Biostatistics (2006), **7**, 4, *pp.* 551–568
doi:10.1093/biostatistics/kxj025
Advance Access publication on February 17, 2006

Bayesian dynamic modeling of latent trait distributions

DAVID B. DUNSON

Biostatistics Branch, National Institute of Environmental Health Sciences, MD A3-03, PO Box 12233, Research Triangle Park, NC 27709, USA
dunson1@niehs.nih.gov

SUMMARY

Studies of latent traits often collect data for multiple items measuring different aspects of the trait. For such data, it is common to consider models in which the different items are manifestations of a normal latent variable, which depends on covariates through a linear regression model. This article proposes a flexible Bayesian alternative in which the unknown latent variable density can change dynamically in location and shape across levels of a predictor. Scale mixtures of underlying normals are used in order to model flexibly the measurement errors and allow mixed categorical and continuous scales. A dynamic mixture of Dirichlet processes is used to characterize the latent response distributions. Posterior computation proceeds via a Markov chain Monte Carlo algorithm, with predictive densities used as a basis for inferences and evaluation of model fit. The methods are illustrated using data from a study of DNA damage in response to oxidative stress.

Keywords: Dynamic Dirichlet process; Factor analysis; Hierarchical model; Latent variables; Measurement error; Random effect; Surrogate data.

1. INTRODUCTION

In many applications, the primary response variable of interest cannot be measured directly and one must instead rely on multiple surrogates. For example, in studying DNA damage and repair, it is not feasible to directly measure the frequency of DNA strand breaks for each cell in a sample. However, using single-cell gel electrophoresis (also known as the comet assay), one can obtain multiple measures that relate directly to the frequency of strand breaks. In such settings, it is natural to use a latent response model in which the different measured outcomes are assumed to be manifestations of a latent variable, which in turn may depend on covariates, such as the dose of an exposure.

Often, in applying such models, one assumes that both the latent and manifest variables are normally distributed (see Roy and Lin, 2000, 2002; Xu and Zeger, 2001, for recent references). However, a number of approaches have been proposed which allow the measured outcomes to have different parametric distributions, typically restricted to be underlying normal (Muthén, 1984; Shi and Lee, 2000) or in the exponential family (Muthén, 1984; Sammel *et al.*, 1997; Moustaki and Knott, 2000; Dunson, 2000, 2003). In addition, one can potentially use a latent class model in which the underlying response is categorical (see Miglioretti, 2003, for a recent reference). Since full likelihood approaches are often difficult to implement and may be sensitive to distributional assumptions, some authors have advocated the use of robust score tests (Sammel and Ryan, 2002) or estimating equation-based approaches

(Reboussin *et al.*, 1999; Roy *et al.*, 2003). Such methods are most useful when interest focuses on assessing changes in the overall mean response profile with covariates. However, in certain applications, one may anticipate possible changes in not only the mean but also the distributional shape across levels of a predictor. For example, in the molecular epidemiology studies which motivated this article, the distribution of DNA damage across cells in a sample may have different shapes depending on the level of oxidative stress induced by a chemical exposure, the amount of time after exposure damage is measured, and the presence of polymorphisms in genes involved in the base excision repair pathway.

To allow the surrogate outcome distributions to be unknown, Dunson *et al.* (2003) proposed an approximate Bayesian approach for quantile regression. Following Lavine (1995), they replaced the likelihood function with a substitution likelihood based on quantiles. Covariate effects were incorporated on the level of the surrogate outcomes and residual dependency was accommodated through a shared latent normal variable. In order to reduce dimensionality in assessing covariate effects on the latent response of interest, it may be preferable to allow covariates to affect the location of the latent variable, while avoiding parametric assumptions about the latent variable distribution.

This article proposes a Bayesian semiparametric approach to this problem. The surrogate outcomes are related to a latent response variable through a factor analytic model, with a scale mixture (West, 1987) of underlying normals used to characterize flexibly the measurement error distributions. Our primary focus is on developing an approach for assessing dynamic changes in the latent response distribution across levels of a predictor, $X \in \{1, \ldots, d\}$. For example, X may represent the level of a treatment, age, or time since exposure. To allow for uncertainty in the latent response distribution conditional on X, we propose a dynamic mixture of Dirichlet processes (DMDP). In particular, the latent response distribution in group h is represented as a mixture of the distribution in group $h-1$ and an unknown innovation distribution, which is assigned a Dirichlet process (DP) prior (Ferguson, 1973, 1974). This structure accommodates autocorrelation in the distributions, and results in a flexible dynamic mixture structure for the surrogate outcomes.

The proposed approach is an alternative to the dependent Dirichlet process (DDP) of MacEachern (1999, 2000), which is a class of priors for a collection of unknown distributions (see also De Iorio *et al.*, 2002, 2004; Gelfand *et al.*, 2004). The DDP characterizes dependence through a stochastic process for a fixed number of atoms in the unknown distributions. The proposed DMDP instead allows evolving changes in the number of atoms through a weighted mixture of independent DP measures. This approach extends the weighted mixture formulation of Müller *et al.* (2004), which was used to borrow strength across studies, to a time series setting. An innovative alternative approach to defining dependent nonparametric measures was recently proposed by Griffin and Steel (2006). Their order-based DDP allows the weights in the Sethuraman (1994) stick-breaking representation of the DP to be dependent on covariates. They considered a nonparametric time series model as a special case, though our proposed DMDP has advantages in terms of ease of implementation.

For a recent review of Bayesian nonparametric inference, refer to Müller and Quintana (2004). Several authors have used DP priors for intermediate variables in hierarchical models, without allowing the unknown distributions to vary with covariates. Bush and MacEachern (1996) used a DP mixture for a random block factor, and Kleinman and Ibrahim (1998) applied a related approach to random effects distributions in mixed effects models. Also considering semiparametric linear mixed models, Ishwaran and Takahara (2002) developed an iid weighted Chinese restaurant algorithm for inference. Mukhopadhyay and Gelfand (1997) proposed a general class of DP mixtures of hierarchical generalized linear models. Recent authors have considered improved approaches for computation and inference (Neal, 2000; Gelfand and Kottas, 2002; Ishwaran and James, 2002).

Section 2 proposes the semiparametric latent response model and prior structure. Section 3 outlines a hybrid Gibbs sampler and Metropolis algorithm for posterior computation, and discusses inferences.

Section 4 applies the approach to data from a study of DNA damage in relationship with oxidative stress, and Section 5 discusses the results.

2. Semiparametric hierarchical model

2.1 *Data structure and measurement model*

Let $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hip})'$ denote a $p \times 1$ vector of surrogate measurements for the latent response of the ith ($i = 1, \ldots, n_h$) subject in group h ($h = 1, \ldots, d$). For example, in the DNA damage study, $\mathbf{y}_{hi}$ denotes surrogates of DNA damage for the ith cell in dose group h. The elements of $\mathbf{y}_{hi}$ are ordered so that the first p_1 ($0 \leqslant p_1 \leqslant p$) elements are continuous and the remaining $p_2 = p - p_1$ elements are categorical. To facilitate joint modeling, we link the categorical surrogates to the underlying continuous variables as in Muthén (1984). Formally, let $y_{hij} = g_j(y^*_{hij}; \boldsymbol{\tau}_j)$, for $j = 1, \ldots, p$, where y^*_{hij} is a continuous variable underlying y_{hij}. For the continuous surrogates, we have $y_{hij} = y^*_{hij}$ for $j = 1, \ldots, p_1$. For the categorical surrogates, with $y_{hij} \in \{1, \ldots, d_j\}$, we have $y_{hij} = \sum_{l=1}^{d_j} l \times 1(\tau_{j,l-1} < y^*_{hij} < \tau_{j,l})$ for $j = p_1 + 1, \ldots, p$, where $\boldsymbol{\tau}_j = (\tau_{j,0}, \ldots, \tau_{j,d_j})'$ are thresholds satisfying $-\infty = \tau_{j,0} < \tau_{j,1} = 1 < \tau_{j,2} < \cdots < \tau_{j,d_j-1} < \tau_{j,d_j} = \infty$. Hence, $g_j(\cdot)$ is the identity link for continuous surrogates, and is otherwise a threshold link mapping from $\mathbb{R} \to \{1, \ldots, d_j\}$, where d_j is the number of categories of the jth surrogate.

Letting $\mathbf{y}^*_{hi} = (y^*_{hi1}, \ldots, y^*_{hip})'$, we relate the underlying continuous variables to the latent response through the following measurement model:

$$\mathbf{y}^*_{hi} = \boldsymbol{\mu} + \boldsymbol{\lambda}\eta_{hi} + \boldsymbol{\epsilon}_{hi}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ is a vector of intercept parameters, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ are factor loadings, η_{hi} is a latent response variable for subject i in group h, and $\boldsymbol{\epsilon}_{hi} = (\epsilon_{hi1}, \ldots, \epsilon_{hip})'$ is a vector of independently distributed measurement errors measuring idiosyncratic features of the different surrogates. A primary goal in considering this model is to assess how the latent response distribution changes between groups.

To address this goal, one could potentially use a mean regression model in which $\mathrm{E}(\eta_{hi}) = \mathbf{x}'_{hi}\boldsymbol{\beta}$ and $\mathrm{V}(\eta_{hi}) = 1$, where $\mathbf{x}_{hi} = [1(h = 2), \ldots, 1(h = d)]'$ is a vector of group indicator variables and the variance of the latent variable density is fixed at 1 for identifiability. In fitting the model and performing inferences, one could avoid parametric assumptions on the residual measurement error and latent variable distributions by using two-stage least squares procedures (with some risk of bias). Alternatively, one could follow a full likelihood-based or Bayesian approach after specifying a distribution for $\boldsymbol{\epsilon}_{hi}$ and η_{hi}. For example, an obvious choice that satisfies the moment constraints would be $\boldsymbol{\epsilon}_{hi} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $\eta_{hi} \sim \mathrm{N}(\mathbf{x}'_{hi}\boldsymbol{\beta}, 1)$, where $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$ is the measurement error precision matrix and $\psi_j = 1$ for the categorical surrogates, $j = p_1 + 1, \ldots, p$, to ensure identifiability. Related approaches have been considered by Sammel *et al.* (1997) and Dunson (2003), among others.

As a more flexible approach for modeling of the residual distributions, we use a scale mixture of normal distributions (Fernandez and Steel, 2001) by letting $\epsilon_{hij} \sim \mathrm{N}(0, \sigma^2_{hij})$, where $\sigma^2_{hij} = \kappa^{-1}_{hij}\psi^{-1}_j$ with $\kappa_{hij} \sim \mathcal{G}(\nu_j/2, \nu_j/2)$. This specification results in a t density with ν_j degrees of freedom and, for $\nu_j > 2$, mean 0 and variance $\psi^{-1}_j \nu_j/(\nu_j - 2)$ for the measurement error ϵ_{hij} in the continuous variable underlying the jth surrogate. For continuous surrogates, this accounts for heavier tails than expected under the normal distribution, while for categorical surrogates, the use of a t-distribution results in a more flexible link function than the probit form: $\Pr(y_{hij} \leqslant l|\eta_{hi}) = \Phi(\tau_{j,l} - \mu_j - \lambda_j\eta_{hi})$, where $\Phi(\cdot)$ denotes the standard normal distribution function, implied by assuming $\epsilon_{ij} \sim \mathrm{N}(0, 1)$. Here, we set $\psi_j = 1$, for

$j = p_1 + 1, \ldots, p$, for identifiability. See Johnson and Albert (1999) for an overview of the Bayesian modeling of categorical data using latent variable formulations.

2.2 *Nonparametric latent response model*

Our focus is primarily on developing a nonparametric specification for the latent response distribution. Let $\eta_{1i} \sim G_1$, where $G_1 \sim \mathcal{D}(\alpha_0 G_0)$ is an unknown distribution drawn from a DP centered on nonatomic base distribution G_0 and with precision parameter α_0 (as in Antoniak, 1974). Following Sethuraman's (1994) stick-breaking representation, we specify

$$G_1 = \sum_{l=1}^{\infty} p_{1l}\delta_{\theta_{1l}}, \quad \frac{p_{1l}}{\prod_{m=1}^{l-1}(1-p_{1m})} \overset{\text{iid}}{\sim} \text{beta}(1, \alpha_0), \quad \text{and} \quad \theta_{1l} \overset{\text{iid}}{\sim} G_0, \tag{2.2}$$

where δ_θ denotes the degenerate distribution with all its mass at θ, $\{p_{1l}, l = 1, 2, \ldots, \infty\}$ is an infinite sequence of random weights, and $\{\theta_{1l}, l = 1, 2, \ldots, \infty\}$ is a corresponding sequence of random atoms generated from G_0. It can be shown that G_1 is almost surely discrete.

Letting $\boldsymbol{\eta}_1 = (\eta_{11}, \ldots, \eta_{1,n_1})'$, the DP structure implies that the elements of $\boldsymbol{\eta}_1$ are allocated to $k_1 \leqslant n_1$ unique values (or clusters), which we denote as $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_{k_1})'$. Letting $\boldsymbol{\eta}_1^{(i)}$ denote the subvector of $\boldsymbol{\eta}_1$ excluding the ith element, the conditional distribution of η_{1i} given $\boldsymbol{\eta}_1^{(i)}$ is

$$\left(\frac{\alpha_0}{\alpha_0 + n_1 - 1}\right) G_{01} + \sum_{l=1}^{k_1^{(i)}} \left(\frac{n_{1l}^{(i)}}{\alpha_0 + n_1 - 1}\right) \delta_{\theta_{1l}^{(i)}}, \tag{2.3}$$

where the unique values of $\boldsymbol{\eta}_1^{(i)}$ are denoted as $\boldsymbol{\theta}_1^{(i)} = (\theta_{1l}^{(i)}, l = 1, \ldots, k_1^{(i)})'$, and $n_{1l}^{(i)}$ denotes the number of elements of $\boldsymbol{\eta}_1^{(i)}$ having value $\theta_{1l}^{(i)}$. This distribution is the mixture of the base distribution G_0 and a uniform distribution with support on $\boldsymbol{\eta}_1^{(i)}$, with the mixture weights depending on α_0 and the sample size n_1.

The form of (2.3) simplifies posterior computation and prediction of the latent response for an additional subject in group 1, denoted η_{1,n_1+1}. In particular, the conditional predictive density of η_{1,n_1+1} given $\boldsymbol{\eta}_1$ is simply

$$\left(\frac{\alpha_0}{\alpha_0 + n_1}\right) G_0 + \sum_{l=1}^{k_1} \left(\frac{n_{1l}}{\alpha_0 + n_1}\right) \delta_{\theta_{1l}}. \tag{2.4}$$

Potentially, this predictive distribution could be used as a reasonable best guess for G_2, the distribution of η_{2i}, the latent response for a subject in the second group. Such an approach would indirectly account for dependency between G_1 and G_2 by modeling G_2 conditionally on $\boldsymbol{\eta}_1$. We prefer to specify explicitly the dependence between G_2 and G_1. In particular, it is reasonable to assume that G_2 shares features with G_1 but that innovations may have occurred. This can be modeled using the mixture structure $G_2 = (1-\pi_1)G_1 + \pi_1 H_1$, where $0 \leqslant \pi_1 \leqslant 1$ and $H_1 \sim \mathcal{D}(\alpha_1 H_{01})$ is a DP-distributed 'innovation' distribution. Note that this formulation randomly modifies the discrete distribution G_1 by (i) reducing the probabilities allocated to the atoms in G_1 by a multiplicative factor $(1 - \pi_1)$ and (ii) incorporating new atoms drawn from the nonatomic base distribution H_{01}.

Letting $\mathcal{B}_1, \ldots, \mathcal{B}_K$ denote Borel sets partitioning $\mathbb{R}$, we have

$$\begin{aligned}[G_2(\mathcal{B}_1, \ldots, \mathcal{B}_K)|\pi_1, G_1] &\sim (1-\pi_1)G_1(\mathcal{B}_1, \ldots, \mathcal{B}_K) + \pi_1 \mathcal{D}_K(\alpha_1 H_{01}(\mathcal{B}_1), \ldots, \alpha_1 H_{01}(\mathcal{B}_K)) \\ &\overset{\text{d}}{=} G_1(\mathcal{B}_1, \ldots, \mathcal{B}_K) + \Delta_2(\mathcal{B}_1, \ldots, \mathcal{B}_K),\end{aligned} \tag{2.5}$$

where $D_K(\cdot)$ denotes the finite K-dimensional Dirichlet distribution, and

$$\Delta_2(\mathcal{B}_1, \ldots, \mathcal{B}_K) = \pi_1\{\mathcal{D}_K(\alpha_1 H_{01}(\mathcal{B}_1), \ldots, \alpha_1 H_{01}(\mathcal{B}_K)) - G_1(\mathcal{B}_1, \ldots, \mathcal{B}_K)\}$$

is a random innovation on $G_1(\mathcal{B}_1, \ldots, \mathcal{B}_K)$. For any $\mathcal{B} \subset \mathbb{R}$, we have

$$\begin{aligned}
\mathrm{E}\{\Delta_2(\mathcal{B})|\pi_1, G_1, \alpha_1\} &= \pi_1\{H_{01}(\mathcal{B}) - G_1(\mathcal{B})\} \\
\mathrm{V}\{\Delta_2(\mathcal{B})|\pi_1, G_1, \alpha_1\} &= \frac{\pi_1^2 H_{01}(\mathcal{B})\{1 - H_{01}(\mathcal{B})\}}{(1+\alpha_1)}.
\end{aligned}$$

The hyperparameters π_1 and H_{01} control the magnitude of the expected change from G_1 to G_2, with $G_2 = G_1$ in the limit as $\pi_1 \to 0$ and $\mathrm{E}\{G_2(\mathcal{B})|\pi_1, G_1, \alpha_1\} = G_1(\mathcal{B})$ as $H_{01} \to G_1$. The variance of the change is controlled by π_1 and α_1, with $V\{\Delta_2(\mathcal{B})|\pi_1, G_1, \alpha_1\} \to 0$ in the limit as $\alpha_1 \to \infty$ or $\pi_1 \to 0$. We do not consider the case in which $\alpha_1 \to 0$ because that corresponds to the degenerate case in which H_1 places all its mass at a single point.

Extending this approach to later groups ($h = 2, \ldots, d$), we let $\eta_{hi} \sim G_h$, with

$$\begin{aligned}
G_h &= (1 - \pi_{h-1})G_{h-1} + \pi_{h-1}H_{h-1} \\
&= \left\{\prod_{l=1}^{h-1}(1-\pi_l)\right\} G_1 + \sum_{l=1}^{h-1}\left\{\prod_{m=l+1}^{h-1}(1-\pi_m)\right\}\pi_l H_l \\
&= \omega_{h1}G_1 + \omega_{h2}H_1 + \cdots + \omega_{hh}H_{h-1} \\
H_l &\sim \mathcal{D}(\alpha_l H_{0l}), \quad \text{for } l = 1, \ldots, h-1,
\end{aligned} \tag{2.6}$$

where $\omega_{hl} = \pi_{l-1}\prod_{m=l}^{h-1}(1-\pi_m)$, for $l = 1, \ldots, h$, with $\pi_0 = 1$ and $\boldsymbol{\omega}_h = (\omega_{h1}, \ldots, \omega_{hh})'$, are probability weights on the different components in the mixture. Note that this model can be expressed equivalently as

$$\begin{aligned}
\eta_{hi} &= \sum_{l=1}^{h} 1(M_{hi} = l)\xi_{hil} \\
M_{hi} &\sim \text{Multinomial}(1, \ldots, h; \omega_{h1}, \ldots, \omega_{hh}) \\
\xi_{hil} &\sim G_l^*, \quad G_l^* \sim \mathcal{D}(\alpha_l G_{0l}^*), \quad \text{for } l = 1, \ldots, h,
\end{aligned} \tag{2.7}$$

where $G_l^* = G_1$, $G_{0l}^* = G_0$ for $l = 1$ and $G_l^* = H_{l-1}$, $G_{0l}^* = H_{0,l-1}$ for $l = 2, \ldots, h$. This formulation expresses η_{hi} as equal to a randomly selected element out of a set of independent DP-distributed latent factors $\boldsymbol{\xi}_{ih} = \{\xi_{ih1}, \ldots, \xi_{ihh}\}$.

The Appendix derives the correlation coefficient between $G_{h-1}(\mathcal{B})$ and $G_h(\mathcal{B})$ and the marginal mean and variance of $G_h(\mathcal{B})$. Focusing on the special case in which $G_{0l}^*(\mathcal{B}) = G_0^*(\mathcal{B})$, for $l = 1, \ldots, d$, so that the same base distribution is chosen for each component in the mixture, we have

$$\mathrm{Corr}(G_{h-1}(\mathcal{B}), G_h(\mathcal{B})) = \frac{\sum_{l=1}^{h-1}\omega_{hl}\omega_{h-1,l}(\alpha_l+1)^{-1}}{\left[\sum_{l=1}^{h}\omega_{hl}^2(\alpha_l+1)^{-1}\right]^{1/2}\left[\sum_{l=1}^{h-1}\omega_{h-1,l}^2(\alpha_l+1)^{-1}\right]^{1/2}}, \tag{2.8}$$

with the $(\alpha_l+1)^{-1}$ terms dropping out in the special case in which $\alpha_l = \alpha$, for $l = 1, \ldots, d$. Because the ω values depend on $\boldsymbol{\pi}$, it is clear from this expression that the correlation in the unknown distributions is driven by these mixture weights. This expression is particularly useful due to its simplicity and its lack of dependency on $\mathcal{B}$.

Because factors drawn from a common DP cluster together as characterized in the Pólya urn scheme of Expression (2.3), it is clear that latent variables for subjects in different groups can belong to the same cluster. For example, if $M_{hi} = M_{h'i'}$, then subjects i and i' can potentially belong to the same cluster, so that $\eta_{hi} = \eta_{h'i'}$. The prior probability of clustering together two subjects h, i and h', i' in the same or different groups is simply

$$\Pr(\eta_{hi} = \eta_{h'i'}) = \sum_{l=1}^{\min(h,h')} \frac{\omega_{hl}\omega_{h'l}}{\alpha_l + 1}, \tag{2.9}$$

which is the probability that they are sampled from the same mixture component and are then grouped together, summed across the different possibilities for the mixture component. It is clear from the above expressions that $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)'$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)'$ are key hyperparameters controlling the clustering process and dynamic changes in the latent variable distribution across groups. Potentially, a reasonable simplifying assumption in some applications may be $\pi_h = \pi$ and $\alpha_h = \alpha$, for $h = 1, \ldots, d$. This special case may be particularly useful when group sizes are small. However, for greater flexibility, we choose hyperprior distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\alpha}$ as follows:

$$\pi(\boldsymbol{\pi}) = \prod_{h=1}^{d} \text{beta}(\pi_h; a_{\pi_h}, b_{\pi_h}) \quad \text{and} \quad \pi(\boldsymbol{\alpha}) = \prod_{h=1}^{d} \mathcal{G}(\alpha_h; a_{\gamma_h}, b_{\gamma_h}), \tag{2.10}$$

where $\mathcal{G}(a, b)$ denotes the gamma density with mean a/b and variance a/b^2. In most applications, one may expect a priori that correlation between G_{h-1} and G_h is moderate to high. Such belief corresponds to the expectation that the π values are less than 0.5 and may be close to 0, which can be reasonably expressed using $a_{\pi_h} = a_\pi = 1$ and $b_{\pi_h} = b_\pi = 4$. It is also reasonable, in most cases, to anticipate that a small to moderate number of atoms are added in moving between two groups, which can be expressed by choosing a prior that assigns high probability to small values of α (e.g. $a_{\gamma_h} = a_\gamma = 1$ and $b_{\gamma_h} = b_\gamma = 1$).

2.3 *Identifiability and prior specification*

An important issue in latent variable models is the incorporation of constraints to ensure identifiability of the model from the observed data. Although this is different from formal Bayesian identifiability, it is nonetheless an appealing property for a Bayesian model. As a starting point for a discussion of identifiability, consider the expectation and covariance of $\mathbf{y}^*_{hi}$ integrating out the latent variables $\boldsymbol{\eta}_{hi}$ and $\boldsymbol{\kappa}_{hi}$

$$\begin{aligned} \mathrm{E}(\mathbf{y}^*_{hi}) &= \boldsymbol{\mu} + \boldsymbol{\lambda}\,\mathrm{E}(\eta_{hi}), \\ \mathrm{V}(y^*_{hij}) &= \lambda_j^2 \mathrm{V}(\eta_{hi}) + \psi_j^{-1}\left(\frac{\nu_j}{\nu_j - 2}\right), \quad \text{for } j = 1, \ldots, p, \\ \mathrm{cov}(y^*_{hij}, y^*_{hij'}) &= \lambda_j \lambda_{j'} \mathrm{V}(\eta_{hi}), \quad \text{for all } j \neq j'. \end{aligned} \tag{2.11}$$

The correlation coefficient between the underlying variables, y^*_{hij} and $y^*_{hij'}$, denoted $\rho_h(j, j')$, can be used as a measure of the correlation between y_{hij} and $y_{hij'}$. Clearly, there is a potential nonidentifiability problem, since the model is invariant to transformations that (i) multiply $\mathrm{V}(\eta_{hi})$ by any positive constant c_1 while dividing λ_j by $\sqrt{c_1}$ for $j = 1, \ldots, p$ or (ii) add any real number c_2 to $\mathrm{E}(\eta_{hi})$ while subtracting $\lambda_j c_2$ from μ_j for $j = 1, \ldots, p$. To eliminate this problem, we recommend fixing the values of one of the elements of both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$; say by letting $\mu_1 = 0$ and $\lambda_1 = 1$.

It is important to consider carefully the sources of information about the group-specific latent variable distributions, $G_1, \ldots, G_d$. For purposes of discussion, first consider the simple case in which $p_1 = p = 1$

and $d = 1$, so there is one continuous outcome and a single group. Since the factor loadings, $\boldsymbol{\lambda}$, characterize dependency among the outcomes, we recommend fixing $\boldsymbol{\lambda} = \mathbf{1}$ for identifiability when $p = 1$. With this constraint, there is information in the data about the shape of G_1, since the distribution of y_{1i1} is characterized as the mixture of a t-distribution across G_1. Lack of fit of the t-distribution, such as a positively skewed shape or multimodality, can be accommodated through a nonnormal mixing distribution, G_1. Extending to the $d > 1$ group case, the mixing distribution will change dynamically from G_1 to G_d across the range of the group index (e.g. across dose groups or time points). Hence, the model can accommodate systematic differences in lack of fit. For example, in the presence of heterogeneity in a dose or treatment effect, there may be increasing skewness in higher treatment groups.

Finally, considering the general case in which $p \geqslant 1$ and $d \geqslant 1$, the density of the jth surrogate in group h can be expressed as the mixture of a t-distribution with mean $\mu_j + \lambda_j \eta_h$ across the G_h-distribution for η_h:

$$f_{Y_{hj}}(y) \propto \int \{1 + \psi_j (y - \mu_j - \lambda_j \eta_h)^2 / \nu_j\}^{-(\nu_j+1)/2} \, \mathrm{d}G_h(\eta_h).$$

Hence, the marginal distribution of Y_{hj} will be increasingly driven by the characteristics of G_h as the factor loading λ_j and the correlation with the other surrogates increase. If the surrogates in group h tend to be positively skewed or to have other characteristics inconsistent with the t-distribution, these features will be reflected in G_h. Common features of the surrogate distributions which tend to change across the groups will be reflected in differences among $G_1, G_2, \ldots, G_d$. In this manner, the data clearly inform about the latent variable distributions $G_1, \ldots, G_d$.

A Bayesian specification of the model is completed with priors for the threshold parameters $\boldsymbol{\tau}$ and additional unknowns in the measurement model (2.1). Letting $\mathbf{U}_{hi} = [\mathbf{I}_p \ \eta_{hi} \mathbf{I}_p]$ and $\boldsymbol{\gamma} = (\boldsymbol{\mu}', \boldsymbol{\lambda}')'$, Expression (2.1) is equivalent to $\mathbf{y}^*_{hi} = \mathbf{U}_{hi} \boldsymbol{\gamma} + \boldsymbol{\epsilon}_{hi}$. Following Albert and Chib (1993), we choose a uniform improper prior for the threshold parameters, $\boldsymbol{\tau}_j$, for all $j \in \{p_1 + 1, \ldots, p\}$ such that $d_j > 2$, with the τ values known for $d_j = 2$. For the remaining parameters, our prior can be expressed as follows:

$$\pi(\boldsymbol{\gamma}) \propto \mathrm{N}(\boldsymbol{\gamma}_0, \Sigma_\gamma) 1(\lambda_j > 0, j = 1, \ldots, p), \quad \pi(\nu_j) \stackrel{\mathrm{d}}{=} \mathcal{G}(a_\nu, b_\nu), \quad \pi(\psi_j) \stackrel{\mathrm{d}}{=} \mathcal{G}(a_\psi, b_\psi). \tag{2.12}$$

We choose the first and $(p + 1)$st diagonal elements of Σ_γ to be ≈ 0 in effect fixing μ_1 and λ_1, for identifiability purposes. This is done to simplify book-keeping in developing the computational algorithm, and yields essentially identical results to treating μ_1 and λ_1 as strictly fixed constants. Focusing on the case in which the surrogates all have the same direction, we constrain $\lambda_j > 0$, though this sign restriction is not necessary for identifiability.

3. Posterior computation and inferences

This section outlines a Markov chain Monte Carlo algorithm for posterior computation and predictive inference. In the absence of outside information about G_1 and systematic changes that occur across groups, a natural choice for $G_0, H_{01}, \ldots, H_{0,d-1}$ is the standard normal distribution. This choice results in a semiparametric model, which is centered on a parametric model having normally distributed factors. Given the focus in the literature on normal latent variable models, this form is particularly appealing and will be our focus in developing a computational algorithm. The form of the algorithm is motivated by efficiency considerations, and we utilize approaches for efficient sampling in DP models while also using a block-updating approach for the unknowns in the measurement model. The steps involved in these two components are described separately in the following two subsections. We focus on the case in which all the surrogates are continuous since the extension to the general case is straightforward using the Albert and Chib (1993) approach.

3.1 *Updating the unknowns in the latent response model*

In this section, we describe our algorithm for updating the latent variables $\boldsymbol{\eta}_h$, the mixture weights π_h, and precisions α_h, for $h = 1, \ldots, d$, integrating out the infinite-dimensional $\{G_1, \ldots, G_d\}$. Our approach is related to the Pólya urn Gibbs sampler described by MacEachern (1994) and West *et al.* (1994). Complications arise due to the DMDP structure of Expression (2.6), but most of these are alleviated by using the characterization of Expression (2.7).

Let $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{l,k_l})'$ denote the unique values of the latent variable in the lth mixture component $\boldsymbol{\xi}_l = \{\eta_{hi}\colon M_{hi} = l, h = 1, \ldots, d, i = 1, \ldots, n_h\}$, and let $\mathcal{S}_{hi} = (l, r)$ denote that $M_{hi} = l$ and $\eta_{hi} = \theta_{lr}$, so that subject h, i belongs to the rth cluster in the lth mixture component, with $\mathbf{S} = \{\mathcal{S}_{hi}, h = 1, \ldots, d, i = 1, \ldots, n_h\}$. Let m_l and m_{lr} denote the total number of subjects having $M_{hi} = l$ and $\mathcal{S}_{hi} = (l, r)$, respectively. Also, let $\boldsymbol{\theta}_l^{(hi)}$, $k_l^{(hi)}$, $\mathbf{S}^{(hi)}$, $m_l^{(hi)}$, and $m_{lr}^{(hi)}$ denote the values obtained excluding subject h, i. Then, the conditional prior distribution of η_{hi} given the latent variable values and mixture component indicators for all other subjects is

$$\sum_{l=1}^{h} \omega_{hl} \left[\left(\frac{\alpha_l}{\alpha_l + m_l^{(hi)}} \right) G_{0l}^* + \sum_{r=1}^{k_l^{(hi)}} \left(\frac{m_{lr}^{(hi)}}{\alpha_l + m_l^{(hi)}} \right) \delta_{\theta_{lr}^{(hi)}} \right]. \tag{3.1}$$

We first derive the conditional posterior distribution of η_{hi}, updating this prior with the data. Introducing shorthand notation, let w_{hl0} and w_{hlr} denote the respective multipliers on G_{0l}^* and $\delta_{\theta_{lr}^{(hi)}}$ in Expression (3.1). Multiplying the conditional prior (3.1) by the conditional likelihood, $\prod_{j=1}^{p} \mathrm{N}(\tilde{y}_{hij}; \lambda_j \eta_{hi}, \sigma_{hij}^2)$, with $\tilde{y}_{hij} = y_{hij} - \mu_j$, and normalizing results in the following full conditional posterior distribution for η_{hi}:

$$\tilde{w}_{hl0}\, \mathrm{N}(\tilde{\eta}_{hi}, \tilde{V}_{\eta_{hi}}) + \sum_{l=1}^{h} \sum_{r=1}^{k_l^{(hi)}} \tilde{w}_{hlr} \delta_{\theta_{lr}^{(hi)}}, \tag{3.2}$$

where $\tilde{V}_{\eta_{hi}} = (1 + \sum_{j=1}^{p} \sigma_{hij}^{-2} \lambda_j^2)^{-1}$ and $\tilde{\eta}_{hi} = \tilde{V}_{\eta_{hi}} \sum_{j=1}^{p} \sigma_{hij}^{-2} \lambda_j \tilde{y}_{hij}$ are the conditional posterior variance and mean derived under the base parametric prior $\eta_{hi} \sim \mathrm{N}(0, 1)$, the updated component weights are defined as follows:

$$\tilde{w}_{hl0} = c \cdot w_{hl0} \cdot \frac{(2\pi)^{-1/2} \prod_{j=1}^{p} \mathrm{N}(\tilde{y}_{hij}; 0, \sigma_{hij}^{-2})}{\mathrm{N}(0; \tilde{\eta}_{hi}, \tilde{V}_{\eta_{hi}})}, \quad \tilde{w}_{hlr} = c \cdot w_{hlr} \cdot \prod_{j=1}^{p} \mathrm{N}(\tilde{y}_{hij}; \lambda_j \theta_{rl}^{(hi)}, \sigma_{hij}^{-2}).$$

Computation can potentially proceed by Gibbs steps, which successively sample from the full conditional distribution (3.2) for each η_{hi}. However, since this approach results in slow mixing, we suggest an alternative approach following MacEachern (1994). In particular, letting $\mathcal{S}_{hi} = (0, l)$ if η_{hi} is allocated to a new cluster in mixture component l and $\mathcal{S}_{hi} = (r, l)$ if $\eta_{hi} = \theta_{rl}^{(hi)}$, we alternate between the following steps:

1. Update $\mathbf{S}$ by sampling each $\mathcal{S}_{hi}$ from its full conditional distribution, which is multinomial with $\Pr\{\mathcal{S}_{hi} = (r, l) | \mathbf{S}^{(hi)}, \boldsymbol{\theta}^{(hi)}, \mathbf{k}^{(hi)}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\psi}\} = \tilde{w}_{hlr}$, for $l = 1, \ldots, h$, $r = 0, 1, \ldots, h$. Whenever $\mathcal{S}_{hi} = (0, l)$, for any l, we replace η_{hi} with a draw from $\mathrm{N}(\tilde{\eta}_{hi}, \tilde{V}_{\eta_{hi}})$ to assign subject h, i to their own cluster in component l.
2. Generate new values for $\boldsymbol{\theta}_l$, for $l = 1, \ldots, d$, conditional on the current configuration of subjects to clusters by sampling θ_{lr}, for $r = 1, \ldots, k_l$, from its full conditional posterior distribution, which is

$\mathrm{N}(\widehat{\theta}_{lr}, \widehat{V}_{\theta_{lr}})$, with

$$\widehat{\theta}_{lr} = \widehat{V}_{\theta_{hl}} \sum_{h=l}^{d} \sum_{i:\mathcal{S}_{hi}=(r,l)} \sum_{j=1}^{p} \sigma_{hij}^{-2} \lambda_j \widetilde{y}_{hij} \quad \text{and} \quad \widehat{V}_{\theta_{hl}} = \left(1 + \sum_{h=l}^{d} \sum_{i:\mathcal{S}_{hi}=(r,l)} \sum_{j=1}^{p} \sigma_{hij}^{-2} \lambda_j^2\right).$$

3. Update π_l, for $l = 1, \ldots, d-1$, from its full conditional posterior distribution which is

$$\text{beta}\left(a_{\pi_l} + \sum_{h=l+1}^{d} \sum_{i=1}^{n_h} 1(M_{ih} = l+1), b_{\pi_l} + \sum_{h=l+1}^{d} \sum_{i=1}^{n_h} \sum_{m=1}^{l} 1(M_{ih} = m)\right),$$

   where $M_{ih} = m$ if $\mathcal{S}_{ih} = (r, m)$ for any r.
4. Update α_l, for $l = 1, \ldots, d$, using the procedure proposed by West (1992), noting that for updating α_l the relevant number of clusters is k_l and the relevant sample size is $m_l = \sum_{h=1}^{d} \sum_{i=1}^{n_h} 1(M_{ih} = l)$, which varies from iteration to iteration.

3.2 *Updating the unknowns in the measurement model*

Sampling of the coefficients, $\boldsymbol{\gamma}$, and unknowns, $\{\psi_j, \kappa_{ij}, \nu_j\}$, proceeds as follows:

Step 2,a. Update $\boldsymbol{\gamma}$ in a single block by sampling from the joint conditional distribution, which is $\mathrm{N}_{2p}(\widehat{\boldsymbol{\gamma}}, \widehat{\Sigma}_\gamma)$ subject to the constraint that $\lambda_j > 0$ for $j = 1, \ldots, p$, where

$$\widehat{\boldsymbol{\gamma}} = \widehat{\Sigma}_\gamma \left(\Sigma_\gamma^{-1} \boldsymbol{\gamma}_0 + \sum_{h=1}^{d} \sum_{i=1}^{n_h} \sum_{j=1}^{p} \sigma_{hij}^{-2} \mathbf{u}'_{hij} y_{hij}\right) \quad \text{and} \quad \widehat{\Sigma}_\gamma = \left(\Sigma_\gamma^{-1} + \sum_{h=1}^{d} \sum_{i=1}^{n_h} \sum_{j=1}^{p} \sigma_{hij}^{-2} \mathbf{u}'_{hij} \mathbf{u}_{hij}\right)^{-1},$$

where $\mathbf{u}_{hij}$ is the jth row vector of $\mathbf{U}_{hi}$.

Step 2,b. Update the measurement error parameters by sampling from the full conditional distribution of ψ_j (for $j = 1, \ldots, p$):

$$\mathcal{G}\left(a_\psi + \frac{n}{2}, b_\psi + \frac{1}{2} \sum_{h=1}^{d} \sum_{i=1}^{n_h} \kappa_{hij} (y_{hij} - \mathbf{u}'_{hij} \boldsymbol{\gamma})^2\right),$$

sampling from the full conditional distribution of κ_{hij} (for all h, i, j):

$$\mathcal{G}\left(\frac{\nu_j + 1}{2}, \frac{\nu_j + \psi_j (y_{hij} - \mathbf{u}'_{hij} \boldsymbol{\gamma})^2}{2}\right),$$

and finally updating the ν_j values in a Metropolis step.

3.3 *Inferences on the latent response distribution*

Posterior summaries of $\rho_h(j, j')$ calculated from the MCMC output can be used as a basis for inferences on correlation between the surrogates. However, the primary focus is typically on assessing changes in the distribution of the latent response as a function of the predictors. For this purpose, it will be useful to

Table 1. *Posterior summaries of the parameters in the DNA damage application*

Parameter	Posterior summary		
	Mean	SD	95% Credible interval
μ_1	0	0	[0, 0]
μ_2	−0.19	0.06	[−0.31, −0.07]
μ_3	−0.14	0.02	[−0.18, −0.11]
μ_4	−0.03	0.02	[−0.07, 0.00]
μ_5	−0.11	0.06	[−0.23, −0.01]
λ_1	1	0	[1, 1]
λ_2	0.16	0.04	[0.09, 0.25]
λ_3	0.84	0.02	[0.80, 0.88]
λ_4	0.97	0.02	[0.93, 1.02]
λ_5	0.12	0.04	[0.03, 0.21]
α_1	3.95	1.38	[1.73, 7.02]
α_2	4.04	1.64	[1.45, 7.95]
α_3	4.29	1.70	[1.46, 8.09]
α_4	3.24	1.55	[0.76, 6.76]
α_5	2.29	1.33	[0.32, 5.38]
π_1	0.47	0.12	[0.24, 0.72]
π_2	0.41	0.14	[0.15, 0.69]
π_3	0.32	0.13	[0.06, 0.59]
π_4	0.19	0.12	[0.03, 0.47]
$\rho_1^{\dagger}$	0.73	0.15	[0.38, 0.94]
ρ_2	0.71	0.17	[0.35, 0.96]
ρ_3	0.74	0.16	[0.39, 0.99]
ρ_4	0.86	0.13	[0.52, 1.00]

$^{\dagger}\rho_h = \text{Corr}(G_h(\mathcal{B}), G_{h-1}(\mathcal{B}))$, for $h = 1, \ldots, d-1$

have estimates of the predicted density function of η in each group, as well as measures of the magnitude, location, and weight of evidence of changes between groups. To address these goals, we recommend collecting draws from the conditional predictive distributions of η_{h,n_h+1} for a future subject in dose group h, for $h = 1, \ldots, d$. Generalizing Expression (2.4), this distribution is simply

$$g(\eta_{h,n_h+1}) = \sum_{l=1}^{h} \omega_{hl} \left[\left(\frac{\alpha_l}{\alpha_l + m_l} \right) \text{N}(\eta_{h,n_h+1}; 0, 1) + \sum_{r=1}^{k_l} \left(\frac{m_{lr}}{\alpha_l + m_l} \right) \delta_{\theta_{lr}}(\eta_{h,n_h+1}) \right]. \tag{3.3}$$

One can collect samples from this distribution after apparent convergence, along with the mean and selected percentiles. Samples can be obtained easily due to the simple mixture structure, the mean is available in closed form, and percentiles can be estimated by calculating the cdf at a dense grid of values.

After convergence, the samples of η_{h,n_h+1} represent draws from the predictive density of the latent response in group h, and inferences can be based on comparing these densities between groups. One can also estimate marginal posterior densities of differences in quantiles between groups, which is useful in summarizing group differences, as we illustrate in Section 4. In addition, by also collecting draws from the conditional distributions of the surrogate outcomes for additional subjects in each group, we can estimate predictive densities of the surrogates. By comparing these predictive densities to the empirical distributions, one can assess goodness of fit of the procedure. In particular, it is of interest to look for surrogates

which do not follow the trajectory predicted by the model, since this may suggest that the simple one-factor structure may be insufficient. Although we have focused on the one-factor case throughout this article, the generalization to multiple factors is straightforward.

4. Application to DNA damage study

We illustrate the methodology using data from a genotoxicity experiment analyzed previously by Dunson *et al.* (2003). The study assessed the effect of oxidative stress, induced by hydrogen peroxide exposure, on the frequency of DNA strand breaks using single-cell gel electrophoresis. Human lymphoblastoid cells ($n = 500$) drawn from an immortalized cell line were randomized to one of the five dose groups (0, 5, 20, 50, or 100 micromoles H_2O_2), resulting in 100 cells per dose group. For each cell ($i = 1, \ldots, 500$), we have $p = 5$ surrogate measures of DNA damage, including (1) % tail DNA, (2) tail extent divided by head extent, (3) extent tail moment, (4) Olive tail moment, and (5) tail extent. For a detailed description of these variables and diagnostics demonstrating nonnormality, refer to Dunson *et al.* (2003).

Letting $h = 1, \ldots, 5$ index the dose group and normalizing each of the surrogates, $\mathbf{y}_i = (y_{i1}, \ldots, y_{i5})'$, prior to analysis, we apply the approach proposed in Sections 2 and 3 to assess the effect of hydrogen peroxide exposure on the frequency of DNA strand breaks. To complete a Bayesian specification of the model, we let $\boldsymbol{\gamma}_0 = (0, 0, 0, 0, 0, 1, 1, 1, 1, 1)'$, $\sigma_{\gamma,j}^{-2} = 10^{10}$ for $j = 1, 6$, and $\sigma_{\gamma,j}^{-2} = 2.0$ for $j = 2, 3, 4, 5, 7, 8, 9, 10$. A prior mean of 0 for the μ_j values seems reasonable since the surrogates are normalized. The prior for the factor loadings expresses our belief in a moderate degree of dependency in the surrogates. In addition, we let $a_\nu = 10.0$ and $b_\nu = 1.0$, since we anticipate heavier tails than normal, and we let $a_\psi = 5.0$ and $b_\psi = 1.0$ to express our belief in the magnitude of the residual variance component. Finally, for the hyperparameters in the priors for $\boldsymbol{\pi}$ and $\boldsymbol{\alpha}$ in Expression (2.10), we follow the recommendation of Section 2.2.

We ran the MCMC algorithm for 50 000 iterations, discarding the first 5000 iterations as a burn-in, and collecting every 10th sample to thin the chain. Based on examination of trace plots, convergence was rapid and autocorrelation was low to moderate, reflecting good computational efficiency of our MCMC algorithm under the constraints imposed on the model to ensure identifiability. Posterior summaries are provided in Table 1.

Applying the approach described in Section 3.3, we estimated the predictive distribution of the latent response variable η_{hi} for cells in each of the dose groups. The results are shown in Figure 1. Interestingly, the densities change considerably in shape as dose increases, with the density in the control group having an approximately log-normal shape with relatively low variance. As dose increases, the mean and variance increase substantially, the distribution flattens out, and the right tail becomes increasingly fat. Figure 2 shows boxplots for samples from the posterior distributions for changes in the 10th, 25th, 50th, 75th, and 90th percentiles of the predictive distribution for η_{hi} attributable to increasing H_2O_2 exposure from 0 ($h = 1$) to the maximum dose of 100 ($h = 5$). The differences clearly increase as one moves further into the right tail, reflecting failure of a mean or median regression model. This pattern likely reflects heterogeneity among the cells in their response to oxidative stress induced by hydrogen peroxide exposure, with some cells having little or no induced damage. The ability to make inferences on changes in quantiles of the response distribution is an appealing feature of our approach.

It is important to assess how well the model fits the data since we make some parametric assumptions and use informative priors. In particular, we assume that a single latent response variable can be used to characterize departures from the t-distribution and effects of hydrogen peroxide exposure on each of the surrogate variables. To assess model fit, we estimated predictive distributions for each of the surrogate variables at each dose level and compared these estimates to histograms of the raw data. The results for the first four surrogates are included in Figure 3, with the fifth excluded to make the plot readable (though the results are similar).

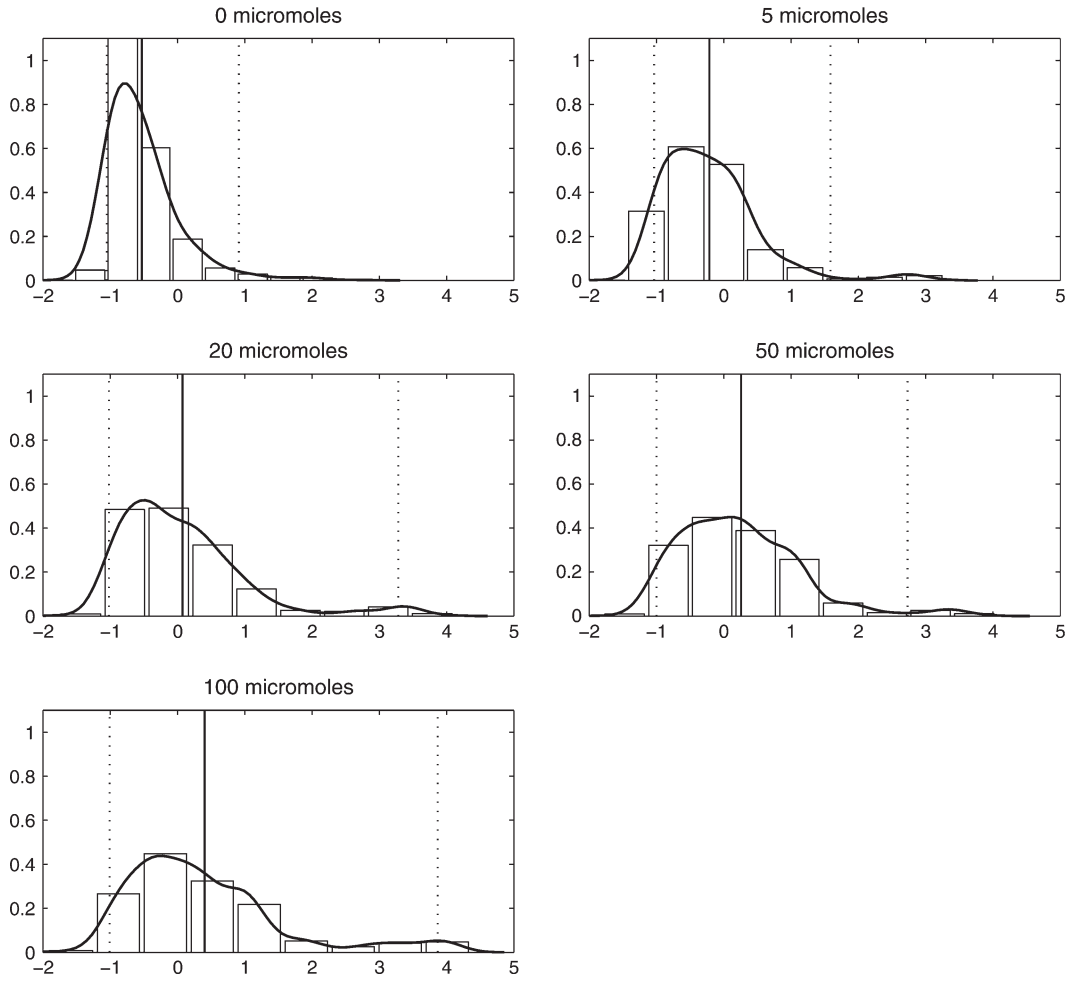

Fig. 1. Histogram and kernel smoothed density estimates for the latent DNA damage variable (η_i) among cells in each of the dose groups. The solid vertical lines represent the mean and the dashed lines represent 2.5th and 97.5th percentiles.

Overall, the model-based predictive distributions do a good job at capturing the distributional shifts that occur as dose increases. The best fit is for the Olive tail moment, which was recommended by Dunson *et al.* (2003) as the best single surrogate of DNA damage. In contrast, the fit is not as good for tail extent/head extent, which is known to be a poor surrogate due to sensitivity to individual pixels in the tail of the image. At 100 micromoles, there is some evidence of a second mode in the right tail which is picked up in Figure 1 but not in Figure 3, possibly due to over-smoothing by the t-kernel.

An important issue is sensitivity of the results to the choice of hyperparameters. To assess robustness, we repeated the analysis using alternative priors with (i) $a_\pi = 5$ and $b_\pi = 1$ to correspond to lower autocorrelation across dose groups, (ii) $a_\alpha = 5$ to correspond to less uncertainty in the normal base distribution and a higher rate of adding atoms between dose groups, (iii) the prior variance doubled for all the parameters in the measurement model, and (iv) the prior precision doubled for all the parameters in the measurement model. In each of these cases, the figures were essentially identical to those obtained in the primary analysis. Estimates of quantiles of the predicted latent variable densities in each dose group

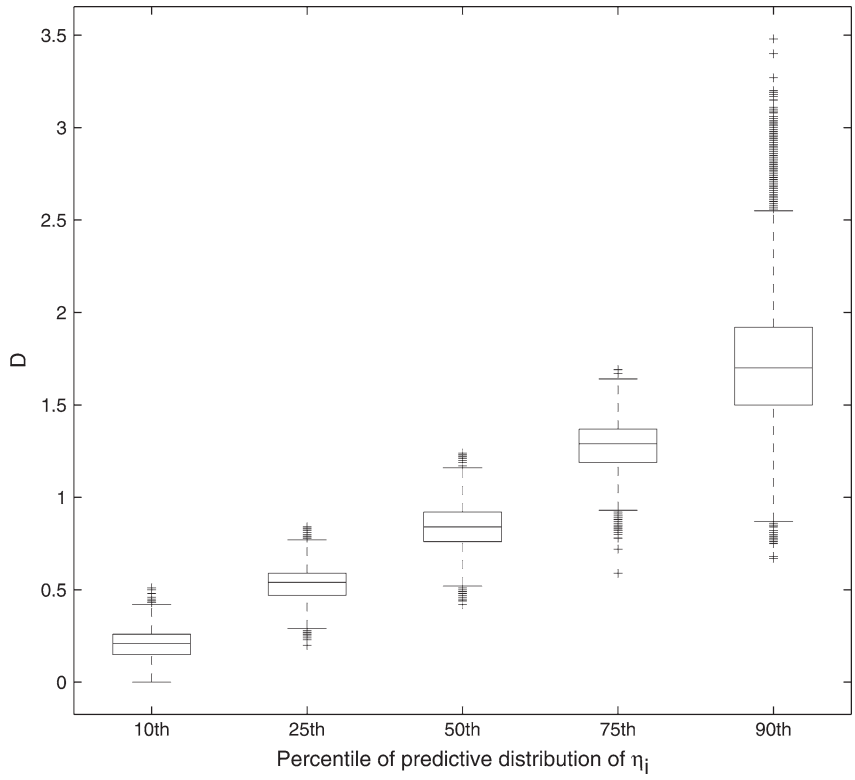

Fig. 2. Boxplots for samples from the posterior distribution of differences in the 10th, 25th, 50th, 75th, and 90th percentiles between the latent response density for $x_i = 0$ and $x_i = 100$. D = difference in quantiles between the conditional distribution of $\eta_i|x_i = 100$ and $\eta_i|x_i = 0$.

under priors (i) and (ii) are shown in Table 2; results for priors (iii) and (iv) differed by only 1.6%, on average, from the primary analysis estimates and are not shown.

5. Discussion

This article has proposed a Bayesian semiparametric latent response model in which the latent variable density can shift dynamically across groups. Although inferences on covariate effects on the mean response profile may be somewhat robust to the parametric form for the latent variable density, linear mean regression structures may be insufficiently flexible in many applications. For example, this is often the case in epidemiologic and toxicologic studies in which there may be increasing variance and skewness at higher exposure levels. The genotoxicity application presents one striking example of this scenario, though we anticipate many other applications in which the latent linear mean regression model fails. Our proposed Bayesian approach is quite flexible, and should provide a useful alternative to existing methods, such as the semiparametric median regression modeling approach of Kottas and Gelfand (2001).

The proposed DMDP should prove useful in other applications in which a distribution or random function can change across levels of a predictor. Note that one can either apply the DMDP directly to the distribution of interest, in order to model the distribution as discrete with an unknown number and location of atoms, or use the DMDP for a mixture distribution, in order to characterize a continuous density.

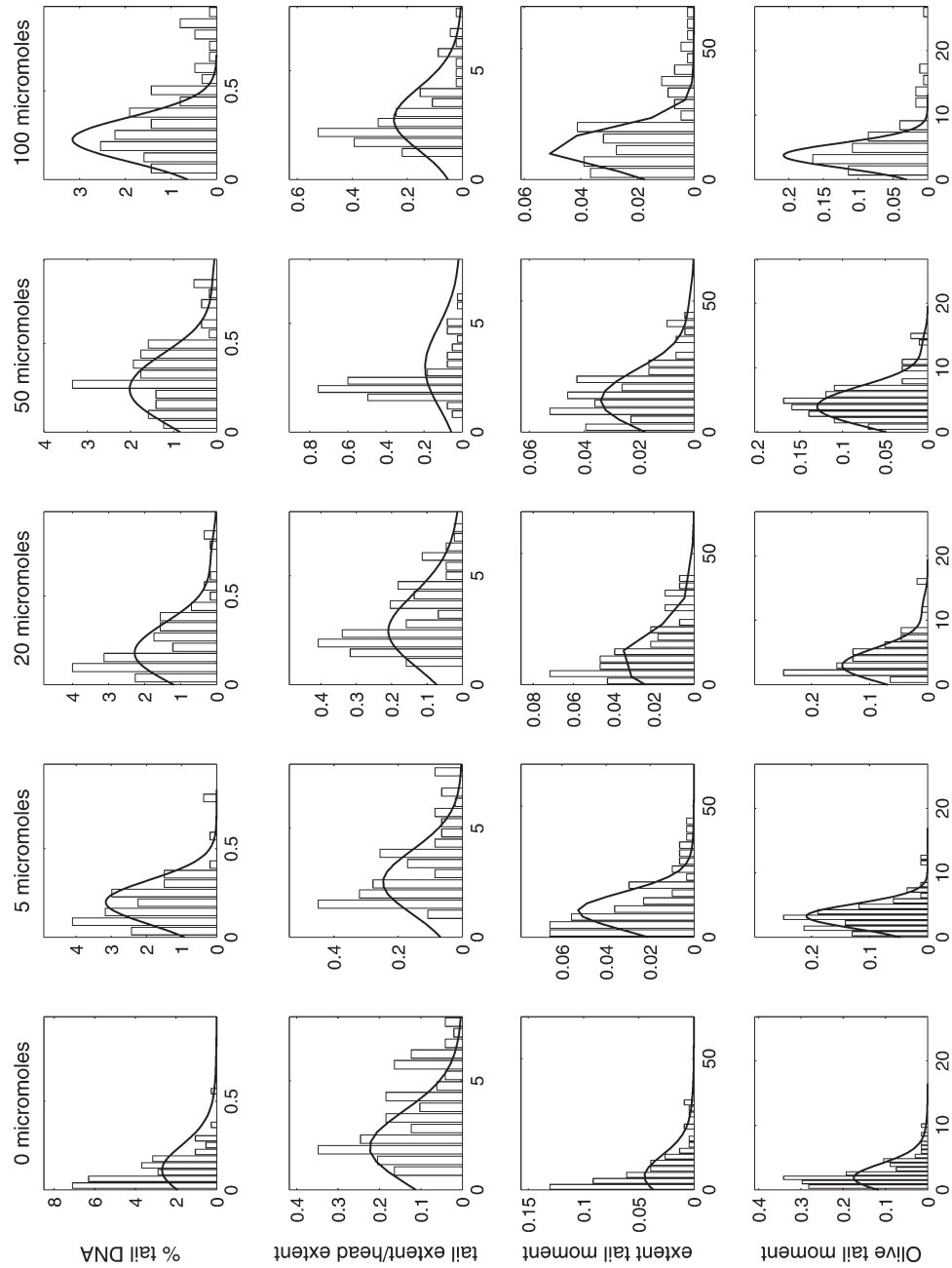

Fig. 3. Empirical histograms and estimated predictive densities for different surrogate variables and dose levels.

Table 2. *Estimated percentiles of the predictive distribution of the latent response for cells in each dose group in the DNA damage application*

Dose	Percentile	Posterior Summary				
		Mean†				
		Main	Prior (i)	Prior (ii)	SD	95% Credible interval
0	10	−0.99	−1.00	−0.98	0.05	[−1.08, −0.90]
0	25	−0.93	−0.97	−0.93	0.06	[−1.03, −0.79]
0	50	−0.68	−0.70	−0.68	0.10	[−0.86, −0.48]
0	75	−0.35	−0.39	−0.38	0.10	[−0.53, −0.14]
0	90	0.16	0.14	0.13	0.14	[−0.15, 0.40]
5	10	−0.97	−0.95	−0.96	0.05	[−1.05, −0.86]
5	25	−0.77	−0.79	−0.76	0.08	[−0.93, −0.61]
5	50	−0.34	−0.31	−0.33	0.08	[−0.48, −0.17]
5	75	0.14	0.13	0.13	0.08	[−0.02, 0.28]
5	90	0.54	0.51	0.50	0.17	[0.24, 0.91]
20	10	−0.90	−0.83	−0.90	0.07	[−1.01, −0.76]
20	25	−0.58	−0.60	−0.55	0.08	[−0.74, −0.43]
20	50	−0.11	−0.16	−0.08	0.10	[−0.31, 0.10]
20	75	0.48	0.50	0.50	0.11	[0.26, 0.70]
20	90	1.07	1.04	1.06	0.14	[0.81, 1.35]
50	10	−0.82	−0.77	−0.83	0.07	[−0.96, −0.69]
50	25	−0.43	−0.34	−0.43	0.07	[−0.56, −0.30]
50	50	0.14	0.23	0.14	0.08	[−0.01, 0.29]
50	75	0.81	0.88	0.84	0.12	[0.56, 1.03]
50	90	1.23	1.21	1.21	0.18	[1.01, 1.80]
100	10	−0.79	−0.74	−0.80	0.07	[−0.93, −0.65]
100	25	−0.40	−0.37	−0.40	0.06	[−0.52, −0.26]
100	50	0.16	0.15	0.17	0.07	[0.01, 0.31]
100	75	0.93	0.96	0.95	0.09	[0.72, 1.09]
100	90	1.91	2.09	1.95	0.40	[1.26, 2.96]

†Main = main analysis; (i) = prior with $a_\pi = 5, b_\pi = 1$; (ii) = prior with $a_\alpha = 5, b_\alpha = 1$

The DMDP should also prove useful when interest focuses on clustering of observations within and across groups. For example, in a time course or dose response gene expression study, one may want to cluster genes having similar levels of differential expression, both within a given time or dose group and across times or doses.

An interesting area for future research is the generalization to broader classes of factor analytic and structural equation models, allowing for uncertainty in the number of factors as in Lopes and West (2004) for the normal linear factor model. It is possible that fewer factors may be needed to characterize the covariance structure if one allows the factors to have nonparametric distributions. However, issues of interpretation and identifiability need to be carefully considered. Potentially, a rich class of multivariate distributions could be generated by including a few factors having unknown distributions. This idea is conceptually related to models based on mixtures of factor analyzers (Utsugi and Kumagai, 2001; Fokoue and Titterington, 2003; McLachlan *et al.*, 2003), though previous methods have assumed fully parametric mixture structures.

ACKNOWLEDGMENTS

The author thanks Mary Watson and Jack Taylor for generously providing the data used in the example. Thanks also to Alan Gelfand, David Umbach, and Gregg Dinse for their helpful comments on an early draft of the manuscript.

APPENDIX

A.1 *Characterization of correlation in the unknown distributions*

From (2.6) and (2.7), it is straightforward to derive $\mathrm{Corr}(G_h(\mathcal{B}), G_{h-1}(\mathcal{B}))$:

$$\mathrm{Corr}(G_h(\mathcal{B}), G_{h-1}(\mathcal{B})) = \frac{\mathrm{E}\{G_h(\mathcal{B})G_{h-1}(\mathcal{B})\} - \mathrm{E}\{G_h(\mathcal{B})\}\mathrm{E}\{G_{h-1}(\mathcal{B})\}}{[\mathrm{V}\{G_h(\mathcal{B})\}\mathrm{V}\{G_{h-1}(\mathcal{B})\}]^{1/2}}. \tag{A.1}$$

The expectation and variance terms are as follows:

$$\mathrm{E}\{G_h(\mathcal{B})\} = \sum_{l=1}^{h} \omega_{hl} G_{0l}^*(\mathcal{B}) \quad \text{and} \quad \mathrm{V}\{G_h(\mathcal{B})\} = \sum_{l=1}^{h} \left(\frac{\omega_{hl}^2}{\alpha_l + 1}\right) G_{0l}^*(\mathcal{B})\{1 - G_{0l}^*(\mathcal{B})\}, \tag{A.2}$$

for $h = 1, \ldots, d$. The numerator in Expression (A.1) can be calculated as follows:

$$\begin{aligned}
&\mathrm{E}\{G_h(\mathcal{B})G_{h-1}(\mathcal{B})\} - \mathrm{E}\{G_h(\mathcal{B})\}\mathrm{E}\{G_{h-1}(\mathcal{B})\} \\
&\quad = \mathrm{E}[\{\omega_{h1}G_1^*(\mathcal{B}) + \cdots + \omega_{hh}G_h^*(\mathcal{B})\}\{\omega_{h-1,1}G_1^*(\mathcal{B}) + \cdots + \omega_{h-1,h-1}G_{h-1}^*(\mathcal{B})\}] \\
&\qquad - \mathrm{E}\{\omega_{h1}G_1^*(\mathcal{B}) + \cdots + \omega_{hh}G_h^*(\mathcal{B})\}\mathrm{E}\{\omega_{h-1,1}G_1^*(\mathcal{B}) + \cdots + \omega_{h-1,h-1}G_{h-1}^*(\mathcal{B})\} \\
&\quad = \sum_{l=1}^{h-1} \omega_{hl}\omega_{h-1,l}[\mathrm{E}\{G_l^*(\mathcal{B})^2\} - \mathrm{E}\{G_l^*(\mathcal{B})\}^2] \\
&\quad = \sum_{l=1}^{h-1} \omega_{hl}\omega_{h-1,l}\mathrm{V}\{G_l^*(\mathcal{B})\} \\
&\quad = \sum_{l=1}^{h-1} \left(\frac{\omega_{hl}\omega_{h-1,l}}{\alpha_l + 1}\right) G_{0l}^*(\mathcal{B})\{1 - G_{0l}^*(\mathcal{B})\}.
\end{aligned} \tag{A.3}$$

Hence, focusing on the special case in which $G_{0l}^*(\mathcal{B}) = G_0^*(\mathcal{B})$, for $l = 1, \ldots, d$, Expression (2.8) follows from straightforward algebra.

REFERENCES

ALBERT, J. H. AND CHIB, S. (1993). Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

BUSH, C. A. AND MACEACHERN, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. AND MACEACHERN, S. N. (2002). ANOVA DDP models: a review. In Denison, D. D., Hansen, M. H., Holmes, C. C., Mallick, B. and Yu, B. (eds), *Nonlinear Estimation and Classification*. New York: Springer, p. 467.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. AND MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

DUNSON, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B* **62**, 355–366.

DUNSON, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association* **98**, 555–563.

DUNSON, D. B., WATSON, M. AND TAYLOR, J. A. (2003). Bayesian latent variable models for median regression on multiple outcomes. *Biometrics* **59**, 296–304.

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.

FERNANDEZ, C. AND STEEL, M. F. J. (2001). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory* **16**, 80–101.

FOKOUE, E. AND TITTERINGTON, D. M. (2003). Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation. *Machine Learning* **50**, 73–94.

GELFAND, A. E. AND KOTTAS, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **11**, 289–305.

GELFAND, A. E., KOTTAS, A. AND MACEACHERN, S. N. (2004). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Technical Report AMS 2004-5*. Department of Applied Math and Statistics, University of California, Santa Cruz.

GRIFFIN, J. E. AND STEEL, M. F. J. (2006). Order-based dependent Dirichlet process. *Journal of the American Statistical Association* (in press).

ISHWARAN, H. AND JAMES, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

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.

JOHNSON, V. E. AND ALBERT, J. H. (1999). *Ordinal Data Modeling*. New York: Springer.

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.

LOPES, H. F. AND WEST, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

MACEACHERN, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **23**, 727–741.

MACEACHERN, S. N. (1999). *Dependent nonparametric process*, ASA Proceeding of the Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association.

MACEACHERN, S. N. (2000). *Dependent Dirichlet processes* (unpublished). Department of Statistics, The Ohio State University.

MCLACHLAN, G. J., PEEL, D. AND BEAN, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* **41**, 379–388.

MIGLIORETTI, D. L. (2003). Latent transition models for mixed outcomes. *Biometrics* **59**, 710–720.

MOUSTAKI, I. AND KNOTT, M. (2000). Generalized latent trait models. *Psychometrika* **65**, 391–411.

MUKHOPADHYAY, S. AND GELFAND, A. E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

MÜLLER, P. AND QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

MÜLLER, P., QUINTANA, F. A. AND ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B* **66**, 735–749.

MUTHÉN, D. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**, 115–132.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture. *Journal of Computational and Graphical Statistics* **9**, 249–265.

REBOUSSIN, B. A., LIANG, K. Y. AND REBOUSSIN, D. M. (1999). Estimating equations for a latent transition model with multiple discrete indicators. *Biometrics* **55**, 839–845.

ROY, J. AND LIN, X. H. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* **56**, 1047–1054.

ROY, J. AND LIN, X. H. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropout and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association* **97**, 40–52.

ROY, J., LIN, X. H. AND RYAN, L. M. (2003). Scaled marginal models for multiple continuous outcomes. *Biostatistics* **4**, 371–383.

SAMMEL, M. D. AND RYAN, L. M. (2002). Effects of covariance misspecification in a latent variable model for multiple outcomes. *Statistica Sinica* **12**, 1207–1222.

SAMMEL, M. D., RYAN, L. M. AND LEGLER, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society B* **59**, 667–678.

SETHURAMAN, J. (1994) A constructive definition of the Dirichlet process prior. *Statistica Sinica* **2**, 639–650.

SHI, J. Q. AND LEE, S. Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, Series B* **62**, 77–87.

UTSUGI, A. AND KUMAGAI, T. (2001). Bayesian analysis of mixtures of factor analyzers. *Neural Computation* **13**, 993–1002.

WEST, M. (1987). On scale mixtures of normal-distributions. *Biometrika* **74**, 646–648.

WEST, M. (1992). Hyperparameter estimation in Dirichlet process mixture model. *ISDS Discussion Paper No. 92-03*, Duke University, Durham, NC.

WEST, M., MÜLLER, P. AND ESCOBAR, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In Smith, A. F. M. and Freeman, P. R. (eds), *Aspects of Uncertainty: A Tribute to D.V. Lindley*. London: Wiley.

XU, J. AND ZEGER, S. L. (2001). The evaluation of multiple surrogate endpoints. *Biometrics* **57**, 81–87.

[*Received December 4, 2004; first revision June 7, 2005; second revision December 4, 2005; third revision January 25, 2006; accepted for publication February 15, 2006*]