_{ih}^2)\right),$$

where $\Sigma_{ih} = (\psi_h^{-1} + \sigma^{-2} (\mathbf{X}_i \boldsymbol{\Gamma}_h)' (\mathbf{X}_i \boldsymbol{\Gamma}_h))^{-1}$ and $\mu_{ih} = \Sigma_{ih} (\psi_h^{-1} \beta_h + \sigma^{-2} (\mathbf{X}_i \boldsymbol{\Gamma}_h)' \mathbf{y}_i^{(1)})$, we sample β_{ih}^* from $N(\beta_{ih}^*; \mu_{ih}, \Sigma_{ih})$.

- The full conditional posterior distribution of θ_{sh} is

$$N(\theta_{sh}; \mu_{sh}^*, \Sigma_{sh}^*), \quad s = 1, \dots, k_h,$$

where $\Sigma_{sh}^* = (\psi_h^{-1} + \sigma^{-2} \sum_{i \in I_{sh}} (\mathbf{X}_i \boldsymbol{\Gamma}_h)' (\mathbf{X}_i \boldsymbol{\Gamma}_h))^{-1}$ and $\mu_{sh}^* = \Sigma_{sh}^* (\sigma^{-2} \sum_{i \in I_{sh}} (\mathbf{X}_i \boldsymbol{\Gamma}_h)' \mathbf{y}_i^{(1)} + \psi_h^{-1} \beta_h)$.

- The full conditional posterior distribution of α_h is

$$\pi(\alpha_h | \phi_h, k_h) \propto \lambda_h \mathcal{G}(a + k_h, b - \log(\phi_h)) + (1 - \lambda_h) \mathcal{G}(a + k_h - 1, b - \log(\phi_h)),$$

where $\lambda_h = \frac{a + k_h - 1}{n(b - \log(\phi_h)) + a + k_h - 1}$, and $\pi(\phi_h | \alpha_h) \propto \phi_h^{\alpha_h} (1 - \phi_h)^{n-1}$.

- The full conditional posterior distribution of $\boldsymbol{\gamma}_\psi$ is

$$\pi(\boldsymbol{\gamma}_\psi | \boldsymbol{\beta}_i^*, \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) \propto N(\hat{E}_\psi, \hat{V}_\psi),$$

where $\hat{V}_\psi = (\sigma^{-2} \sum_{i=1}^n \mathbf{V}_i^* \mathbf{V}_i^* + V_\psi^{-1})^{-1}$, $\hat{E}_\psi = \hat{V}_\psi (\sigma^{-2} \sum_{i=1}^n \mathbf{V}_i^* (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_i^*) + V_\psi^{-1} E_\psi)$, $\mathbf{V}_i = (V_{i1}, \dots, V_{in_i})'_{n_i \times P}$ with $V_{ij} = (x_{ijl}(\beta_{ih}^* - \beta_h) : h = 1, \dots, p-1; l = h+1, \dots, p)'$, and $P = \frac{1}{2}p(p-1)$, and \mathbf{V}_i^* denotes \mathbf{V}_i removing the elements corresponding to zero of $\boldsymbol{\psi}$.

- In (12), the conditional probability of $\beta_h = 0$ is

$$\hat{\lambda}_{1,h} = \frac{\lambda_{1,0h}}{\lambda_{1,0h} + (1 - \lambda_{1,0h}) BF_1}$$

with

$$BF_1 = \frac{\hat{V}_h^{\frac{1}{2}}}{\sigma_{0h}} \exp\left(\frac{1}{2} \left(\hat{V}_h^{-1} \hat{E}_h^2 - \sigma_{0h}^{-2} \beta_{0h}^2 \right)\right),$$

where $\hat{V}_h = (\tilde{V}_h^{-1} + \sigma_{0h}^{-2})^{-1}$, $\hat{E}_h = \hat{V}_h(\tilde{V}_h^{-1}\tilde{E}_h + \sigma_{0h}^{-2}\beta_{0h})$, $\tilde{V}_h = (\sigma^{-2}\sum_{i=1}^n(\mathbf{X}_i\boldsymbol{\Gamma}_h^*)'(\mathbf{X}_i\boldsymbol{\Gamma}_h^*) + k_h\psi_h^{-1})^{-1}$, $\tilde{E}_h = \tilde{V}_h(\sigma^{-2}\sum_{i=1}^n(\mathbf{X}_i\boldsymbol{\Gamma}_h^*)'\mathbf{y}_i^{(2)} + \psi_h^{-1}\sum_{s=1}^{k_h}\theta_{sh})$, $\boldsymbol{\Gamma}^* = \mathbf{I} - \boldsymbol{\Gamma}$, $\mathbf{y}_i^{(2)} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\Gamma}\boldsymbol{\beta}_i^* - \mathbf{X}_i\boldsymbol{\Gamma}_{-h}^*\boldsymbol{\beta}_{-h}$, $\boldsymbol{\Gamma}_h^*$ denotes the h th column of $\boldsymbol{\Gamma}^*$, $\boldsymbol{\Gamma}_{-h}^*$ denotes the submatrix of $\boldsymbol{\Gamma}^*$ excluding the h th column, and $\boldsymbol{\beta}_{-h}$ denotes the subvector of $\boldsymbol{\beta}$ with β_h excluded.

When $\psi_h = 0$, $\tilde{V}_h = \sigma^2(\sum_{i=1}^n\sum_{j=1}^{n_i}x_{ijh}^2)^{-1}$, $\tilde{E}_h = \sigma^{-2}\tilde{V}_h\sum_{i=1}^n\sum_{j=1}^{n_i}x_{ijh}y_{ij}^{(2)}$, $y_{ij}^{(2)} = y_{ij} - \mathbf{x}'_{ij(-h)}\boldsymbol{\beta}_{-h} - \mathbf{x}'_{ij}\boldsymbol{\Gamma}_{-h}(\boldsymbol{\beta}_{i,-h}^* - \boldsymbol{\beta}_{-h})$.

- In (13), the conditional probability of $\psi_h = 0$ is

$$\hat{\lambda}_{2,h} = \frac{\lambda_{2,0h}}{\lambda_{2,0h} + (1 - \lambda_{2,0h})BF_2}$$

with

$$BF_2 = \frac{\mathbf{L}(\boldsymbol{\theta}_h, \boldsymbol{\beta}_{i,-h}^*, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \sigma^{-2}; \mathbf{y})\Gamma(\hat{c}_h)d_{0h}^{c_{0h}}}{\mathbf{L}(\boldsymbol{\theta}_h = \boldsymbol{\beta}_h, \boldsymbol{\beta}_{i,-h}^*, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \sigma^{-2}; \mathbf{y})\Gamma(c_{0h})d_h^{\hat{c}_h}},$$

where $\mathbf{L}(\boldsymbol{\theta}_h, \boldsymbol{\beta}_{i,-h}^*, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \sigma^{-2}; \mathbf{y}) = \prod_{s=1}^{k_h}\prod_{i \in I_{sh}}\mathbf{N}(\mathbf{y}_i^{(1)}; \mathbf{X}_i\boldsymbol{\Gamma}_h\theta_{sh}, \sigma^2\mathbf{I}_{n_i})$, $\hat{c}_h = c_{0h} + \frac{k_h}{2}$, $\hat{d}_h = d_{0h} + \frac{1}{2}\sum_{s=1}^{k_h}(\theta_{sh} - \beta_h)^2$.

- The full conditional posterior distribution for σ^{-2} is

$$\pi(\sigma^{-2} | \boldsymbol{\beta}_i^*, \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) \propto \mathcal{G}\left(c + \frac{1}{2}\sum_{i=1}^n n_i, d + \frac{1}{2}\sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)\right).$$

Table 1: Comparison of the estimates of the parameters in simulation study.

Parameter	True value	LME	CDM	NME
β_1	0	-0.07 _{(-0.41,0.26)^a}	0.12 _{(-0.11,0.32)^b}	0.09 _{(-0.07,0.21)^b}
β_2	2	1.91 _(1.66,2.16)	2.03 _(1.71,2.32)	2.02 _(1.71,2.27)
β_3	3	2.89 _(2.50,3.38)	3.10 _(2.69,3.50)	3.06 _(2.64,3.39)
β_4	6	5.96 _(5.40,6.52)	6.01 _(5.61,6.46)	6.03 _(5.67,6.51)
d_{11}	4	3.46	4.22 _(3.58,5.16)	4.17 _(3.63,5.05)
d_{22}	0	0.02	0.02 _(0.00,0.03)	0.00 _(0.00,0.02)
d_{33}	3.94	3.92	3.96 _(3.52,4.37)	3.96 _(3.63,4.30)
d_{44}	12	10.88	12.27 _(11.53,13.18)	12.24 _(11.68,13.10)
ρ_{21}	0	-0.89	0.01 _(0.00,0.02)	0.01 _(0.00,0.01)
ρ_{31}	0.60	0.71	0.65 _(0.59,0.69)	0.63 _(0.59,0.70)
ρ_{32}	0	-0.72	0.04 _(0.00,0.06)	0.02 _(0.00,0.05)
ρ_{41}	0.86	0.92	0.89 _(0.81,0.91)	0.88 _(0.80,0.91)
ρ_{42}	0	-0.95	0.00 _(0.00,0.05)	0.00 _(0.00,0.01)
ρ_{43}	0.85	0.83	0.88 _(0.79,0.91)	0.88 _(0.81,0.91)
σ^{-2}	1	0.96	0.97 _(0.89,1.07)	0.98 _(0.92,1.06)

^a 95% confidence interval

^b 95% credible interval

Table 2: Estimated model posterior probabilities in simulation study.

Model	CDM	NME	DIC
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij3}, z_{ij4}^a$	0.356 ^b _{(0.310,0.472)^c}	0.483 _(0.349,0.522)	3630.4
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij4}$	0.147 _(0.125,0.200)	0.161 _(0.120,0.214)	3661.7
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij3}$	0.168 _(0.130,0.247)	0.129 _(0.093,0.187)	3659.0
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij2}, z_{ij4}$	0.079 _(0.058,0.092)	0.081 _(0.061,0.092)	3710.6
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$	0.083 _(0.069,0.096)	0.056 _(0.038,0.078)	3711.4
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$	0.045 _(0.037,0.059)	0.024 _(0.020,0.036)	3809.0
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij2}, z_{ij4}$	0.040 _(0.030,0.048)	0.020 _(0.015,0.028)	3860.0
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij3}, z_{ij4}$	0.012 _(0.010,0.018)	0.012 _(0.011,0.020)	3908.4
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}$	0.021 _(0.019,0.025)	0.010 _(0.007,0.016)	3917.2
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, z_{ij3}, z_{ij4}$	0.021 _(0.017,0.024)	0.008 _(0.003,0.011)	3906.9
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij3}$	0.013 _(0.009,0.015)	0.008 _(0.004,0.010)	3930.5
$x_{ij2}, x_{ij3}, x_{ij4}, z_{ij1}$	0.010 _(0.008,0.013)	0.008 _(0.005,0.012)	3945.9
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, z_{ij3}$	0.005 _(0.000,0.007)	0	3970.0

^a True model

^b Posterior probability

^c Range

Table 3: Estimated model posterior probabilities of top ten models in the OECD data.

Model	Posterior Probability	Range
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}$	0.183	(0.157,0.226)
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}$	0.112	(0.098,0.139)
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}, z_{ij6}$	0.107	(0.077,0.131)
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}$	0.094	(0.069,0.114)
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, z_{ij1}, z_{ij3}$	0.062	(0.041,0.089)
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij4}$	0.048	(0.025,0.067)
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$	0.040	(0.029,0.052)
$x_{ij1}, x_{ij3}, x_{ij4}, z_{ij1}, z_{ij3}, z_{ij4}$	0.029	(0.014,0.041)
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}, z_{ij5}, z_{ij6}$	0.026	(0.018,0.039)
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij4}$	0.025	(0.017,0.040)

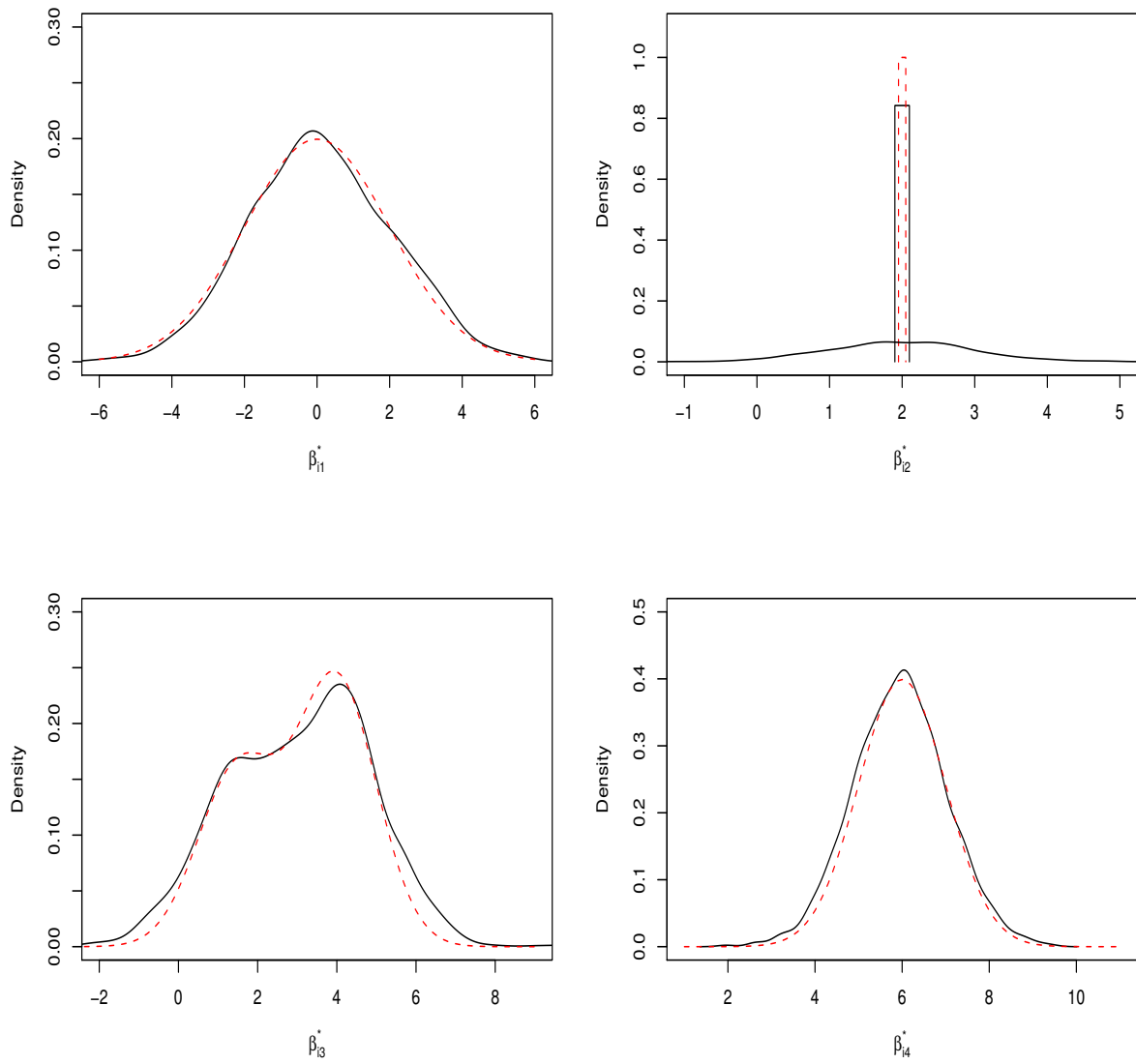


Figure 1: Posterior densities (solid lines) and true densities (dashed lines) of the parameters β_i^* in the simulation study. The vertical bars denote the point masses at 2.

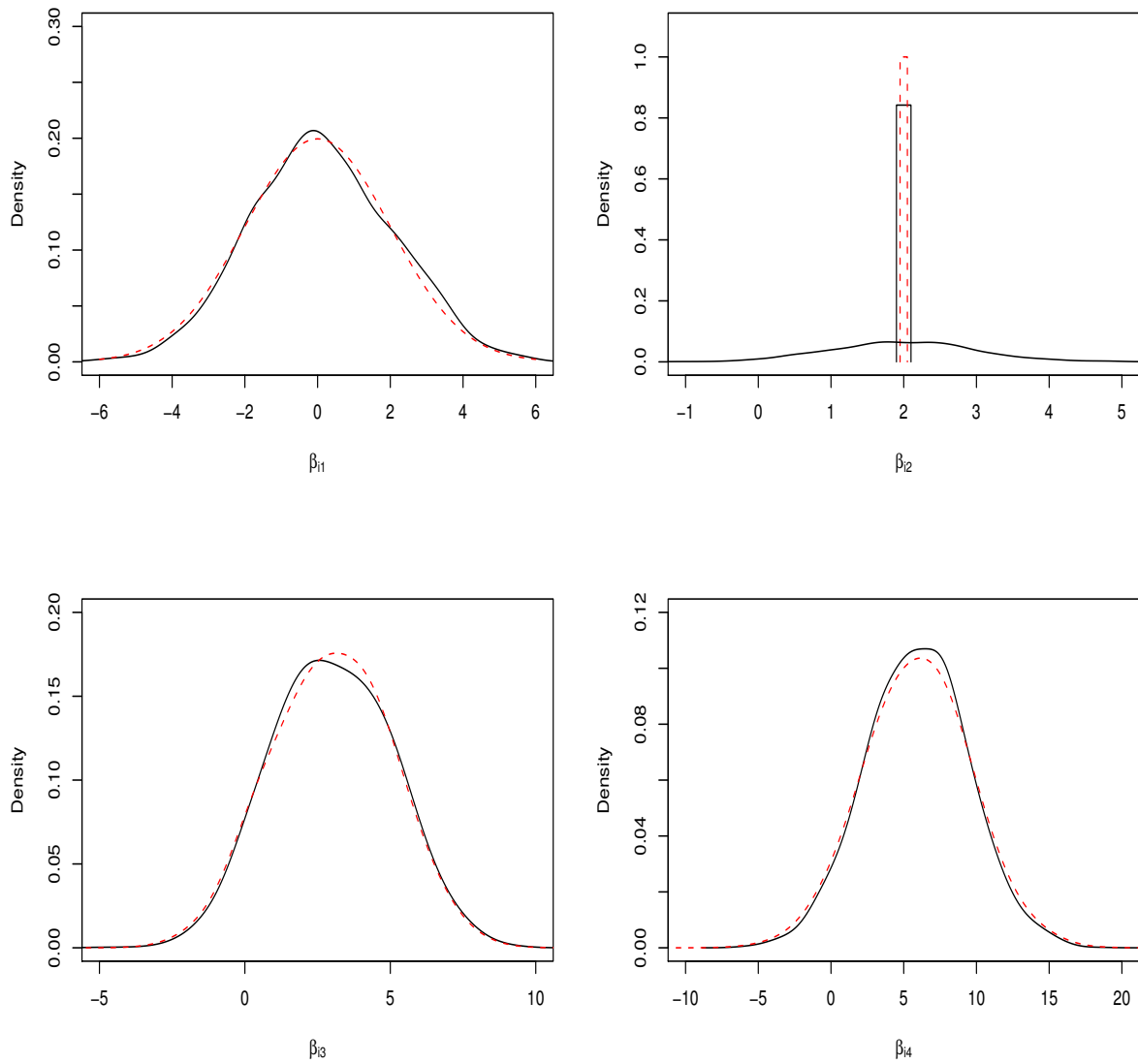


Figure 2: Posterior densities (solid lines) and true densities (dashed lines) of the regression coefficients β_i in the simulation study. The vertical bars denote the point masses at 2.

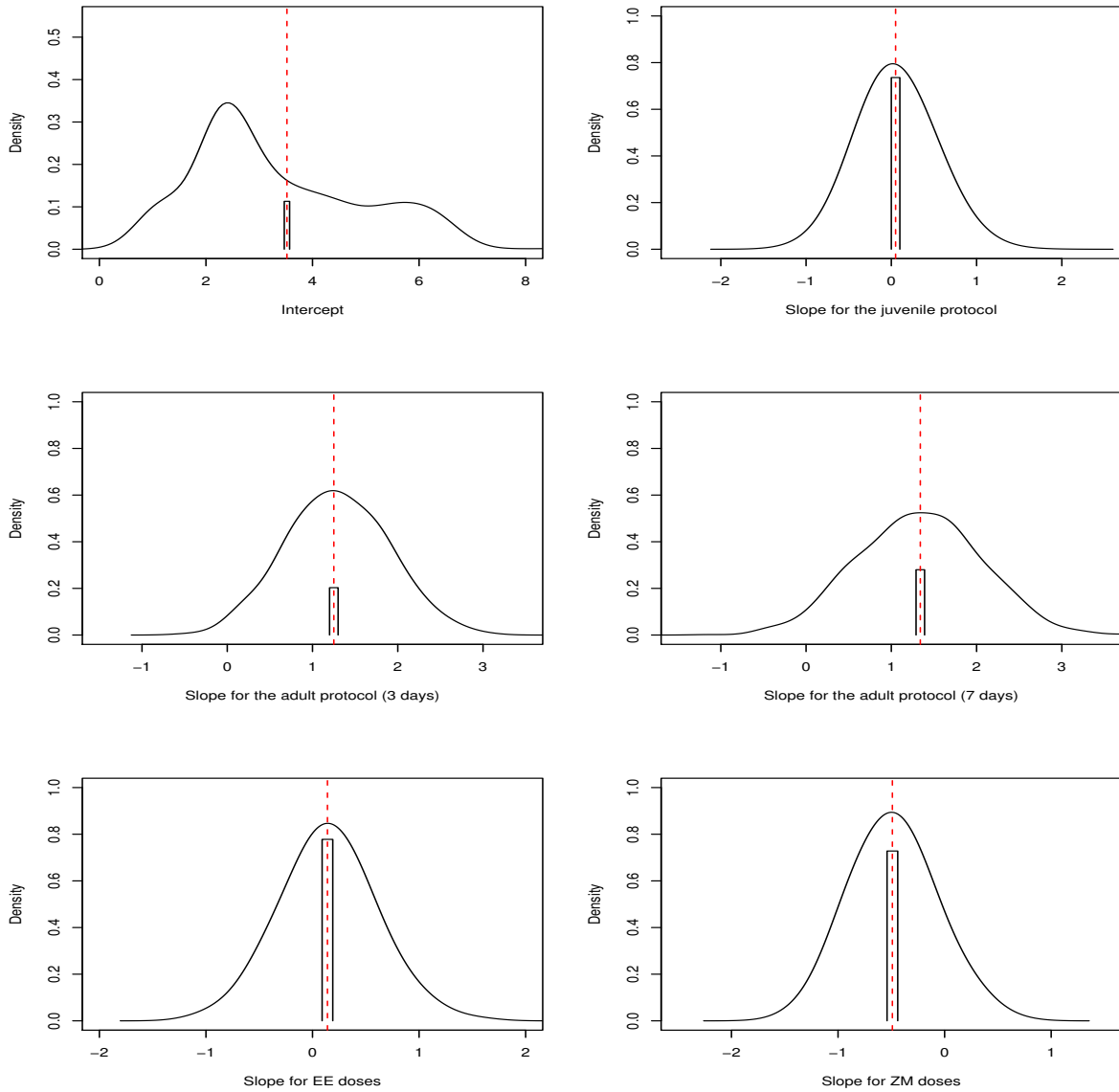


Figure 3: Posterior densities (solid lines) of the intercept and slopes for protocols, EE doses and ZM doses in the application. The vertical bars denote the point masses and the vertical lines denote the posterior means.