

Semiparametric Bayesian latent trajectory models

David B. Dunson¹ and Amy H. Herring²

¹Biostatistics Branch, MD A3-03

National Institute of Environmental Health Sciences

P.O. Box 12233, RTP, NC 27709

²Department of Biostatistics

The University of North Carolina at Chapel Hill

Summary Latent trajectory models (LTMs) characterize longitudinal data using a finite mixture of curves. We address uncertainty in the number of latent classes and in the form of the class-specific curves using a semiparametric Bayesian approach. A mixture of functional Dirichlet processes (FDP) is used to characterize the distribution of longitudinal trajectories. The FDP is defined by replacing the atoms in the stick-breaking representation of a Dirichlet process with random functions. Based on the FDP, subjects are automatically clustered into an unknown number of groups based on their latent trajectories. To allow joint nonparametric modeling with a multivariate response, we generalize the FDP to a class of joint FDPs (JFDP). The proposed approach allows the response distribution to be unknown and varying with trajectory class. An MCMC algorithm is developed for posterior computation. The methods are motivated by an epidemiologic study of water quality and pregnancy outcomes.

Key Words: Dependent Dirichlet process; Dynamic factor model; Functional data; Gaussian process; Joint model; Latent class, Latent trajectory; Nonparametric Bayes.

1. Introduction

In longitudinal data analysis, a common focus is characterization of the distribution of trajectories across time among individuals. Widely used Gaussian linear mixed effects models (cf. Laird and Ware, 1982) may be insufficiently flexible in assuming a known parametric form for the mean function as well as normally distributed random effects. A rich body of literature has focused on relaxing these assumptions by allowing a nonparametric mean function (Rice and Wu, 2001; Wu and Zhang, 2002; Zhang, 2004) and/or a nonparametric distribution for the random effects (Davidian and Gallant, 1993; Bush and MacEachern, 1996; Kleinman and Ibrahim, 1998; Ishwaran and Takahara, 2002; Burr and Doss, 2005). Methods that cluster the longitudinal trajectories into groups, with the group status unknown, provide a useful dimensionality reduction technique and aid in interpreting results.

Latent class trajectory models (Muthén and Shedden, 1999; Lin et al., 2000; Muthén et al., 2002; Elliott et al., 2005) provide a useful approach. These models combine latent class and random effects models, assuming individuals can be grouped into a finite number of classes having distinct random effects. Such approaches can be used for joint modeling of longitudinal predictor data with a response by including the latent class indicators in the outcome model (Lin et al., 2002). In addition, a number of modifications are possible, such as allowing the latent class status to change dynamically with time (Miglioretti, 2003).

Although latent class trajectory models are very flexible, difficult issues include choice of the number of latent classes and selection of trajectory models within each class. The typical strategy fixes the number of classes in advance at a small value, such as 2-4, assessing goodness-of-fit using a criteria, such as the BIC or AIC, frequentist diagnostics (Formann, 2003) or graphical posterior checks (Garrett and Zeger, 2000). The class probabilities are modeled using a multinomial response model, while the trajectories are modeled parametrically, say with a polynomial function.

This article proposes a more flexible semiparametric Bayesian approach. Viewing the

trajectories as random functions, we treat the distribution of trajectories as unknown using a functional Dirichlet process (FDP). The FDP defines a random probability measure with support on a function space by replacing the atoms in the Sethuraman (1994) stick breaking representation of the Dirichlet process (DP) (Ferguson, 1973; 1974) with random functions generated from a Gaussian process. A closely related formulation to the FDP, the dependent DP (DDP), was used by MacEachern (1999; 2001) to define dependency in a collection of random probability measures. The DDP has been used to induce ANOVA-type dependency structures (De Iorio et al., 2004) and to define a nonparametric spatial process (Gelfand et al. 2005).

The longitudinal data for a subject are assumed to arise from the convolution of a smooth latent trajectory with a noisy Gaussian process residual. By assuming an FDP prior for the distribution of latent trajectories, we can automatically cluster subjects into an unspecified number of latent classes, with the class-specific curves treated nonparametrically. This formulation is then generalized for joint nonparametric modeling of a longitudinal predictor with a multivariate response variable. For example, in the application motivating this work, interest focuses on relating the trajectory in a time-varying exposure in pregnancy to the joint distribution of duration of gestational and birth weight.

An alternative infinite mixture of Gaussian processes was proposed by Rasmussen and Ghahramani (2002). Bigelow and Dunson (2005) considered a different strategy for nonparametric Bayesian clustering of functional data. They used a DP applied to the random effects distribution in a hierarchical multivariate adaptive spline model, with reversible jump MCMC (Green, 1995) used to allow uncertainty in the basis functions. By avoiding the need to select basis functions, our approach should have advantages in terms of computational speed and interpretability.

Section 2 describes the motivating application to drinking water disinfection by-products and pregnancy outcomes. Section 3 describes the model for the latent trajectories and pro-

vides background on the FDP. Section 4 considers joint modeling of longitudinal trajectories with a multivariate response. Section 5 outlines an MCMC algorithm for posterior computation. Section 6 contains simulated data examples, Section 7 applies the approach to the water quality example, and Section 8 discusses the results.

2. Motivating Application

Epidemiologists often study the relationship between a time-varying predictor, such as an environmental exposure, and one or more health outcomes. For example, in the Right from the Start (RFTS) study (Promislow et al., 2004), interest centered on the relationship between disinfection by-products (DBPs) in the water in early pregnancy and later outcomes, such as gestational age at delivery and birth weight. DBPs include a variety of chemicals formed when organic matter interacts with disinfection agents added to the water. For illustration, we focus on the DBP bromodichloromethane (BDCM). Figure 1 plots the observed data for 10 randomly-selected women from among the 1742 women in the study.

The BDCM levels tend to oscillate up and down over the weeks, leading to a variety of trajectories for women, with some women having low levels all of the time. A difficult issue is how to model the effects of these data on pregnancy outcomes, including gestational age at delivery in weeks (GAD) and birth weight in grams (BW). A common approach is to average BDCM levels within a variety of time windows corresponding to different stages of fetal development. However, reproductive epidemiologists are uncertain about what aspect of the trajectory is most predictive of pregnancy outcomes if any. In addition, the joint distribution of GAD and BW may shift in unanticipated ways according to the trajectory. Figure 2 plots the observed GAD and BW values for the 1742 women under study. Standard parametric models are known to provide a poor fit, generating considerable controversy in the epidemiologic literature about how to analyze GAD and BW. The typical approach is to separately analyze indicators of preterm birth (GAD dichotomized using a 37 week cutoff)

and small for gestational age (BW dichotomized using the 10th percentile of the population distribution stratified by GAD, race, gender and mother’s parity).

The approach taken in this article is to cluster the BDCM trajectories into latent classes. For example, the 10 women in Figure 1 could possibly be clustered into 4 classes: (1) a flat trajectory close to zero (3 women); (2) a slowly increasing trajectory starting around 12 $\mu\text{g/L}$ (5 women); (3) a steadily decreasing trajectory starting around 31 $\mu\text{g/L}$ (1 woman); and (4) a decreasing trajectory starting at 25 $\mu\text{g/L}$ which flattens out rapidly (1 woman). The cluster indicators can then be included as predictors in a joint model for GAD and BW. The goal of this article is to develop a Bayesian nonparametric approach, which automatically allocates the trajectories into an unspecified number of classes, with the response density varying nonparametrically across trajectory classes.

3. Latent Trajectory Models

3.1 *Mixtures of Gaussian Processes*

Let y_i denote a continuous-time stochastic process $\{y_i(t), t \in \mathfrak{R}^+\}$ specific to subject i , for $i = 1, \dots, n$. In a longitudinal study, observations are collected on subject i at times $\mathbf{t}_i = (t_{i1}, \dots, t_{ini})'$, so that the observed data consist of the vector $\mathbf{y}_i = y_i(\mathbf{t}_i) = [y_i(t_{i1}), \dots, y_i(t_{ini})]'$. The collection of random functions $\{y_i, i = 1, \dots, n\}$ for the different individuals are initially assumed to arise from the following process:

$$y_i = \eta_i + \epsilon_i, \quad \eta_i \sim G, \quad \epsilon_i \sim GP(\mathcal{C}_\nu), \quad (1)$$

where η_i is a latent trajectory curve drawn from G and ϵ_i is a residual function drawn from a Gaussian process, with covariance function \mathcal{C}_ν parameterized in terms of the finite-dimensional parameter ν . We assume that η_i and ϵ_i are independent.

To allow for uncertainty in the distribution of latent trajectories, we assume that G is a random probability measure over (Ω, \mathcal{B}) , with Ω a space of functions and \mathcal{B} the Borel σ -algebra of subsets of Ω . We initially consider the case in which G is characterized as a

finite mixture of k latent trajectories, with

$$\eta_i \sim G(\cdot) = \sum_{h=1}^k p_h \delta_{\Theta_h}(\cdot), \quad \Theta_h \sim GP(\mathcal{C}_{\kappa_h}), \quad (2)$$

with $\mathbf{p} = (p_1, \dots, p_k)'$ mixture weights summing to one, δ_{Θ} denoting the Dirac probability measure that assigns one to Θ and zero to all subsets of Ω not containing Θ , and $\Theta = (\Theta_1, \dots, \Theta_k)'$ is a collection of k latent trajectories, which are assigned Gaussian process priors. The covariance function \mathcal{C}_{κ_h} can be chosen based on the desired properties of Θ_h .

For example, suppose that $\Theta_h(t) = a_h + b_h t$, so that the latent trajectories η_i are linear for all subjects, but the intercept and slope vary independently depending on the latent class membership. By placing a distribution on the coefficients, $a_h \sim N(0, \kappa_{h1})$ and $b_h \sim N(0, \kappa_{h2})$, we can consider the linear function Θ_h as random. Then, following Rasmussen (1996), we have

$$\begin{aligned} \mathbb{E}\{\Theta_h(t)\} &= \int \int \Theta_h(t) N(a_h; 0, \kappa_{h1}) N(b_h; 0, \kappa_{h2}) da_h db_h = 0 \\ \mathcal{C}_{\kappa_h}(t, t') &= \mathbb{E}\{\Theta_h(t)\Theta_h(t')\} = \int \int (a_h + b_h t)(a_h + b_h t') N(a_h; 0, \kappa_{h1}) N(b_h; 0, \kappa_{h2}) da_h db_h \\ &= \kappa_{h1} + \kappa_{h2} t t', \end{aligned} \quad (3)$$

potentially with $\kappa_{h1} = \kappa_1$ and $\kappa_{h2} = \kappa_2$ for simplicity. Hence, linear latent trajectories are a special case of the general framework of expressions (1) and (2). In general, one may prefer that the latent trajectories have unknown, smooth, nonlinear shapes, which can be accomplished through more flexible covariance functions. Rasmussen and Williams (2006) provide a detailed overview of different choices of covariance function.

Note that formulation (1)-(2) corresponds to a finite mixture of Gaussian processes. Let $\mathcal{M}_i = h$ for $\eta_i = \Theta_h$ denote that subject i is drawn from mixture component h . To allow the probabilities \mathbf{p} allocated to each of the latent classes to be unknown, we can choose a Dirichlet prior, $\mathbf{p} \sim \text{Diri}(\alpha_1, \dots, \alpha_k)$. In the special case in which Θ_h is drawn from a finite-dimensional Gaussian instead of a Gaussian process, Ishwaran and Zarepour (2002a)

showed that the choice $\alpha_k = \alpha/k$ has appealing theoretical properties, including consistency and a limiting distribution, for $k \rightarrow \infty$, corresponding to a Ferguson Dirichlet process for G .

Theorem 1. Choosing $\kappa_h = \kappa$ and $\mathbf{p} \sim \text{Diri}(\alpha/k, \dots, \alpha/k)$, in the limit as $k \rightarrow \infty$, we have

$$G(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\Theta_h}(\cdot), \quad \Theta_h \sim GP(\mathcal{C}_\kappa)$$

$$p_h = V_h \prod_{l=1}^{h-1} (1 - V_l), \quad V_h \stackrel{iid}{\sim} \text{beta}(1, \alpha).$$

The proof follows from the proof of Theorem 2 in Ishwaran and Zarepour (2002b). The specification shown in Theorem 1 corresponds to a functional Dirichlet process (FDP). Unlike the finite mixture model, this infinite mixture model avoids bounding the number of latent classes. However, the finite mixture model (2) can be considered to approximate the FDP. In addition, note that the finite mixture specification does not require all k classes to be occupied, so that k effectively serves as an upper bound on the number of classes, which is treated as unknown.

3.2 Identifiability Issues

A potential concern in fitting model (1)-(2) is non-identifiability. For sake of discussion, first consider the case in which $k = 1$, so $\eta_i = \Theta_1$ for all i , with Θ_1 a mean function assigned a Gaussian process prior. Then, the curve Θ_1 is a smooth mean function, which is shared by the different subjects, while ϵ_i is a potentially-noisy subject-specific deviation from this curve. We recommend choosing different forms for the covariance functions, \mathcal{C}_ν and \mathcal{C}_κ , to reflect the differences in expected shape of the mean and residual trajectories. For example, in the water quality application, we used the following form for \mathcal{C}_ν :

$$\mathcal{C}_\nu(t, t') = \nu_1 \exp \{ -\nu_2 * (t - t')^2 \} + 0.04\nu_1 1(t = t'), \quad (4)$$

where $\nu = (\nu_1, \nu_2)'$ are unknown parameters. For computational reasons, we follow Neal (1997) in adding a small amount of jitter to the Gaussian covariance function. Without this

term, the condition number of the covariance matrix at the observation times may be large, leading to inaccurate matrix computations.

We place log-normal hyperpriors on ν_1 and ν_2 for added flexibility. Examining the data in Figure 1, we note that the posterior for ν_1 and ν_2 will tend to concentrate on values of ν such that realizations from $GP(\mathcal{C}_\nu)$ have a similar shape to the empirical curves. In particular, these realizations will tend to be bumpy curves centered on a flat line at zero on average. The mean curve Θ_1 then allows systematic deviations from this flat line across the subjects. We anticipate that Θ_1 is much smoother than the ϵ_i 's and choose the following covariance function:

$$\mathcal{C}_\kappa(t, t') = \kappa_1(1 + tt') + \kappa_2 \left| \frac{1}{2}(t - t') \min(t, t')^2 + \frac{1}{3} \min(t, t')^3 \right|, \quad (5)$$

where $\kappa = (\kappa_1, \kappa_2)'$ are unknown parameters. Rasmussen (unpublished observation) showed that this covariance function leads to a posterior mean curve, which is asymptotically equivalent to a natural cubic spline. Realizations from a Gaussian process with this covariance function lead to curve shapes in agreement with our expectation for the latent trajectories in the water quality application. We chose log-normal hyperprior densities for κ_1 and κ_2 to allow the data to inform about the covariance function.

Note that with a single mean trajectory, Θ_1 , we clearly have Bayesian learning in that the prior for Θ_1 and for the covariance parameters, ν and κ , will be updated by the data to produce a posterior distribution that may differ substantially from the prior. However, in examining Figure 1 and additional trajectories from the study, it is clear that a single mean curve is insufficient. Hence, to better fit the data, we can add additional latent classes, each with a different mean trajectory. The main problem arises in determining how many trajectory classes, k , to use. If we include too few, subjects with substantively different trajectories may be grouped together, while including too many will lead subjects with very similar trajectories to be split into subgroups. The FDP effectively solves this problem by

treating k as a random variable.

3.3 Properties of FDP Prior

Assume that $\eta_i \sim G$, for $i = 1, \dots, n$, where $G \sim FDP(\alpha, \mathcal{C}_\kappa)$ denotes that G is assigned an FDP prior centered on the Gaussian process, $GP(\mathcal{C}_\kappa)$, with precision α . This implies that the random variable $\eta_i(t) \sim G(t)$, with $G(t) \sim DP(\alpha G_0(t))$ and $G_0(t)$ obeying a univariate Gaussian law. Hence, for any time t , the unknown latent variable distribution at t is assigned a Dirichlet process prior.

To obtain additional insight into the clustering process, we consider the conditional prior distribution of the latent trajectory for subject i , η_i , given the trajectories for other subjects, $\boldsymbol{\eta}^{(i)} = \{\eta_1, \dots, \eta_{i-1}, \eta_{i+1}, \dots, \eta_n\}$, marginalizing over the FDP prior. This conditional distribution follows the Blackwell and MacQueen (1973) Pólya urn scheme:

$$(\eta_i | \boldsymbol{\eta}^{(i)}, \alpha) = \left(\frac{\alpha}{\alpha + n - 1} \right) GP(\mathcal{C}_\kappa) + \left(\frac{1}{\alpha + n - 1} \right) \sum_{j \neq i} \delta_{\eta_j}. \quad (6)$$

Hence, subject i is either assigned to a new class having a trajectory sampled at random from a Gaussian process, with probability $\alpha/(\alpha + n - 1)$, or is grouped with one of the other subjects chosen at random. Because many of these other subjects are also grouped together, one can re-express (6) as

$$(\eta_i | \boldsymbol{\theta}^{(i)}, \alpha) = \left(\frac{\alpha}{\alpha + n - 1} \right) GP(\mathcal{C}_\kappa) + \left(\frac{1}{\alpha + n - 1} \right) \sum_{h=1}^{k^{(i)}} n_h^{(i)} \delta_{\theta_h^{(i)}}, \quad (7)$$

where $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \dots, \theta_{k^{(i)}}^{(i)})'$ represents the $k^{(i)}$ unique values of $\boldsymbol{\eta}^{(i)}$, $\mathcal{S}_j^{(i)} = h$ denotes that $\eta_j = \theta_h^{(i)}$, $j \neq i$, and $n_h^{(i)} = \sum_{j \neq i} 1(\mathcal{S}_j^{(i)} = h)$ is the number of elements of $\boldsymbol{\eta}^{(i)}$ in cluster h . This expression implies directly that k , the number of unique trajectories represented in $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_n\}$, increases stochastically with α and n .

In fact, the prior for k in terms of α and n can be expressed as described in Antoniak (1974) for the Dirichlet process, as the clustering properties of the FDP are identical with

those of the DP:

$$\Pr(k = h | \alpha, n) = c_n(h)n!\alpha^h \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad \text{for } h = 1, \dots, n, \quad (8)$$

where $c_n(h) = \Pr(k = h | \alpha = 1, n)$ are free from α and can be calculated using recurrence equations for Stirling numbers. Updating this prior with the data, we obtain a posterior distribution for the number of latent class trajectories that is driven by the data but is also critically dependent on α .

The property of increasing k stochastically with n can be viewed as an appealing property, both substantively, as new subjects may have different attributes than subjects currently in the sample, and theoretically, as increasing k slowly with n may be needed for consistency. For work on posterior consistency in Gaussian process models, refer to Ghosal and Roy (2006) and Choi (2005). However, from the perspective of interpretability and computational feasibility, it is necessary to limit the number of clusters. Certainly, results become difficult to interpret as k becomes much greater than 20-30, but the DP-type stick-breaking form will tend to force the addition of new clusters with increasing n , regardless of interpretability.

Recognizing the sensitivity of the prior for k to the choice of α , a common strategy is to choose a hyperprior density for α (Escobar and West, 1995), though this can still result in very many clusters in practice. We instead advocate setting $\alpha = g/\log(n)$ motivated by the results of Theorem 1, which shows that the asymptotic prior distribution of k depends only on the fixed hyperparameter g and is free of n . In particular, for large n , the number of clusters is approximately distributed as $1 + \text{Poisson}(g)$.

Theorem 2. Letting $\alpha = g/\log(n)$, with g a pre-specified constant, we have

$$\lim_{n \rightarrow \infty} \Pr(k = h | \alpha, n) = \pi_h = \frac{g^{h-1} \exp(-g)}{(h-1)!}, \quad h = 1, 2, \dots, \infty,$$

The proof follows trivially from the large sample Poisson approximation of West (1992).

4. Joint Nonparametric Modeling

4.1 Motivation

The FDP prior described in Section 3 is useful for continuous time nonparametric modeling and flexible clustering of longitudinal data. However, our primary motivation, drawn from the water quality application described in Section 2, is to assess the relationship between a longitudinal predictor and a multivariate response. In particular, using a dimensionality reduction device to aid in interpretation, we would like to be able to classify the time-varying trajectories in BDCM levels during pregnancy into a modest number of categories, with the category indicators then predicting (or not) the joint distribution of GAD and BW.

Potentially, one could apply a two-stage approach, first fitting the latent class trajectory model to classify the disinfection by-product curves, and then plugging in the class indicators into a regression model for the outcome variables. However, such an approach may underestimate uncertainty in estimating the classes, does not allow the response values to inform about clustering, and can lead to bias. Instead, we propose a joint nonparametric modeling approach. In Section 4.2, we propose an approach based on assigning a parametric model to the class-specific response distributions. In Section 4.3, we generalize the method in Section 4.3 to treat class-specific response distributions nonparametrically.

4.2 Joint Functional Dirichlet Process

Our goal is to develop a joint nonparametric latent class model for the longitudinal trajectory, y_i , and a multivariate response, $\mathbf{r}_i = (r_{i1}, \dots, r_{is})' \in \mathfrak{R}^s$. We initially consider the model

$$y_i = \eta_i + \epsilon_i, \quad \epsilon_i \sim GP(\mathcal{C}_\nu), \quad \text{and} \quad \mathbf{r}_i = \boldsymbol{\mu}_i + \boldsymbol{\tau}_i, \quad \boldsymbol{\tau}_i \sim N_s(\mathbf{0}, \boldsymbol{\Sigma}), \quad (9)$$

so that the trajectory y_i is characterized as in (1), and we assume that the joint distribution of \mathbf{r}_i is multivariate normal with a subject-specific mean. Then, to group the subjects into

latent classes nonparametrically, we let

$$\phi_i = \{\eta_i, \boldsymbol{\mu}_i\} \sim H = \sum_{h=1}^{\infty} p_h \delta_{\Psi_h}(\cdot), \quad \Psi_h = \{\Theta_h, \boldsymbol{\mu}_h^*\} \sim H_0, \quad (10)$$

where $\mathbf{p} = (p_h, h = 1, \dots, \infty)'$ is an infinite sequence of stick-breaking weights, defined as in Theorem 1, and Ψ_h collects the function Θ_h together with the s -dimensional finite mean $\boldsymbol{\mu}_h^*$. The known probability measure H_0 has support on Φ , the Cartesian product of the function space Ω with \mathfrak{R}^s . For example, a convenient choice of H_0 corresponds to the product measure of $GP(\mathcal{C}_\kappa)$ with $N_s(\mathbf{0}, \mathbf{D})$. In this case, Ψ_h is sampled by drawing a function $\Theta_h \sim GP(\mathcal{C}_\kappa)$ and then independently drawing $\boldsymbol{\mu}_h^* \sim N_s(\mathbf{0}, \mathbf{D})$.

Note that H is a random probability measure on (Φ, \mathcal{A}) , with \mathcal{A} the Borel σ -algebra of subsets of Φ . We refer to prior (10) as a joint functional DP (JFDP) to denote that the functional distribution is modeled jointly with the response distribution. The JFDP shares many properties with the DP, including almost sure discreteness and identical clustering properties to those discussed in Section 3.3, as the stick-breaking process for the weights remains unchanged. Let $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_k\}$ denote the values of $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_h, h = 1, \dots, \infty\}$ represented in the sample of n subjects, with $\boldsymbol{\psi}_h = \{\theta_h, \tilde{\boldsymbol{\mu}}_h\}$, for $h = 1, \dots, k$. In addition, let $\mathcal{S}_i = h$ denote that individual i is in latent class h , implying $\eta_i = \theta_h$ and $\boldsymbol{\mu}_i = \tilde{\boldsymbol{\mu}}_h$. Note that expression (9) can be rewritten as:

$$(y_i | \mathcal{S}_i = h) \sim GP(\theta_h, \mathcal{C}_\nu) \quad \text{and} \quad (\mathbf{r}_i | \mathcal{S}_i = h) \sim N_s(\tilde{\boldsymbol{\mu}}_h, \boldsymbol{\Sigma}). \quad (11)$$

Hence, the longitudinal trajectories for subjects in latent class h are drawn from a Gaussian process centered on the class-specific trajectory mean, θ_h , while the responses are drawn from a multivariate normal centered on the class-specific response mean, $\tilde{\boldsymbol{\mu}}_h$.

In this manner, we allow a systematic difference in the mean of the response distribution according to the latent trajectory class. To clarify the structure, consider the water quality application and the data in Figure 1. The longitudinal trajectories are assumed to arise

from one of k classes, with the classes varying in the smooth mean trajectory. For example, it seems likely that a class will be introduced corresponding to a flat trajectory near zero, with three of the women represented in Figure 1 assigned to this class. This flat trajectory class will also have a class-specific mean representing the average gestational age at delivery and offspring birth weight for women in that class. If BDCM has an adverse effect, then one might expect the flat trajectory class to have a higher mean (longer gestation, heavier babies) than other classes, which correspond to higher exposure levels. By updating the JFDP prior with the data, one allows uncertainty in the number of classes, the class assignment, the shapes of the trajectories in each class, and the corresponding means.

4.3 *Nonparametric Class-Specific Response Densities*

Although the approach described in Section 4.2 is very flexible and should be useful in many applications, one critical assumption that is common to many model-based clustering methods is violated for the water quality application. This assumption is that the class-specific response distribution is multivariate normal. As is illustrated in Figure 2, the joint distribution of gestational age at delivery and birth weight is not well approximated by a bivariate normal. Hence, in applying the approach of Section 4.2, many classes are introduced to accommodate lack of fit of the bivariate normal model, instead of variability in the longitudinal trajectories. For example, in fitting the approach to data from women having very similar flat BDCM trajectories, one introduces multiple classes, because a mixture of normals fits the response data better than a single normal. This greatly inflates the number of classes and makes interpretation difficult.

To solve this problem, we model the distribution of the residual, $\boldsymbol{\tau}_i$, in expression (9) as a Dirichlet process location mixture of normals instead of as normal:

$$\boldsymbol{\tau}_i \sim N_s(\boldsymbol{\xi}_i, \boldsymbol{\Sigma}), \quad \boldsymbol{\xi}_i \sim F, \quad F \sim DP(\alpha_F F_0), \quad (12)$$

where α_F is the DP precision and F_0 is the base measure. Under this structure, we obtain

the following expression for the response density in the h th latent trajectory class:

$$f_h(\mathbf{r}) = \int N_s(\mathbf{r}; \tilde{\boldsymbol{\mu}}_h + \boldsymbol{\xi}, \boldsymbol{\Sigma}) F(d\boldsymbol{\xi}) = \sum_{l=1}^{\infty} \pi_l N_s(\mathbf{r}; \tilde{\boldsymbol{\mu}}_h + \boldsymbol{\xi}_l, \boldsymbol{\Sigma}), \quad \boldsymbol{\xi}_l \sim N_s(0, \Gamma), \quad (13)$$

with $\boldsymbol{\pi} = (\pi_l, l = 1, \dots, \infty)'$ DP stick-breaking weights, and $N_s(0, \Gamma)$ chosen for F_0 . Hence, the joint distribution of the response is modeled as an infinite mixture of normals, with $\tilde{\boldsymbol{\mu}}_h$ providing a class-specific deviation. For example, latent trajectory classes having a greater proportion of preterm births and lighter babies would have a lower value of $\tilde{\boldsymbol{\mu}}_h$ relative to other classes. In this manner, we avoid the problem of splitting trajectory classes to better fit the response distribution, while not sacrificing interpretability, as we have a single parameter ranking the trajectory classes relative to each other for each outcome type.

5. Posterior Computation

5.1 MCMC algorithm

We propose a Gibbs sampling algorithm for posterior computation. We follow a common strategy in computation for DP mixture models (MacEachern, 1998; MacEachern and Müller, 1998) in alternating between updating the cluster allocation indicators and number of clusters separately from updating the cluster-specific parameters. We focus initially on the case described in section 4.2, and then describe extensions to accommodate fixed predictors and allow a nonparametric residual distribution in the response model.

Note that the data for subject i consist of $\mathbf{z}_i = [y_i(\mathbf{t}_i)', \mathbf{r}_i']'$, which is the $n_i + s$ vector of longitudinal observations along with the response. Let $\phi_i = (\eta_i(\mathbf{t}_i)', \boldsymbol{\mu}_i)'$, $\epsilon_i^* = (\epsilon_i(\mathbf{t}_i)', \boldsymbol{\tau}_i)'$, and $\boldsymbol{\Sigma}_i = \text{block-diag}(\mathcal{C}_\nu(\mathbf{t}_i), \boldsymbol{\Sigma})$, with $\mathcal{C}_\nu(\mathbf{t}_i)$ denoting the covariance matrix of the random variable $\epsilon_i(\mathbf{t}_i)$. Then, the likelihood for the data from subject i conditional on $\phi_i, \nu, \boldsymbol{\Sigma}$ is

$$L(\mathbf{z}_i; \phi_i, \nu, \boldsymbol{\Sigma}) = N_{n_i+s}(\mathbf{z}_i; \phi_i, \boldsymbol{\Sigma}_i). \quad (14)$$

Let $\phi_i(\mathbf{t}) = (\eta_i(\mathbf{t})', \boldsymbol{\mu}_i)'$, for any finite dimensional vector \mathbf{t} having elements in \mathfrak{R}^+ , and use the notation $\phi_i(\cdot)$ to denote $\{\eta_i(\cdot), \boldsymbol{\mu}_i\}$, with $\eta_i(\cdot)$ a stochastic process in time. Then,

$\boldsymbol{\psi}(\cdot) = \{\psi_1(\cdot), \dots, \psi_k(\cdot)\}$ denotes the $k \leq n$ unique values of $\{\phi_i(\cdot), i = 1, \dots, n\}$. We let $\mathcal{S}_i = h$ if $\phi_i(\cdot) = \psi_h(\cdot)$ index the allocation of subject i to trajectory class h , for $h = 1, \dots, k$, with $\mathbf{S} = (\mathcal{S}_1, \dots, \mathcal{S}_n)'$. Furthermore, excluding subject i , let $\boldsymbol{\psi}^{(i)}(\cdot)$ denote the $k^{(i)}$ unique values of $\{\phi_j(\cdot), j \neq i\}$, $\mathcal{S}_j^{(i)} = h$ if $\phi_j(\cdot) = \psi_h(\cdot)$, for $h = 1, \dots, k^{(i)}$, and $\mathbf{S}^{(i)} = (\mathcal{S}_j, j \neq i)'$.

Because it is impossible to do computation for an infinite-dimensional parameter, we focus on the finite realization at times $\mathbf{t} = (t_1, \dots, t_J)'$, with $\mathbf{t}_i \subset \mathbf{t}$, for $i = 1, \dots, n$. Then, the conditional prior distribution of $\phi_i(\mathbf{t})$ given $\mathbf{S}^{(i)}$, $k^{(i)}$ and $\boldsymbol{\psi}^{(i)}(\mathbf{t})$ is

$$(\phi_i(\mathbf{t}) | \mathbf{S}^{(i)}, k^{(i)}, \boldsymbol{\psi}^{(i)}(\mathbf{t}), \alpha) \sim a_{i0} H_{0(\mathbf{t})} + \sum_{h=1}^{k^{(i)}} a_{ih} \delta_{\psi_h^{(i)}(\mathbf{t})}, \quad (15)$$

where $a_{i0} = c_0 \alpha$, $a_{ih} = c_0 n_h^{(i)}$, $n_h^{(i)} = \sum_{j \neq i} 1(\mathcal{S}_j = h)$, $c_0 = 1/(\alpha + n - 1)$, and $H_{0(\mathbf{t})}$ is the probability measure corresponding to $N_{J+s}(\mathbf{0}, \mathbf{C}_\kappa)$, with $\mathbf{C}_\kappa = \text{block-diag}(\mathbf{C}_\kappa(\mathbf{t}), \mathbf{D})$. Updating prior (15) with data \mathbf{z}_i , we obtain the conditional posterior

$$(\phi_i(\mathbf{t}) | \mathbf{z}_i, \mathbf{S}^{(i)}, k^{(i)}, \boldsymbol{\psi}^{(i)}(\mathbf{t}), \alpha) \sim q_{i0} H_{i,0(\mathbf{t})} + \sum_{h=1}^{k^{(i)}} q_{ih} \delta_{\psi_h^{(i)}(\mathbf{t})}, \quad (16)$$

where the terms are derived below. Letting $\phi^i(\mathbf{t}) = T_i\{\phi(\mathbf{t})\} = [\eta(\mathbf{t}_i)', \boldsymbol{\mu}'_i, \eta(\bar{\mathbf{t}}_i)']$, with $\bar{\mathbf{t}}_i = \mathbf{t}/\mathbf{t}_i$, $\phi(\mathbf{t}) \sim N_{J+s}(\mathbf{0}, \mathbf{C}_\kappa)$ implies $\phi^i(\mathbf{t}) \sim N_{J+s}(\mathbf{0}, \mathbf{C}_\kappa^i)$, with $T_i\{\cdot\}$ a reordering operator and \mathbf{C}_κ^i obtained from \mathbf{C}_κ by applying $T_i\{\cdot\}$ to the rows and columns. Then

$$\begin{aligned} H_{i,0(\mathbf{t})}(\phi_i(\mathbf{t})) &= \frac{1}{h_i(\mathbf{z}_i)} N_{n_i+s}(\phi_i; \mathbf{0}, \mathbf{C}_{11}) N_{n_i+s}(\mathbf{z}_i; \phi_i, \boldsymbol{\Sigma}_i) \\ &\quad \times N_{J-n_i}(\eta_i(\bar{\mathbf{t}}_i); \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \phi_i, \mathbf{C}_{11} - \mathbf{C}_{12} \mathbf{C}_{22}^{-1} \mathbf{C}_{21}), \end{aligned} \quad (17)$$

where $\phi_i = [\eta_i(\mathbf{t}_i)', \boldsymbol{\mu}'_i]'$, $h_i(\mathbf{z}_i)$ is the normalizing constant, and

$$\mathbf{C}_\kappa^i = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}, \quad \text{with } \mathbf{C}_{11} = \text{first } n_i + s \text{ rows \& columns of } \mathbf{C}_\kappa^i.$$

The mixture weights in (16) can then be shown to be $q_{i0} = c h_i(\mathbf{z}_i)$ and $q_{ih} = c L(\mathbf{z}_i; \psi_h^{(i)}(\mathbf{t}_i), \boldsymbol{\Sigma}_i)$, for $h = 1, \dots, k^{(i)}$, with c a normalizing constant.

Then, to implement the MCMC sampling algorithm:

1. Sample \mathcal{S}_i from the multinomial conditional posterior distribution with

$$\Pr(\mathcal{S}_i = h \mid \mathbf{z}_i, \mathbf{S}^{(i)}, k^{(i)}, \boldsymbol{\psi}^{(i)}(\mathbf{t})) = q_{ih}, \quad \text{for } h = 0, 1, \dots, k^{(i)}.$$

If $\mathcal{S}_i = 0$, let $k = k^{(i)} + 1$ and assign subject i to cluster k , with $\phi_i(\mathbf{t}) = \psi_k(\mathbf{t}) \sim H_{i,0}(\mathbf{t})$.

2. Update $\psi_h(\mathbf{t})$, for $h = 1, \dots, k$, by sampling from the Gaussian full conditional

$$(\psi_h(\mathbf{t}) \mid \mathbf{z}, \mathbf{S}, k,) \propto N_{J+s}(\psi_h(\mathbf{t}); \mathbf{0}, \mathbf{C}_\kappa) \prod_{i:\mathcal{S}_i=h} N_{n_i+s}(\mathbf{z}_i; \psi_h(\mathbf{t}_i), \boldsymbol{\Sigma}_i).$$

We also include Metropolis-Hastings steps for updating the covariance parameters, ν and κ , and a Gibbs step for updating $\boldsymbol{\Sigma}$ from its inverse-Wishart full conditional (assuming an inverse-Wishart prior).

5.2 Generalization 1: Fixed Predictors

It is straightforward to modify the model and computation algorithm to accommodate a $w \times 1$ vector of fixed predictors, $\mathbf{x}_i = (x_{i1}, \dots, x_{iw})'$, of the response, \mathbf{r}_i , by replacing $\mathbf{r}_i = \boldsymbol{\mu}_i + \boldsymbol{\tau}_i$ in expression (9) with $\mathbf{r}_i = \boldsymbol{\mu}_i + \boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\tau}_i$, where $\boldsymbol{\beta}$ is an $s \times w$ matrix of coefficients. We assume a matrix-normal prior for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}, \mathbf{K}) = \frac{|\mathbf{K}|^{s/2}}{2\pi^{s|w|}} \exp \left[-\frac{1}{2} \text{tr} \left\{ (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \mathbf{K} \right\} \right],$$

with $\boldsymbol{\beta}_0$ an $s \times w$ mean matrix, $\boldsymbol{\Sigma}$ is an $s \times s$ row covariance matrix, and \mathbf{K} is a $w \times w$ column covariance matrix.

An appealing default choice corresponds to $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{K} = \mathbf{X}\mathbf{X}'/n$, where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$. Letting $\tilde{\mathbf{r}}_i = \mathbf{r}_i - \boldsymbol{\mu}_i$, for $i = 1, \dots, n$, with $\tilde{\mathbf{R}} = [\tilde{\mathbf{r}}_1, \dots, \tilde{\mathbf{r}}_n]$, we obtain the following posterior for $\boldsymbol{\beta}$ under the default prior:

$$(\boldsymbol{\beta} \mid \tilde{\mathbf{R}}, \mathbf{X}, \boldsymbol{\Sigma}) \sim \mathcal{N} \left(\tilde{\mathbf{R}}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} \left(\frac{n}{n+1} \right), \boldsymbol{\Sigma}, \mathbf{X}\mathbf{X}' \left(\frac{n+1}{n} \right) \right). \quad (18)$$

Note that one can sample from this distribution using the fact that $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}, \mathbf{K})$ implies $\tilde{\boldsymbol{\beta}} \sim N_{s \times w}(\tilde{\boldsymbol{\beta}}_0, \boldsymbol{\Sigma} \otimes \mathbf{K}^{-1})$, with $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}_0$ denoting the stacked columns of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$,

respectively. The algorithm of Section 5.2 is then generalized by simply including a step to sample (18), and then replacing \mathbf{r}_i by $\mathbf{r}_i^* = \mathbf{r}_i - \boldsymbol{\beta}\mathbf{x}_i$ in the other steps.

5.3 Generalization 2: Nonparametric Residual

Finally, the algorithm can be generalized to accommodate the extension of Section 4.3 in which the residual distribution in the response model is characterized as a DP mixture of normals. We start by letting $\mathbf{r}_i = \boldsymbol{\mu}_i + \boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\xi}_i + \boldsymbol{\tau}_i$, where $\boldsymbol{\xi}_i \sim F$ and $F \sim DP(\alpha_F F_0)$. Replacing \mathbf{r}_i with $\mathbf{r}_i - \boldsymbol{\xi}_i$, the steps in Sections 4.1 and 4.2 can proceed without modification. Then, holding $\boldsymbol{\mu}_i, \boldsymbol{\beta}, \mathbf{x}_i$ as fixed, we have

$$\tilde{\mathbf{r}}_i = \mathbf{r}_i - \boldsymbol{\mu}_i - \boldsymbol{\beta}\mathbf{x}_i = \boldsymbol{\xi}_i + \boldsymbol{\tau}_i, \quad \boldsymbol{\tau}_i \sim N_s(\mathbf{0}, \boldsymbol{\Sigma}),$$

which is a typical DP location mixture of normals. Hence, updating of $\boldsymbol{\xi}_i$ can proceed as for a typical DP mixture model. One possibility is to use a simplification of the algorithm of Section 5.1 in which we alternate between (i) updating indicators $\mathbf{T} = (\mathcal{T}_1, \dots, \mathcal{T}_n)$ of the configuration of subjects to the k_ξ unique values of $\boldsymbol{\xi} = \{\boldsymbol{\xi}_i, i = 1, \dots, n\}$; and (ii) updating the cluster-specific means by sampling from the posterior conditional on \mathbf{T} , which has a simple multivariate normal form.

5.4 Post-Processing MCMC Output

Although the JFDP is extremely flexible, there can be some difficulties in interpretation arising due to the fact that the number of trajectory classes, k , will vary across the samples from the posterior distribution. Such variability is not a problem if one wants to perform inferences on the number of classes, or predict data for future subjects. However, our goal in the water quality study is to examine the identified class-specific trajectories and to assess how pregnancy outcomes vary across these classes.

One possibility is to apply a simulated annealing algorithm to estimate a posterior modal configuration of subjects to a single set of trajectory classes. We find that it is difficult to

obtain reliable results for such an approach due to the multimodal nature of the posterior, and a tendency to converge to a local mode. In addition, even if it could be obtained consistently, the modal configuration may be of limited interest in that there is likely a large number of alternative configurations having similar posterior probabilities, making inferences on any one configuration unreliable.

Instead, we propose an approach to estimate a single set of clusters based on post-processing of the MCMC output. In particular:

1. At iteration t after burn-in, initialize the clusters by letting $k_0 = k$ and $\boldsymbol{\psi}_0 = \{\psi_h(\mathbf{t}), l = 1, \dots, k_0\}$.
2. At iteration $t + u * i$, $i = 1, 2, \dots$, ($u =$ thinning integer), update the clusters by assigning $\psi_h(\mathbf{t})$ to the *closest* element of $\boldsymbol{\psi}_0$, for $h = 1, \dots, k$. *Closeness* is based on L_2 distance from the *cluster seed*, which is equal to the weighted average of all the values allocated to the cluster so far, with weights proportional to the number of subjects allocated.

This algorithm is then applied for a large number of iterations, obtaining a posterior distribution for the number of subjects allocated to each class and for the class-specific trajectories. The final cluster seed provides an estimate of the predictive mean of ϕ_{n+1} for a new subject $i = n + 1$ in class l , for $l = 1, \dots, k_0$, while realizations of $\psi_h(\mathbf{t})$ within each class provide a measure of within-class variability in the trajectories.

A potential concern is sensitivity to the value of k_0 sampled at iteration t . However, we find that all the common trajectory shapes tend to be represented at each iteration after convergence. The main changes that occur across iterations tend to be merging of similar classes, and splitting classes into sub-classes that are quite similar in shape, with an occasional substantively different outlying cluster introduced. Our post-processing approach tends to do a good job at identifying clusters having more than a few members based on our

simulation results and careful consideration of real data examples.

6. Simulation Examples

Using a small simulation study, we evaluated the MCMC algorithm and gauged performance of the approach. We let $J = 20$ (observation times) and $s = 2$ (number of outcomes) and chose priors motivated by the water quality and pregnancy outcomes study. In addition, we focused on $n = 500$ subjects and $k = 4$ classes, with the following true mean functions:

$$\theta_1(t) = 0, \quad \theta_2(t) = \left(\frac{t}{J+1}\right)^2, \quad \theta_3(t) = 1 - \cos\left(\frac{5 * t}{J+1}\right), \quad \theta_4(t) = 0.5 * \log\left(\frac{t}{J+1-t}\right).$$

Individuals were drawn from classes with equal probability. In each case, trajectories were drawn from model (1), with $\eta_i = \theta_h$ for individuals in class h and with residual GP covariance function (4) having $\nu_1 = 0.005$ and $\nu_2 = 1$. This choice corresponds to a moderate amount of variability among individuals within a class.

The response is bivariate, with a single fixed predictor $x_i \sim \text{Bernoulli}(0.5)$ having $\beta = 0$. Four different cases were considered for the response densities: (i) $N_2(\mathbf{0}, \mathbf{I}_2)$ for each class; (ii) same as case (i) but with class 2 having mean $(-1, -2)$; (iii) response distribution

$$0.3 N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) + 0.5 N_2\left(\begin{pmatrix} -1 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right) + 0.2 N_2\left(\begin{pmatrix} 1.5 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}\right)$$

in each class, and (iv) same as (iii) but with class 2 having shift $(-1, -2)'$ from common mean $(-0.2, -0.8)'$.

In each case, we simulated one data set and implemented the MCMC algorithm of Section 5.1-5.3 for 30,000 iterations, with the first 5,000 discarded as a burn-in. Priors for the covariance parameters were chosen as:

$$\log \nu_1 \sim N(-5.3, 25), \quad \log \nu_2 \sim N(0, 25), \quad \log \kappa_1 \sim N(-6.9, 1), \quad \text{and} \quad \log \kappa_2 \sim N(-6.9, 1),$$

with the hyperparameters chosen to generate curves in agreement with our prior expectation for the range of trajectories and residual behavior in the water quality data. In addition, we

chose the default priors for β , α and α_F as recommended in Sections 3.3 and 5.2 with $g = 1$, and we let $\mathbf{D} = \Gamma = \mathbf{I}_2$, $\Sigma^{-1} \sim \mathcal{W}(\nu_0, \Sigma_0)$ with $\nu_0 = 1,000$ and $\Sigma_0 = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{pmatrix}$, with \mathcal{W} denoting the Wishart distribution.

Figure 3 shows the results in case (i). Because the cluster index is arbitrary, we plot the clusters in order of allocation probability, matched to the closest true cluster. The trajectories were all estimated accurately, with the 5,000 draws from the posterior within each class all very close to the true trajectory. In addition, 99% credible intervals for the class-specific response means included the true values of zero in each case. The correct number of trajectory classes was identified with high posterior probability, $\Pr(k = 4 | \text{data}) > 0.99$, and the estimate of β was close to the true value, $\hat{\beta} = -0.028$ with 95% interval $[-0.20, 0.14]$. Performance was similar in cases (ii)-(iv), with the results in case (iv) shown in Figure 4.

7. Water Quality and Pregnancy Outcomes

We applied the same approach used in the simulation examples of Section 6 to the water quality and pregnancy outcomes data described in Section 2, with the exception that we included $w = 7$ predictors in the \mathbf{x}_i vector, including indicators of male offspring, current smoking, recently quit smoking, distant quit smoking, African American ethnicity, Latino ethnicity and previous pregnancy (multiparous). For sake of comparison, we first applied a simpler analysis that assumed $\epsilon_i(t) \sim N(0, \sigma^2)$ instead of $\epsilon_i \sim GP(\mathcal{C}_\nu)$. As expected, we obtained a large number of latent trajectory classes, because the residual dependency structure was not rich enough to characterize dependent deviations from the trajectory means.

Data were standardized prior to analysis separately for BDCM, GAD and BW. Posterior summaries of the regression coefficients, β , are included in Table 1. As expected, male babies tended to be born earlier but were heavier, while current smokers and African Americans had more early deliveries and lighter babies. Women with a past pregnancy had

significantly shorter gestational lengths. MLEs from a normal linear regression model fitted to the separate outcomes are shown for comparison.

Figure 5 shows posterior realizations from each of the $k = 19$ estimated trajectories, the posterior mean number of subjects assigned to each trajectory class, and 95% credible intervals for the (standardized) class-specific mean response. Approximately one third of the women were assigned to a flat trajectory, with mean close to zero, which is as expected based on examination of the empirical curves. All of the $n = 656$ assigned to the class have water drawn from an underground aquifer, which has little contamination with organic matter and hence low disinfection by-product levels. These women have close to average pregnancy outcomes. Other trajectories occurred with much lower frequency, representing a wide variety of smooth hill-type shapes.

Posterior summaries of gestational length and birth weight for the women in each trajectory class are shown in Table 2. Most of these classes had quite similar pregnancy outcomes, in agreement with the null results produced by simple frequentist analyses of these data, which relied on collapsing the exposure data into summary statistics. However, a notable exception was class 19, which had a spike in BDCM at week 20 of gestation and significant shorter gestation and lighter babies. Although less than 1% of the women in the sample are assigned to this class, these results suggest that levels of BDCM of $38 \mu/L$ (corresponding to 3 in standardized units) at week 20 or later of gestation may have clinically important adverse effects on pregnancy outcomes. Hence, it would be interesting to run a study on a population with more women exposed in this range.

8. Discussion

This article has proposed a semiparametric Bayesian approach to latent trajectory modeling motivated by an epidemiologic study of water quality and pregnancy outcomes. The method should be useful in many other contexts for studying the relationship between the

trajectory in a time-varying predictor and the joint distribution of a multivariate outcome. The proposed approach provides a unified nonparametric characterization of the distribution of the latent trajectories and the class-specific response distribution, while maintaining interpretability.

Clustering is inherently a subjective task in that goodness-of-fit can be improved by adding clusters for each subject. Hence, we view clustering of trajectories as an aid in interpreting complex multivariate relationships, while reducing dimensionality. In fact, in most applications including the water quality example, the truth is likely that subjects fall along a continuum. The choice of the number of clusters is necessarily dependent upon the prior choice. In particular, the choice of covariance function in the Gaussian process prior for the class-specific trajectories plays an important role. If this covariance function is not chosen to generate smooth trajectory curves, one may obtain a large number of clusters that provide a good fit to the data, while leading to problems in interpretation. In future work, we will consider alternative covariance functions and theoretical properties, such as posterior consistency.

References

- Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes with Application to Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.
- Blackwell, D. and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353-355.
- Bigelow, J. and Dunson, D.B. (2005), "Semiparametric Classification in Hierarchical Functional Data Analysis," *ISDS Discussion Paper*, 2005-18, Duke University.
- Burr, D. and Doss, H. (2005), "A Bayesian Semiparametric Model for Random-Effects Meta Analysis," *Journal of the American Statistical Association*, 100, 242-251.

- Bush, C.A. and MacEachern, S.N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275-285.
- Choi, T. (2005), "Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors," Dissertation, Department of Statistics, Carnegie Mellon University.
- Davidian, M. and Gallant, A.R. (1993), "The Nonlinear Mixed Effects Model with a Smooth Random Effects Density," *Biometrika*, 80, 475-488.
- 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.
- Elliott, M.R., Gallo, J.J., Ten Have, R.R., Bogner, H.R. and Katz, I.R. (2005), "Using a Bayesian Latent Growth Curve Model to Identify Trajectories of Positive Affect and Negative Events Following Myocardial Infarction," *Biostatistics*, 6, 119-143.
- 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), "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.
- Formann, A.K. (2003), "Latent Class Model Diagnosis from a Frequentist Point of View," *Biometrics*, 59, 189-196.
- Garrett, E.S. and Zeger, S.L. (2000), "Latent Class Model Diagnosis," *Biometrics*, 56, 1055-1067.

- Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2005), “Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing,” *Journal of the American Statistical Association*, 100, 1021-1035.
- Ghosal, S. and Roy, A. (2006), “Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression,” , *Annals of Statistics*, to appear.
- Green, P.J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711-732.
- Ishwaran, H. and Takahara, G. (2002), “Independent and Identically Distributed Monte Carlo Algorithms for Semiparametric Linear Mixed Models,” *Journal of the American Statistical Association*, 97, 1154-1166.
- Ishwaran, H. and Zarepour, M. (2002a), “Dirichlet Prior Sieves in Finite Normal Mixture Models,” *Statistica Sinica*, 12, 941-963.
- Ishwaran, H. and Zarepour, M. (2002b), “Exact and Approximate Sum-Representations for the Dirichlet Process,” *Canadian Journal of Statistics*, 30, 1-15.
- Kleinman, K.P. and Ibrahim, J.G. (1998), “A Semiparametric Bayesian Approach to the Random Effects Model,” *Biometrics*, 54, 921-938.
- Laird, N.M. and Ware, J.H. (1982), “Random-Effects Models for Longitudinal Data,” *Biometrics*, 38, 963-974.
- Lin, H.Q., McCulloch, C.E., Turnbull, B.W., Slate, E.H. and Clark, L.C. (2000), “A Latent Class Mixed Model for Analysing Biomarker Trajectories with Irregularly Scheduled Observations,” *Statistics in Medicine*, 19, 1303-1318.
- Lin, H.Q., Turnbull, B.W., McCulloch, C.E. and Slate, E.H. (2002), “Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to

- Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer. *Journal of the American Statistical Association*, 97, 53-65.
- MacEachern, S.N. (1998), "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, New York: Springer-Verlag, pp. 23-44.
- MacEachern, S.N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.
- MacEachern, S.N. (2001), "Decision Theoretic Aspects of Dependent Nonparametric Processes," in *Bayesian Methods with Applications to Science, Policy and Official Statistics*, ed. E. George. Creta: ISBA, pp. 551-560.
- MacEachern, S.N. and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223-239.
- Miglioretti, D.L. (2003), "Latent Transition Regression for Mixed Outcomes," *Biometrics*, 59, 710-720.
- Muthén, B., and Shedden, K. (1999), "Finite Mixture Modeling with Mixture Outcomes using the EM Algorithm. *Biometrics*, 55, 463-469.
- Muthén, B., Brown, C.H., Masyn, K., Jo, B., Khoo, S.T., Yang, C.C., Wang, C.P., Kellam, S.G., Carlin, J.B., Liao, J. (2002), "General Growth Mixture Modeling for Randomized Preventive Interventions. *Biostatistics*, 3, 459-475.
- Neal, R.M. (1997), "Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification," Technical Report No. 9702, Department of Statistics, University of Toronto.

- Promislow, J.H.E., Makarushka, C.M., Gorman, J.R., Howards, P.P., Savitz, D.A., and Hartmann, K.E. (2004), "Recruitment for a Community-Based Study of Early Pregnancy: the Right From The Start Study," *Paediatric and Perinatal Epidemiology*, 18, 143-152.
- Rasmussen, C.E. (1996), "Evaluation of Gaussian Processes and other Methods for Non-linear Regression," PhD Thesis, Department of Computer Science, University of Toronto.
- Rasmussen, C.E. and Ghahramani, Z. (2002), "Infinite Mixtures of Gaussian Process Experts," In T.G. Diettrich, S. Becker and Z. Ghahramani, eds, *Advances in Neural Information Processing Systems 14*, The MIT Press.
- Rasmussen, C.E. and Williams, C.K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Rice, J.A. and Wu, C.O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253-259.
- Sethuraman, J. (1994), "A Constructive Definition of the Dirichlet Process Prior," *Statistica Sinica*, 2, 639-650.
- West, M. (1992), "Hyperparameter Estimation in Dirichlet Process Mixture Models," *ISDS Discussion Paper 92-03*, Duke University.
- Wu, H.L. and Zhang, J.T. (2002), "Local Polynomial Mixed-Effects Models for Longitudinal Data," *Journal of the American Statistical Association*, 97, 883-897.
- Zhang, H.P. (2004), "Mixed Effects Multivariate Adaptive Spline Model for the Analysis of Longitudinal and Growth Curve Data," *Statistical Methods in Medical Research*, 13, 63-82.

Table 1*Posterior summaries of covariate effects on gestational age at delivery and birth weight.**MLEs under linear regression models are shown for comparison.*

Predictor	Gestational age at delivery				Birth weight			
	MLE [†]	Mean	SD	95% CI	MLE [†]	Mean	SD	95% CI
male	-0.131	-0.126	0.085	-0.292, 0.042	63.2	69.6	53.2	-35.9, 172.0
smoke1	-0.599	-0.327	0.131	-0.586, -0.070	-343.1	-262.1	49.4	-359.8, -166.5
smoke2	0.013	-0.076	0.095	-0.260, 0.111	-29.2	-42.9	42.3	-126.8, 40.5
smoke3	0.200	0.121	0.091	-0.056, 0.300	15.6	31.9	33.3	-34.3, 95.9
black	-0.466	-0.100	0.156	-0.405, 0.208	-337.6	-260.7	37.7	-335.3, -187.2
latina	-0.066	0.111	0.149	-0.177, 0.409	-59.0	-4.40	38.5	-80.7, 70.6
multiparous	-0.265	-0.324	0.130	-0.577, -0.066	98.8	91.5	31.2	30.0, 153.5

†=mle for normal linear model, smoke1=current smoker, smoke2=recent quit,

smoke3=distant quit.

Table 2*Posterior summaries of the mean gestational age at delivery (weeks) and birth weight (gms)**within each of the identified latent trajectory classes.*

Class	Frequency	Gestational age at delivery		Birth weight (gm)	
		Mean	95% CI	Mean	95% CI
1	656 (37.7%)	39.1	38.9, 39.4	3358	3319, 3480
2	133.7 (7.7%)	39.1	38.7, 39.4	3266	3276, 3483
3	127.3 (7.3%)	39.4	39.1, 39.8	3448	3384, 3577
4	110.9 (6.4%)	39.6	39.3, 39.9	3389	3441, 3623
5	96.7 (5.6%)	39.2	38.9, 39.6	3458	3322, 3536
6	92.9 (5.3%)	39.8	39.3, 40.2	3544	3457, 3694
7	89.6 (5.1%)	39.5	39.1, 39.9	3476	3390, 3626
8	80.9 (4.6%)	39.3	38.9, 39.7	3336	3330, 3562
9	58.9 (3.4%)	39.8	39.3, 40.2	3535	3447, 3709
10	53.6 (3.1%)	39.6	39.2, 40.1	3375	3422, 3666
11	49.0 (2.8%)	39.9	39.4, 40.3	3465	3476, 3738
12	44.2 (2.5%)	39.0	38.4, 39.7	3306	3185, 3571
13	43.5 (2.5%)	39.5	39.0, 40.0	3452	3370, 3641
14	22.8 (1.3%)	39.9	39.3, 40.5	3554	3451, 3785
15	21.9 (1.3%)	38.7	37.8, 39.5	3214	3000, 3509
16	19.7 (1.1%)	39.1	38.4, 39.8	3299	3193, 3597
17	17.8 (1.0%)	39.0	38.3, 39.7	3166	3148, 3546
18	12.2 (0.7%)	39.2	38.4, 40.0	3330	3185, 3636
19	10.3 (0.6%)	37.9	37.1, 38.8	2993	2806, 3308

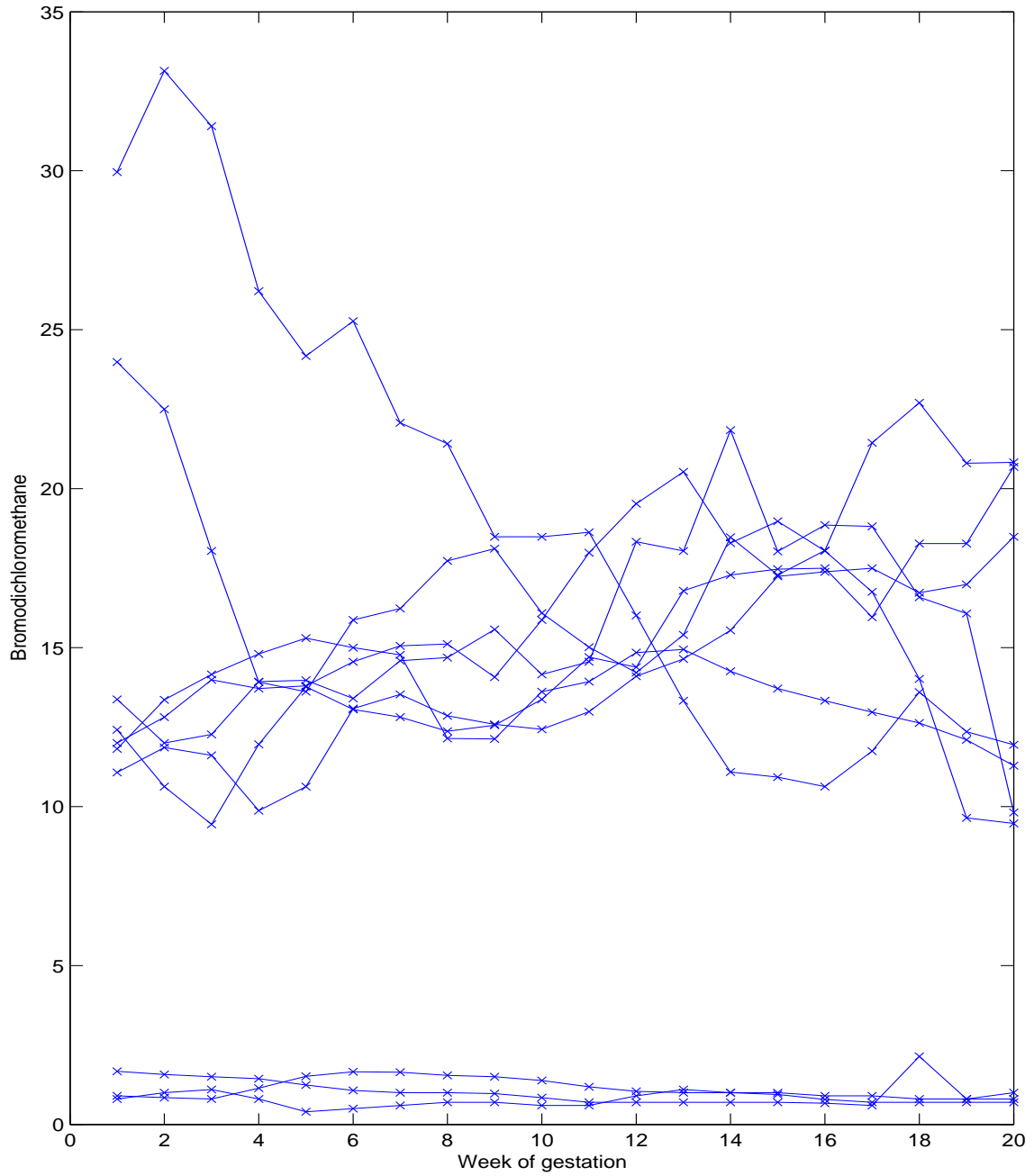


Figure 1: Measured bromodichloromethane values ($\mu\text{g/L}$) in drinking water during the first 20 weeks of pregnancy for 10 randomly-selected women from among the 1742 women in the RFTS study.

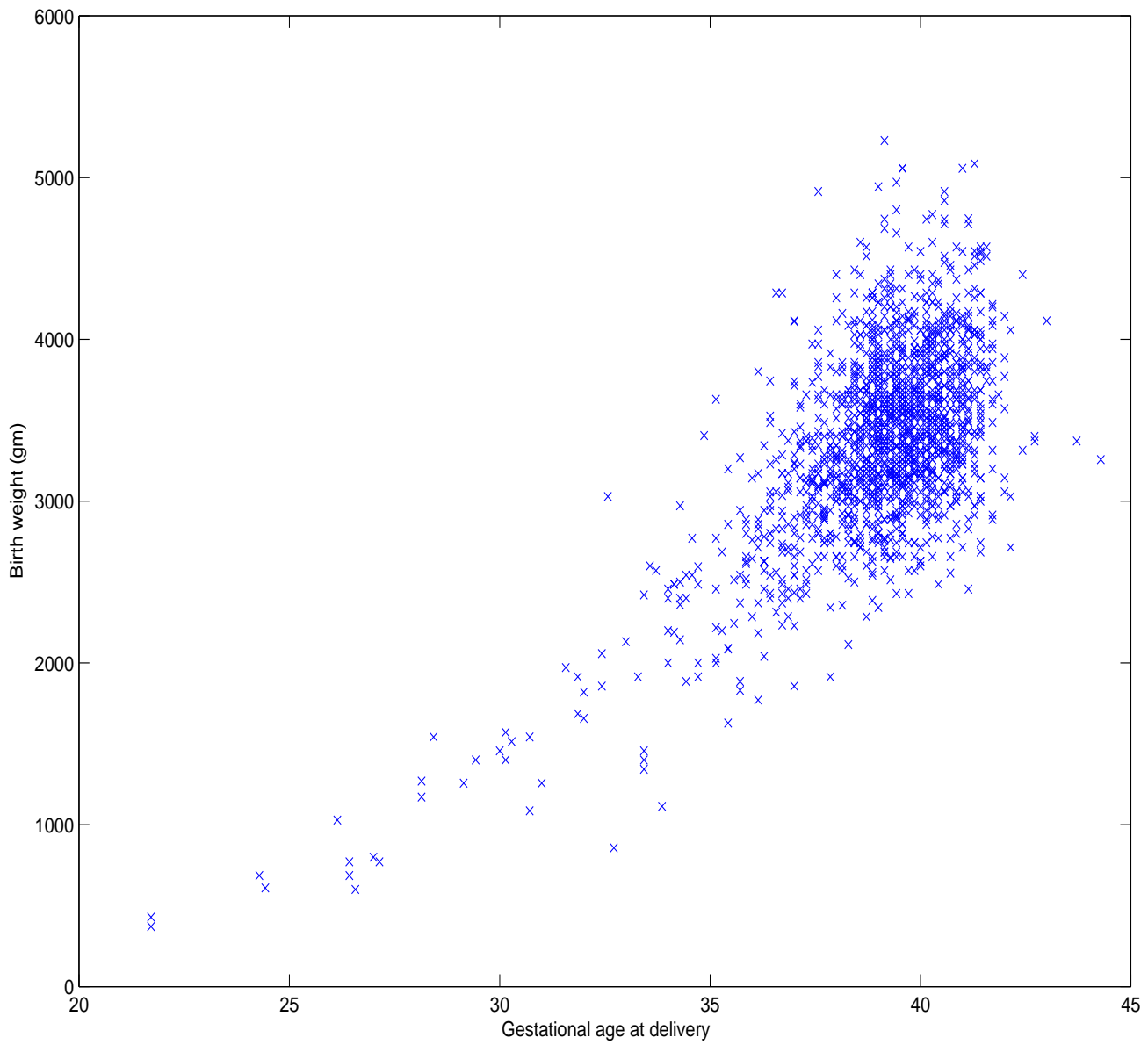


Figure 2: The gestational age at delivery for 1742 women in the RFTS study along with the birth weight of the baby.

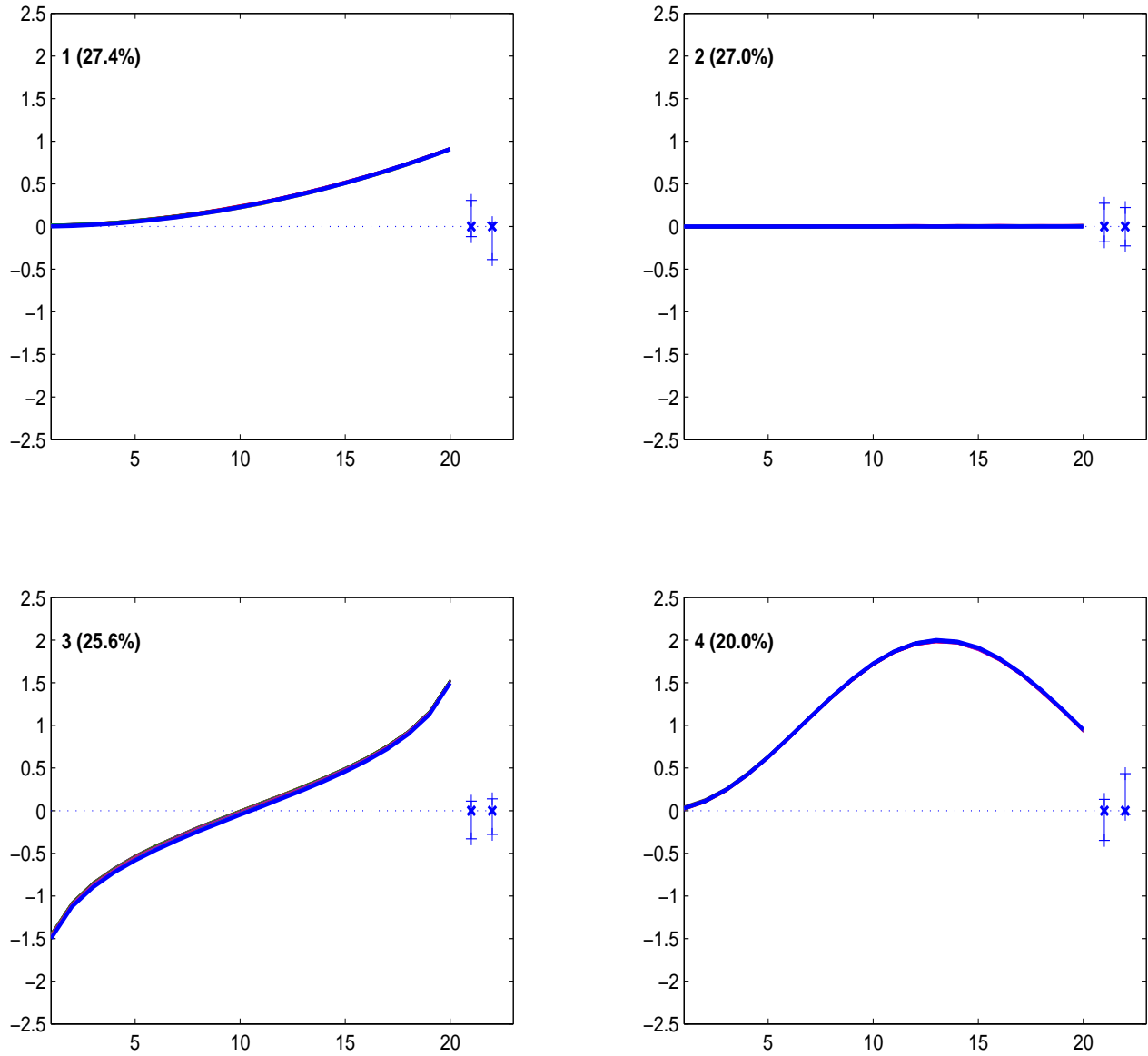


Figure 3: True (red) and estimated (blue) cluster-specific trajectories in simulation case i, along with percentage allocated to each cluster, and 99% credible intervals for cluster-specific mean of bivariate response distribution (+s). True means are \mathbf{x} .

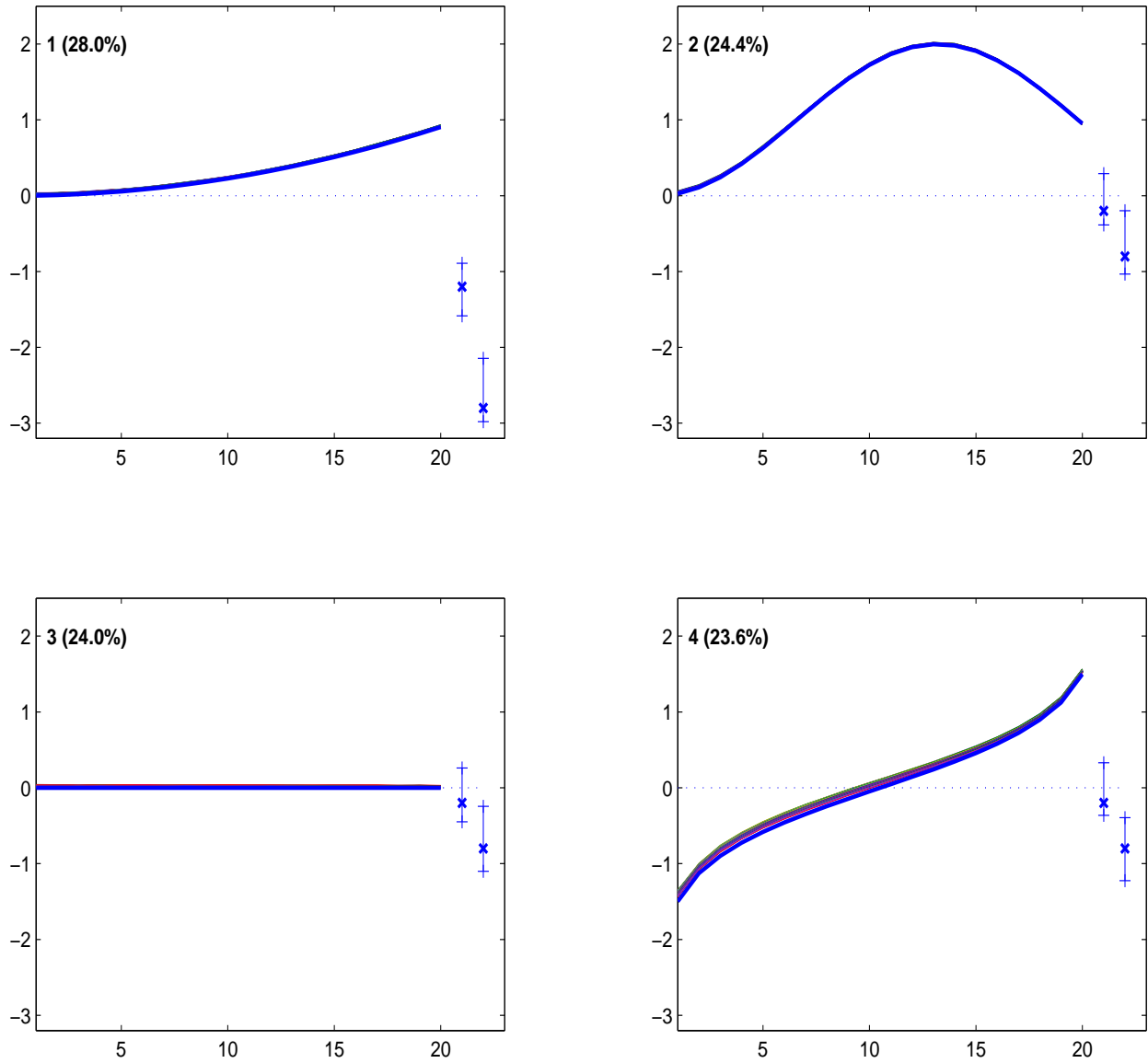


Figure 4: True (red) and estimated (blue) cluster-specific trajectories in simulation case ii, along with percentage allocated to each cluster, and 99% credible intervals for cluster-specific mean of bivariate response distribution (+s). True means are \mathbf{x} .

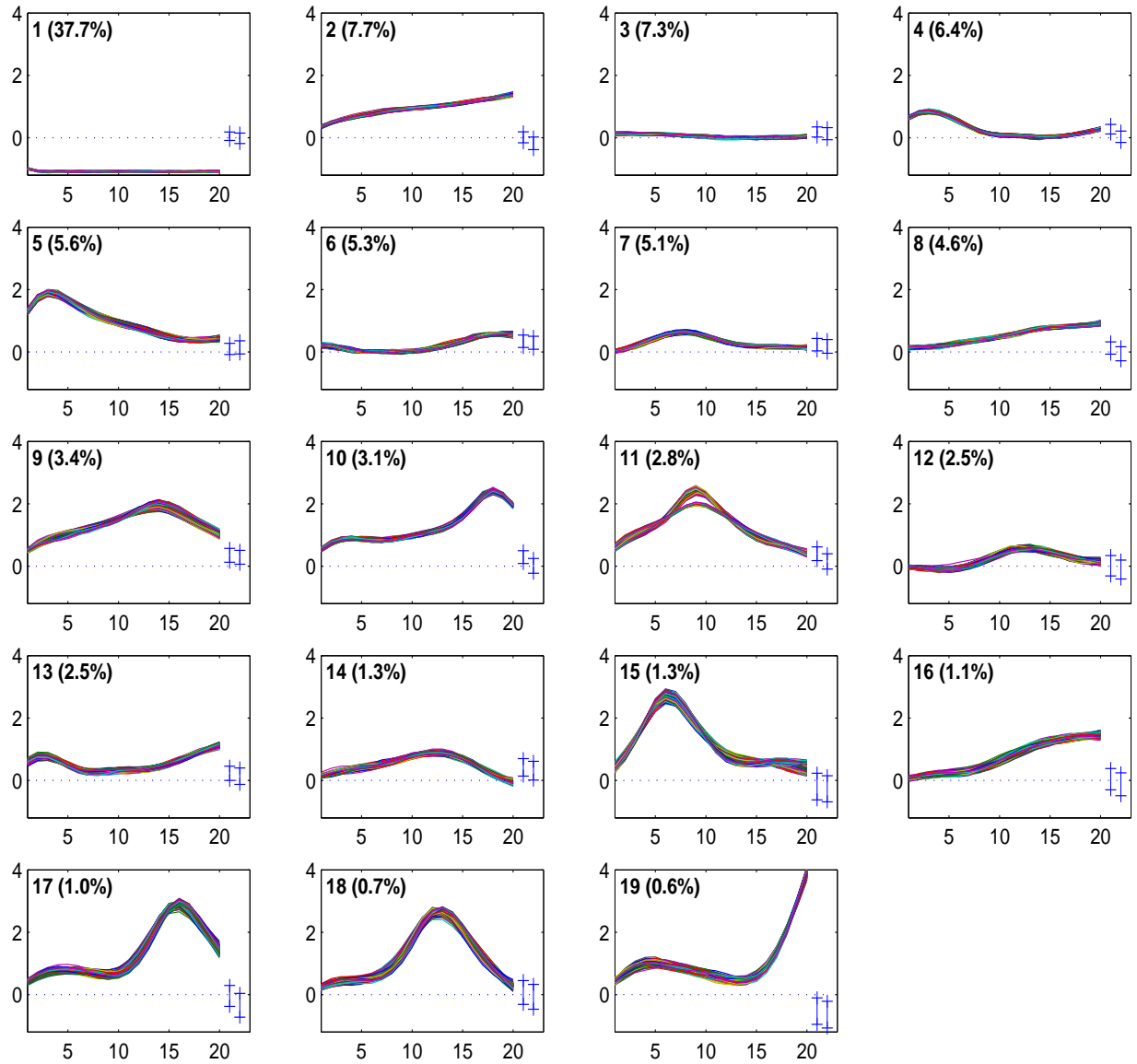


Figure 5: 100 samples from the posterior distribution for the latent trajectory curves (lines) along with 95% credible intervals for the cluster-specific shift in the mean gestational age at delivery & the birth weight (respective +’s)