Biometrics XX, 1–29 DOI: XXXXXXXXXX
XXXXX 20XX

Bayesian Latent Factor Regression for Functional and Longitudinal Data

Silvia Montagna[1,*], Surya T. Tokdar[1], Brian Neelon[2], and David B. Dunson[1]

[1]Department of Statistical Science, Duke University, Durham, NC 27708, U.S.A.

[2]Children's Environmental Health Initiative, Nicholas School of the Environment, Duke University, Durham, NC 27708, U.S.A.

email: sm234@duke.edu

Summary: In studies involving functional data, it is commonly of interest to model the impact of predictors on the distribution of the curves, allowing flexible effects on not only the mean curve but also the distribution about the mean. Characterizing the curve for each subject as a linear combination of a high-dimensional set of potential basis functions, we place a sparse latent factor regression model on the basis coefficients. We induce basis selection by choosing a shrinkage prior that allows many of the loadings to be close to zero. The number of latent factors is treated as unknown through a highly-efficient, adaptive-blocked Gibbs sampler. Predictors are included on the latent variables level, while allowing different predictors to impact different latent factors. This model induces a framework for functional response regression in which the distribution of the curves is allowed to change flexibly with predictors. The performance is assessed through simulation studies and the methods are applied to data on blood pressure trajectories during pregnancy.

Key words: Factor analysis; Functional principal components analysis; Latent trajectory models; Random effects; Sparse data.

1. Introduction

Functional data analysis (FDA) models variables that can be viewed as curves, surfaces or more general functions (Ramsey and Silverman, 2005). Examples include biomarker trajectories, images, videos, genetic codes and hurricane tracks. Although the whole curve itself is not typically observed, a sufficiently high number of measurements per individual is available. Often, these measurements are recorded at identical, equally spaced, locations or time points for all subjects. Functional data observed at irregular time points per subject are usually referred to as longitudinal data. Longitudinal data are often sparse with few measurements per subject. A detailed analysis of the perspectives of FDA and longitudinal data analysis (LDA), as well as a comparison of their methods, can be found in Rice (2004).

Because functional data are infinite dimensional, their statistical analysis necessitates obtaining a low dimensional representation of the individual curves. This becomes absolutely crucial for building a hierarchical model where the curves are to be related to other covariates recorded on the same subjects. A rich framework has been developed in the form of functional principal component analysis (FPCA) (Besse and Ramsay, 1986; Rice and Silverman, 1991; Cardot, 2006), which extends the principal component analysis to deal with infinite dimensional smooth curves, and its extensions to longitudinal data (James, Hastie and Sugar, 2000; Rice and Wu, 2001; Yao, Müller and Wang, 2005). In FPCA, the individual curves are represented by a vector of coefficients with respect to a common functional basis determined from the observed data. By truncating the basis representation at a finite depth, a low dimensional vector of scores is obtained for each subject. Once the common basis is identified and a truncation depth is chosen, all variations between the subject specific curves are reflected through the variations in the score vectors.

The existing literature on FPCA, however, offers little when it comes to incorporating covariates. Some contributions are made in Chiou, Müller and Wang (2003) and Cardot

(2006). However, these approaches are not applicable to sparse longitudinal data, since they rely on the assumption that either the entire curve is observed or that a high number of measurements per subject is available. This has been partially resolved by Jian and Wang (2010), who apply FPCA to curves viewed as functions over their original observation domain augmented with the covariate space. Although extremely flexible, the Jian-Wang approach requires an additional smoothing over the covariate space and can face serious practical difficulties when the covariate dimension is not minuscule. Additional difficulties arise when a direct inference on the relative importance of the covariates is desired. Moreover, the non-parametric, smoothing-based incorporation of the covariates may pose a challenge to robust prediction for future subjects with covariate values near the boundary of the observed data.

Covariate information can be incorporated through some non-FPCA approaches such as latent trajectory models and functional mixed effects models (Nagin, 1999; Lin et. al, 2001; Jones, Nagin and Roeder, 2001; James and Sugar, 2003). Functional mixed effect models extend linear mixed effects models to functional data, and can identify the group-mean trajectories while allowing flexible subject-specific deviations for the mean curve. However, parametric functions may not be flexible enough to capture both the group-mean trajectory and the subject-specific deviations, and nonparametric functions could be needed in such cases. Moreover, it is often desirable to adopt the same functional form for both the group-mean curve and the subject-specific deviations so that they share the same smoothness properties. In addition, these functional models are demanding computationally and difficult to extend to large data sets. On the other hand, latent class trajectory models can be sensitive to the parametric assumptions made about the trajectory cluster-specific response distributions in applications relating time-varying predictors to responses, with violation of the assumptions leading to an increase in the number of clusters and incorrect inferences about cluster-specific response densities. A further limitation of these models is that they

generally assume that a subject assigned to a specific functional cluster is also assigned to the corresponding response cluster.

Recently, Crainiceanu and Goldsmith (2010) proposed methods and software to implement Bayesian analyses using WinBUGS for a variety of functional models represented as mixed effect models. Their methods can be applied to unbalanced and unequally spaced data but particular attention must be paid to the estimation of the covariance operator. In particular, to gain good performance in predicting the underlying curves, the covariance function is estimated using method of moments for any observed pair (t, s) using all available data and then smoothed using penalized spline smoothing (see Staniswalis and Lee, 1998). However, a drawback of performing FPCA using the eigenfunctions obtained from a smooth estimator of the covariance matrix consists in potentially attributing the variability in the estimates to the variability among the underlying curves. In addition, FPCA does not let one to infer the number of eigenfunctions to retain for computation; this has to be chosen using a restricted likelihood ratio test for step-wise testing for zero-variance components, by cross-validation or via Akaike's Information Criterion. And finally, the estimation of the covariance operator is performed without taking into account available covariate information.

The goal of this paper is to provide a new Bayesian latent factor model for functional data. The curve for each subject is characterized as a linear combination of a high-dimensional set of basis functions and a sparse latent factor regression model is placed on the basis coefficients. Several methods have been proposed in the literature to infer the number of latent factors: Lopes and West (2004) propose a reversible jump Markov Chain Monte Carlo algorithm (RJMCMC) whereas Carvalho et al. (2008) allow uncertainty in the locations of the zeros of the factor loading matrix and in the number of factors using Bayesian variable selection methods. We follow an alternative approach due to Bhattacharya and Dunson (2011), who place a shrinkage prior on the loading coefficients that induces basis

selection. The number of latent factors is allowed to be unknown and is estimated using an adaptive blocked Gibbs sampler (Bhattacharya and Dunson, 2011). In contrast to Jian and Wang (2010), our model preserves the modeling goal of FPCA, that is identifying a common basis and assigning low-dimensional scores to individuals with respect to this basis. Variations among individual trajectories are then reflected in the variations between their scores. Furthermore, within our framework, it becomes possible to study the dependence of the curve shapes on covariates.

Our research was motivated by the Healthy Pregnancy, Healthy Baby (HPHB) study, an ongoing prospective cohort study examining the effects of environmental, social, and host factors on racial disparities in pregnancy outcomes. The study is part of the US EPA-funded Southern Center on Environmentally Driven Disparities in Birth Outcomes and enrolls pregnant women from the Duke Obstetrics Clinic and the Durham County Health Department Prenatal Clinic. Demographic, health behavior, and medical history data were obtained by direct patient interview and through electronic medical record review at the time of enrollment. Information on events of the pregnancy, labor, delivery, and health of the neonate were ascertained from maternal and neonatal electronic medical records. Measurements of mean arterial blood pressure (MAP = 2/3 diastolic pressure + 1/3 systolic pressure) during pregnancy were available for 1,027 women for a total of 10,290 measurements. It is of interest to examine the differences in blood pressure trajectories among women and analyze the effect of covariates, e.g., pre-gestational diabetes, parity, maternal age, on average MAP curves for specific groups.

The rest of the paper is organized as follows: Section 2 outlines the functional latent factor regression model (LFRM). Section 3 extends the LFRM to allow joint modeling of a functional response and additional outcomes. Section 4 describes simulation studies for

realistic settings. Section 5 describes the application of our methodology to the blood pressure data. Conclusions are presented in Section 6.

2. Functional latent factor regression model

Suppose that y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, n_i)$ is the response of the ith subject at point t_{ij} (where t is an index such as time or distance). For example, blood pressure measurements during pregnancy may be recorded for patient i at different visits, j's, to a clinic. One may model these data as follows:

$$y_{ij} = f_i(t_{ij}) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathrm{N}(0, \varphi^2) \tag{1}$$

$$f_i(t_{ij}) = \sum_{l=1}^{p} \theta_{il} b_l(t_{ij}) \tag{2}$$

where y_{ij} is the blood pressure measurement for patient i at visit j, f_i is the smooth curve for subject i, ϵ_{ij} is a residual error in the measurement at time t_{ij}, $b_l(t_{ij})$ corresponds to the lth basis function evaluated at time j, θ_{il} is the subject-specific basis function coefficient and φ^2 is the measurement error variance. Characterizing the individual curves by a vector of coefficients with respect to a common functional basis representation with appropriately chosen truncation p, a low dimensional vector of scores is obtained for each subject, and variations between the subject specific curves are reflected through the variations in the score vectors. The choice of the basis functions, which are held fixed in the model, is particularly challenging since the appropriate basis to use is not known in advance. In general, the choice depends on the particular application at hand and, conceptually, any basis function can be chosen. In the blood pressure application, where smooth curves are expected, one would like local basis functions chosen to be sufficiently many and with sufficiently narrow kernels to capture a very high variety of smooth curves without allowing overly-spiky curves. Therefore, after standardizing the time to the $[0, 1]$ interval, $t_{ij} \in [0, 1]$, the basis functions are defined

as

$$b_1(t_{ij}) = 1, \tag{3}$$

$$b_{l+1}(t_{ij}) = \exp(-\nu||t_{ij} - \psi_j||^2), \tag{4}$$

thus $b_l, l = 1, \ldots, p-1$, are set to be fixed Gaussian kernels with $\psi_1, \ldots, \psi_{p-1}$ equally spaced kernel locations and ν denotes the bandwidth of the kernel. We choose an over-complete set of candidate basis functions, with the expectation being that many of these functions are not needed but that a very rich variety of curves can be characterized as a sparse and adaptive linear combination of a rich pre-specified set of potential basis functions.

Denoting as $\mathbf{y}_i$ a n_i-dimensional continuous response, we characterize the individual functions by a linear combinations of a fixed number p of Gaussian kernels:

$$\mathbf{y}_i = \boldsymbol{f}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \psi^2 \boldsymbol{I}_{n_i}), \qquad \text{and} \tag{5}$$

$$f_i(t_{ij}) = \sum_{l=1}^{p} \theta_{il} b_l(t_{ij}) = \mathbf{B}_i(t_{ij})\boldsymbol{\theta}_i \tag{6}$$

with $\mathbf{B}_i(t_{ij})$ denoting the jth row of the design matrix $\mathbf{B}_i$ for subject i with basis functions defined in (4). In addition, a sparse latent factor model is specified for the basis coefficients. In particular, consider the subject-specific vector of coefficient $\boldsymbol{\theta}_i$ related to the continuous latent variables $\boldsymbol{\eta}_i$ for subject i,

$$\boldsymbol{\theta}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i, \qquad \boldsymbol{\zeta}_i \sim \mathrm{N}_p(0, \boldsymbol{\Sigma}), \tag{7}$$

where $\boldsymbol{\Lambda}$ is a $p \times k$ factor loading matrix and $\boldsymbol{\epsilon}_i$ is a residual vector that is uncorrelated with other variables in the model and is normally distributed with mean zero and diagonal covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Modeling the high-dimensional set of basis coefficients for subject i, $\boldsymbol{\theta}_i$, as a sparse linear combination of latent factors induces dimensionality reduction. This structure allows subjects to have basis coefficients near zero for many of the basis functions, inducing subject-specific basis selection, while also allowing certain basis functions to effectively drop out for all subjects. Covariate information is included in the

model via the continuous latent variables $\boldsymbol{\eta}_i$'s by the relation

$$\boldsymbol{\eta}_i = \boldsymbol{\beta}'\boldsymbol{x}_i + \boldsymbol{\delta}_i, \qquad \boldsymbol{\delta}_i \sim \mathrm{N}_k(0, \boldsymbol{I}), \tag{8}$$

where $\boldsymbol{x}_i$ is a $r \times 1$ vector of predictors for subject i (e.g., diabetes, maternal age, parity, etc.), $\boldsymbol{\beta}$ is a $r \times k$ matrix of coefficients and $\boldsymbol{\delta}_i$ is a normally distributed residual vector with identity covariance matrix. This model induces a framework for functional response regression in which the distribution of the curves is allowed to change flexibly with predictors.

Note the similarity with the "random global basis / random subject-specific coefficients" construction of FPCA:

$$f_i(t_{ij}) = \sum_{m=1}^{k} \eta_{im}\phi_m(t_{ij}) + r_i(t_{ij}) \tag{9}$$

with $\phi_m(t_{ij})$ being in our case

$$\phi_m(t_{ij}) = \sum_{l=1}^{p} \lambda_{lm} b_l(t_{ij}). \tag{10}$$

While this corresponds to the FPCA construction (with the eigenfunctions in the linear span of the fixed basis $b_l(t_{ij})$, and where orthogonality is not imposed), we develop a computationally attractive platform by modeling the residual function $r_i(t_{ij})$ as

$$r_i(t_{ij}) = \sum_{l=1}^{p} \zeta_{il} b_l(t_{ij}) \tag{11}$$

which leads to the latent factor model structure given by equations (2) - (7). Therefore, as in FPCA, we obtain a low dimensional representation of the individual curves (with respect to a basis learned from the data), but we utilize the attractive framework of latent factor models for the computation.

We induce dependence among the $\mathbf{y}_i$ by marginalizing over the distribution of the factors and the basis coefficients. Under this model, the n_i-dimensional continuous response $\mathbf{y}_i$ is marginally distributed as

$$\mathbf{y}_i | \mathbf{B}_i, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \boldsymbol{\beta}', \boldsymbol{x_i}, \psi^2 \sim \mathrm{N}(\mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\beta}'\boldsymbol{x}_i, \mathbf{B}_i\boldsymbol{\Omega}\mathbf{B}_i' + \psi^2\mathrm{I}_{n_i}) \tag{12}$$

with $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}$. Therefore, the smooth function $\boldsymbol{f}_i$ for subject i is given a Gaussian

process with mean $\mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\beta}'\boldsymbol{x}_i$ and covariance function $\mathbf{B}_i\boldsymbol{\Omega}\mathbf{B}_i'$

$$\boldsymbol{f}_i \sim \mathrm{GP}(\mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\beta}'\boldsymbol{x}_i, \mathbf{B}_i\boldsymbol{\Omega}\mathbf{B}_i') \tag{13}$$

Our structure is preferred to hierarchical Gaussian process models for curves (i.e., Behseta et al., 2005) because we are effectively learning the covariance kernel and a sparse representation for the mean function, as opposed to Crainiceanu and Goldsmith (2010) who, by performing FPCA using the eigenfunctions obtained from a smooth estimator of the covariance matrix, could mistakenly attribute the variability in the estimates to the variability among functions $\boldsymbol{f}_i$.

2.1 *Bayesian formulation, prior elicitation and posterior computation*

A Bayesian formulation of our sparse latent factor model is completed with priors for the parameters in (5)-(8). Given the dimensionality, it is practically important to choose conditionally conjugate priors that lead to efficient posterior computation via blocked Gibbs sampling. Typical priors for factor analysis constrain $\boldsymbol{\Lambda}$ to be lower triangular with positive diagonal entries using normal and truncated normal priors for the free elements of $\boldsymbol{\Lambda}$ and gamma priors for the residual precisions (Arminger, 1998; Geweke and Zhou, 1996; Aguilar and West, 2000; Lopes and West, 2004). However, following Bhattacharya and Dunson (2011) we note that such constraints are unnecessary and unappealing in leading to order dependence and computational inefficiencies. Hence, we follow their lead in using a multiplicative gamma process shrinkage (MGPS) prior for the loadings as follows:

$$\lambda_{jh}|\phi_{jh}, \tau_h \sim \mathrm{N}(0, \phi_{jh}^{-1}\tau_h^{-1}), \quad \phi_{jh} \sim \mathrm{Gamma}(\upsilon/2, \upsilon/2), \quad \pi_h = \prod_{l=1}^{h} \delta_l \tag{14}$$

$$\delta_1 \sim \mathrm{Gamma}(a_1, 1), \qquad \delta_l \sim \mathrm{Gamma}(a_2, 1), \quad l \geqslant 2 \tag{15}$$

$j = 1, \ldots, p$, $h = 1, \ldots, k$, $\delta_l, l \geqslant 1$, are independent, π_h is a global shrinkage parameter for the hth column and ϕ_{jh}'s are local shrinkage parameters for the elements in the hth column. Under a choice $a_2 > 1$, the π_h's are stochastically increasing favoring more shrinkage as

the column index increases. The choice of this shrinkage prior allows many of the loadings to be close to zero, thus inducing effective basis selection. The number of latent factors, k, is also treated as unknown through a highly-efficient adaptive blocked Gibbs sampler described in Bhattacharya and Dunson (2011). Therefore, an additional attractive feature of our formulation with respect to FPCA is that we can infer the representation size k. The prior structure under our model is completed by

$$\sigma_j^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma), \qquad j = 1, \ldots, p, \tag{16}$$

$$\psi^{-2} \sim \text{Gamma}(a_\psi, b_\psi), \tag{17}$$

and the update of the matrix of coefficients $\boldsymbol{\beta}$ is performed as follows. Consider

$$\begin{pmatrix} \eta_{ij} \\ \vdots \\ \eta_{nj} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \times \boldsymbol{\beta}_j + \boldsymbol{\delta} \tag{18}$$

$$\boldsymbol{\eta}_{\cdot j}' \sim \text{N}(\tilde{\mathbf{X}}'\boldsymbol{\beta}_j, \text{I}_n) \tag{19}$$

where $\boldsymbol{\eta}_{\cdot j}'$ denotes the jth column of the $n \times k$ transpose of the matrix of latent factors $\boldsymbol{\eta}$, $\boldsymbol{\beta}_j$ denotes the jth column of the $r \times k$ matrix of coefficients $\boldsymbol{\beta}$ and $\tilde{\mathbf{X}}'$ denotes the transpose of the matrix of predictors $\tilde{\mathbf{X}}$. Each row $i, i = 1, \ldots, n$, of $\tilde{\mathbf{X}}'$ corresponds to the vector of predictors for subject i, $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ir})$. A Cauchy prior is induced on the matrix of coefficients $\boldsymbol{\beta}$ as follows

$$\boldsymbol{\beta}_j \sim \text{N}(0, \text{Diag}(\omega_{lj}^{-1})), \qquad j = 1, \ldots, k, \qquad l = 1, \ldots, r \tag{20}$$

$$\omega_{lj} \sim \text{Gamma}(1/2, 1/2). \tag{21}$$

Alternatively, one could choose a Gaussian prior distribution for the $\boldsymbol{\beta}$ coefficients but this leads to poorer performance if a subsample of patients has very sparse measurements. This occurrence is common when dealing with longitudinal data, which often consist of few and sparse measurements per subject, and it is verified in the blood pressure data where a group

of women has few observations, usually located in the second half of the pregnancy. For this group of women, the prior becomes more influential and the intercept is pulled closer to zero than for women with more observations, resulting in an undesired low MAP trajectory estimate at early pregnancy.

The posterior computation is similar to the Markov Chain Monte Carlo (MCMC) algorithm for the sparse Bayesian infinite factor model in Bhattacharya and Dunson (2011). Details of the algorithm are provided in the Appendix.

3. Joint model

It is of interest to extend the model in Section 2 to allow joint modeling of a functional response, i.e. the blood pressure trajectories, and one or more additional outcomes. For example, in the study of blood pressure (BP) trajectories during pregnancy, there is substantial interest in relating these trajectories to subsequent pregnancy outcomes, such as gestational age (GA) at delivery and birth weight (BW). By jointly modeling the trajectory in blood pressure and the pregnancy outcomes, one can obtain conditional distributions of interest, such as the conditional density of birth weight adjusted for gestational age at delivery and the trajectory in blood pressure across pregnancy. In addition, covariates can be used to predict blood pressure and/or the pregnancy outcomes.

We start with a simple extension of our model where we include a binary indicator for preterm delivery and allow the probability of premature delivery to depend on the latent factors, $\boldsymbol{\eta}_i$'s. A bivariate probit model for preeclampsia and low birth weight is outlined in Section 3.2, and a joint model for BW, GA and MAP is considered in Section 3.3.

3.1 *Probit model for risk of preterm birth*

Preterm birth refers to the birth of a baby of less than 37 weeks gestational age. Let $y_i = 1$ if preterm birth and $y_i = 0$ if full-term birth. We let $\mathrm{P}(y_i = 1|\alpha, \boldsymbol{\gamma}, \boldsymbol{\eta}_i) = \Phi(\alpha + \boldsymbol{\gamma}'\boldsymbol{\eta}_i)$,

where $\Phi(\cdot)$ denotes the standard normal distribution function. α is an intercept, with prior distribution $\mathrm{N}(\Phi^{-1}(0.123), 0.25)$, where the hyperprior mean was chosen to correspond to the national average of 12.3% in 2008 (Hamilton et al., 2010); $\boldsymbol{\eta}_i$ are the latent factors for subject i; and $\boldsymbol{\gamma}$ is a $k \times 1$ vector of unknown regression coefficients with normal prior distribution, $\boldsymbol{\gamma} \sim \mathrm{N}_k(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$. Therefore, the same set of latent factors impacts on the functional predictor via the basis coefficients $\boldsymbol{\theta}_i$ and on the response variables via the probability of preterm birth.

The full conditional posterior distributions needed for Gibbs sampling are not automatically available, but we can rely on the data augmentation algorithm of Albert and Chib (1993) to facilitate the computation:

$$y_i = \mathbb{1}(W_i > 0)$$

$$W_i \sim \mathrm{N}(\alpha + \boldsymbol{\gamma}'\boldsymbol{\eta}_i, 1),$$

and $\mathrm{P}(y_i = 1|\alpha, \boldsymbol{\gamma}, \boldsymbol{\eta}_i) = \Phi(\alpha + \boldsymbol{\gamma}'\boldsymbol{\eta}_i)$ by marginalizing out W_i.

The MCMC algorithm for the latent factor regression model (LFRM) described in the Appendix is augmented with additional steps to sample from the full conditional posterior distributions of $\alpha, \boldsymbol{\gamma}$, and $W_i, i = 1, \ldots, n$ ((multivariate) normal and truncated normal distributions, respectively), and the update of the latent factors $\boldsymbol{\eta}_i$ described in Step 5 of the Appendix is modified accordingly.

3.2 *Bivariate probit model for preeclampsia and low birth weight*

We developed a bivariate probit model to study the relationship between preeclampsia (hypertension and elevated urine proteins at time of delivery), low birth weight (LBW - weight under 2500 grams) and maternal MAP measured at prenatal clinical visits. The sample proportion of LBW was 12%, thus slightly higher than the corresponding national rate of 8.2% in 2008 (Hamilton et al., 2010), whereas the sample proportion of preeclamptic women was 16%, far above the incidence of preeclampsia which typically affects 5-8% of all pregnancies (Cunningham et al., 2001).

Let us denote the outcome variables for preeclampsia and LBW as z_p^i and z_{lbw}^i, respectively. In particular

$$z_p^i = \begin{cases} 1 & \text{if subject } i \text{ develops preeclampsia} \\ 0 & \text{otherwise} \end{cases}$$

and

$$z_{lbw}^i = \begin{cases} 1 & \text{if birth weight is under 2500 grams} \\ 0 & \text{otherwise} \end{cases}$$

As in Section 3.1, we adopt a data augmentation approach and introduce two underlying normal variables, W_p^i and W_{lbw}^i, such that

$$z_p^i = \mathbb{1}(W_p^i > 0) \qquad \text{and} \qquad z_{lbw}^i = \mathbb{1}(W_{lbw}^i > 0)$$

We assume that W_p^i and W_{lbw}^i follow a bivariate normal distribution

$$\begin{pmatrix} W_p^i \\ W_{lbw}^i \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} \alpha_1 + \boldsymbol{\gamma}_1 \boldsymbol{\eta}_i \\ \alpha_2 + \boldsymbol{\gamma}_2 \boldsymbol{\eta}_i \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

with ρ controlling the dependence between z_p^i and z_{lbw}^i.

The joint probability of preeclampsia and LBW is obtained by double integration of the bivariate normal distribution of the latent variables W_p^i and W_{lbw}^i

$$\Pr(z_p^i = 1, z_{lbw}^i = 1) = \int_0^\infty \int_0^\infty \mathrm{N}_2(W_p^i, W_{lbw}^i; \boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}) dW_p^i, dW_{lbw}^i$$

with $\boldsymbol{\mu} = (\alpha_1 + \boldsymbol{\gamma}_1 \boldsymbol{\eta}_i, \alpha_2 + \boldsymbol{\gamma}_2 \boldsymbol{\eta}_i)'$ and $\tilde{\boldsymbol{\Sigma}} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix} \right)$. Analogously, we can compute the marginal probability of observing preeclampsia and the marginal probability of LBW.

We complete the Bayesian specification of the bivariate probit model by choosing normal and multivariate normal priors for the parameters, that is $\alpha_1 \sim \mathrm{N}(\Phi^{-1}(0.12), 0.25)$, $\alpha_2 \sim \mathrm{N}(\Phi^{-1}(0.082), 0.25)$, $\boldsymbol{\gamma}_1 \sim \mathrm{N}_k(\boldsymbol{\mu}_{\gamma,1}, \boldsymbol{\Sigma}_{\gamma,1})$ and $\boldsymbol{\gamma}_2 \sim \mathrm{N}_k(\boldsymbol{\mu}_{\gamma,2}, \boldsymbol{\Sigma}_{\gamma,2})$. The hyperprior mean for α_1 was set to be moderately high provided that the proportion of preeclamptic women in the sample is over twice the typical incidence range of 5-8%, and that of α_2 was chosen to

correspond to the national average.

The MCMC algorithm for the LFRM described in the Appendix now incorporates additional steps to update $\alpha_1, \alpha_2, \boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ sampling from their (multivariate) normal full conditional distributions, whereas $W_j^i, j = \{p, lbw\}$, is sampled from its full conditional normal distribution truncated below (above) by zero for $z_j^i = 1(z_j^i = 0)$. The sampler also includes a random-walk Metropolis-Hastings step for the update of the correlation coefficient ρ, with proposal density restricted to (-1, 1).

Heterogeneity across subjects and dependence between the smooth function, $\boldsymbol{f}_i$, and the outcomes, z_p^i and z_{lbw}^i, is accommodated through the latent factors, $\boldsymbol{\eta}_i$, which impact on the MAP measurements via the basis coefficients $\boldsymbol{\theta}_i$ and on the probabilities of preeclampsia and LBW via the latent normal variables W_p^i and W_{lbw}^i.

Our goal is to compare sequential predictions of the probability of preeclampsia and LBW for a test sample of women at different times during gestation, say $20th$ week, $25th$ week, $30th$, etc. Predictions are expected to improve over time, and we aim to assess whether we can make a detection with some certainty sufficiently early during gestation or if it is necessary to wait until close to delivery to make an accurate prediction.

3.3 *Joint model of birth weight, gestational age at delivery and blood pressure*

Let $\mathbf{z}_i$ denote the outcome for subject i, $\mathbf{z}_i = (z_{ib}, z_{ig})$, with z_{ib} denoting the birth weight and z_{ig} the gestational age at delivery for subject i. To flexibly joint model gestational age at delivery and birthweight, we consider a two-component mixture-model of bivariate normal distributions as described by McLachlan and Peel (2000) and Marin, Mengersen and Casella (2005)

$$(z_{ig}, z_{ib}) \sim \sum_{h=1}^{2} \pi_{ih} \mathrm{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \tag{22}$$

This model can be equivalently specified as

$$(z_{ig}, z_{ib}) \sim \mathrm{N}(\boldsymbol{\mu}_{T_i}, \boldsymbol{\Sigma}_{T_i}) \tag{23}$$

$$T_i = \mathbb{1}(W_i > 0) \tag{24}$$

where $T_i \in \{1, 2\}$ is a latent variable indicating which class (z_{ig}, z_{ib}) belong to, and $\mathrm{P}(T_i = h) = \pi_{ih}$. We now let the W_i's be independent distributed from t distributions using a scale mixture of normals construction:

$$W_i \sim \mathrm{N}\left(\alpha + \boldsymbol{\gamma}'\boldsymbol{\eta}_i, \tilde{\sigma}^2\tilde{\phi}_i^{-1}\right) \tag{25}$$

$$\tilde{\phi}_i \sim \mathrm{Gamma}(\tilde{\nu}/2, \tilde{\nu}/2) \tag{26}$$

where $\boldsymbol{\gamma}$ a $k \times 1$ vector of unknown regression coefficients with normal prior distribution, $\boldsymbol{\gamma} \sim \mathrm{N}_k(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$, $\boldsymbol{\eta}_i$ are the latent factors for subject i and $\alpha \sim \mathrm{N}(\Phi^{-1}(0.1), 0.25)$. Note that construction (22) - (25) constitutes a t approximation to a logit link function on the mixing weights π_{ih}, and to ensure a good approximation to the univariate logistic distribution we set $\tilde{\sigma}^2 \equiv \pi^2(\tilde{\nu} - 2)/3\tilde{\nu}, \tilde{\nu} \equiv 7.3$ (Albert and Chib, 1993; O'Brien and Dunson, 2004). In addition, this approximation ensures conjugacy of the full conditional distributions, thus allowing efficient posterior update. To complete our Bayesian specification, we choose an inverse-Wishart (I-W) distribution for the covariance matrix $\boldsymbol{\Sigma}_h$, $\boldsymbol{\Sigma}_h \sim \text{I-W}_2(\nu_h, \mathbf{V}_h)$ and a bivariate normal distribution for the mean $\boldsymbol{\mu}_h$, $\boldsymbol{\mu}_h \sim \mathrm{N}_2(\boldsymbol{\mu}_0^h, \boldsymbol{\Sigma}_{\mu 0}^h)$. Studies on birth weight and gestational age usually identify LBW and short gestational age as birth weight < 2.5 Kg and gestational length < 37 weeks (premature), respectively. We applied an EM algorithm MLE using a two-component mixture of bivariate normals to the data without including covariate information to determine the hyperparameters $\boldsymbol{\mu}_0^h$ and $\boldsymbol{\Sigma}_{\mu 0}^h$. The prior mean values resulted in

$$\boldsymbol{\mu}_0^1 = \begin{pmatrix} \mu_{0g}^1 \\ \mu_{0b}^1 \end{pmatrix} = \begin{pmatrix} 36 \\ 2.57 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\mu}_0^2 = \begin{pmatrix} \mu_{0g}^2 \\ \mu_{0b}^2 \end{pmatrix} = \begin{pmatrix} 39 \\ 3.30 \end{pmatrix}$$

with covariance matrices

$$\boldsymbol{\Sigma}_{\mu 0}^{1} = \begin{pmatrix} 7.66 & 1.37 \\ 1.37 & 0.35 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\mu 0}^{2} = \begin{pmatrix} 1.34 & 0.19 \\ 0.19 & 0.22 \end{pmatrix}$$

The data consist of n pairs $(\mathbf{y}_i, \mathbf{z}_i)_{i=1}^n$ and the joint likelihood, conditional on the latent factors $\boldsymbol{\eta}$, can be factored as the product of

$$L(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \propto \varphi^{-N} \exp\left\{ -\frac{\varphi^{-2}}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{B}_i\boldsymbol{\theta}_i)'(\mathbf{y}_i - \mathbf{B}_i\boldsymbol{\theta}_i) \right\}$$
$$\times \prod_{i=1}^{n} \left\{ \sum_{h=1}^{2} \pi_{ih} \times |2\pi\boldsymbol{\Sigma}_h|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_h)'\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_h) \right) \right\}$$

with $N = \sum_{i=1}^{n} n_i$.

Heterogeneity among subjects and dependence between the smooth function, $\boldsymbol{f}_i$, and the outcomes, $\mathbf{z}_i = (z_{ig}, z_{ib})$, is accommodated through the common set of latent factors, $\boldsymbol{\eta}_i$, which impacts on the functional predictors via the basis coefficients, $\boldsymbol{\theta}_i$, and on the class membership probability of the pregnancy outcomes, $\pi_{i1}(\boldsymbol{\eta}_i) = \mathrm{P}(T_i = 1) = \Phi\left(\frac{\alpha + \boldsymbol{\gamma}'\boldsymbol{\eta}_i}{\sqrt{\tilde{\sigma}^2\tilde{\phi}_i^{-1}}}\right)$. In addition, given the structure for the latent factors described in Equation (8), different predictors impact on different latent factors letting the distribution of the curves, as well as the response class, change flexibly with predictors.

The MCMC algorithm for the LFRM can be straightforwardly modified to include steps for the update of the additional model parameters which are sampled from their full conditional distributions, and the update of the latent factors $\boldsymbol{\eta}_i$ is modified accordingly.

4. Simulation example

To evaluate the performance of our model and to compare it with related methods, we considered a simulation example. We assumed $n = 200$ and simulated data under the functional LFRM described in Section 2, with the true parameters set equal to the posterior means from the real data analysis (see Section 5 below). We generated samples of gestational age (in weeks) and birth weight (in Kg) from a two-component mixture of bivariate normal

distributions with true means set equal to $\boldsymbol{\mu}_1 = (34.54, 2.27)'$ and $\boldsymbol{\mu}_2 = (38.17, 3.50)'$ and covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.516 & 0.261 \\ 0.261 & 1.235 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.212 & 0.185 \\ 0.185 & 1.221 \end{pmatrix}$$

We standardized time to the $[0, 1]$ interval, $t_{ij} \in [0, 1]$, and set $b_1(t_{ij}) = 1$ and $b_{l+1}(t_{ij}) = \exp\{-4||t_{ij} - \psi_j||^2\}, l = 1, \ldots, 9$, with ψ_j's equally spaced kernel locations in $[0, 1]$ and $p = 10$. The choice of p resulted in a reasonable default value to ensure smooth curves, with sufficiently many equally-spaced kernels to capture a high variety of smooth trajectory shapes.

To implement our Bayesian analysis, we chose a Ga$(0.5, 0.25)$ prior distribution with mean 2 for the diagonal elements of $\boldsymbol{\Sigma}^{-1}$, and we placed a Ga(0.5, 0.2) with mean 2.5 on ψ^{-2}. The gamma hyperparameter for ϕ_{jh} was set to be $\upsilon = 5$, $a_1 = a_2 = 1.5$ in (14) and a Cauchy prior was induced on the matrix of coefficients $\boldsymbol{\beta}$ ((19) - (20)). We chose $k = 9$ as the starting number of factors, and we adapted k according to the procedure described in Bhattacharya and Dunson (2011). The MCMC algorithm was run for 25,000 iterations including a 5,000 iterations burn-in, and collected every 5th sample to thin the chain and reduce the autocorrelation in the posterior samples. Based on the examination of traceplots of function values at a variety of time locations and for different subjects, the sampler appeared to converge rapidly and to mix efficiently.

The average of the estimated number of factors was 11.18 corresponding to $k_{\text{true}} = 11$, and with empirical 95% credible interval given by $[9, 13]$. The estimated posterior mean of $\boldsymbol{\mu}_1$ was (34.39, 2.35) and the estimated posterior mean for $\boldsymbol{\mu}_2$ was $(37.90, 3.42)$ respectively, with corresponding 95% credible intervals containing the true values of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. The estimates of the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ were

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.436 & 0.424 \\ 0.424 & 1.103 \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.093 & 0.263 \\ 0.263 & 1.1523 \end{pmatrix}$$

with 95% credible intervals containing the true values of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

The left panels of Figure 1 show the data, true curves and estimates for the LFRM. In general, estimates are very close to the true curves even when data are sparse, as for subject 122, and the true curves are always enclosed in the credible bounds.

[Figure 1 about here.]

We then obtained a smooth estimator of the covariance operator and its corresponding eigenfunctions as described by Crainiceanu and Goldsmith (2010). In contrast to the LFRM, FPCA does not allow to learn about the representation size k, thus we need to estimate the dimension of the functional space. As a fast alternative to cross-validation, we decided to retain a number of eigenfunctions such that the cumulative percentage of explained variance was greater than 90% and the explained variance by any single subsequent component was less than 5%. Therefore, we retained the first $k = 4$ eigenfunctions and obtained $\boldsymbol{\Lambda}$ as the least squares estimate of

$$\boldsymbol{\Psi} = \mathbf{B}^* \times \boldsymbol{\Lambda} \tag{27}$$

with $\boldsymbol{\Psi}$ denoting here the matrix of eigenfunctions and $\mathbf{O}_i \times \mathbf{B}^* = \mathbf{B}_i$, $\mathbf{B}_i$ denoting the design matrix for subject i and $\mathbf{O}_i$ representing an $(n_i \times T)$ matrix with column j equal to a column of 1's if subject i was measured at time j, $j = 1, \ldots, T$ (T corresponds to the number of unique time locations). We then repeated the analysis fitting the LFRM with $\boldsymbol{\Lambda}$ and the number of factors $k = 4$ fixed. We will denote this procedure as the modified Crainiceanu-Goldsmith (mCG) approach. Estimates are shown in the right panels of Figure 1. We can notice some deviations of the estimated curves from the true curves along the course of the entire pregnancy, with very wide confidence intervals at early pregnancy when typically no or few measurements are observed and when data are more sparse, as for subject 122. Notice also that for subject 8 the true curve is no longer enclosed within the credible bounds at delivery. The analysis was repeated retaining $k_{true} = 11$ eigenfunctions, but this did not lead

to any significant improvement in the performance.

The estimated posterior mean of $\boldsymbol{\mu}_1$ was $(34.26, 2.31)$ and the estimated posterior mean for $\boldsymbol{\mu}_2$ was $(37.81, 3.39)$ respectively, with corresponding 95% credible intervals containing the true values of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. The estimates of the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ were

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.249 & 0.383 \\ 0.383 & 1.087 \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.227 & 0.305 \\ 0.305 & 1.165 \end{pmatrix}$$

To assess the predictive performance, we repeated the analysis holding out and predicting the MAP measurements collected after the $30th$ week of gestation for 100 randomly selected women having at least 1 observation in the first 30 weeks and 1 observation after the $30th$ week. The mean square predictive errors for the LFRM and the mCG approach were respectively 79.06 and 106.56, the predictive average absolute biases were 7.13 and 8.23, and the predictive maximum absolute biases 25.04 and 28.95. These high values are not surprising given the presence of many outliers in the MAP measurements that are hard to predict. The correlation coefficient between true and predicted values was 0.76 under the LFRM and 0.73 under the mCG approach, respectively. Therefore, the LFRM lead to an overall better predictive performance than the mCG approach.

Figure 2 shows the estimated joint distribution of gestational age (in weeks) and birth weight (in Kg) for subjects 8 and 46 in the simulation example under the LFRM, along with corresponding contour plots. The true values of gestational age at delivery and birth weight corresponded to (38.82, 3.47) and (33.51, 1.41) for subject 8 and subject 46, respectively. The joint distribution is bimodal, with the two components of the Gaussian mixture clearly distinct, and with the joint model assigning higher mass to the true component each subject belongs to, that is, the second component for subject 8 and the first component for subject 46. The posterior probability of being in component 1 was 0.3057 for subject 8, and increased to 0.6025 for subject 46. Analogous results are obtained with the mCG approach, with posterior

probabilities of being in component 1 being equal to 0.2938 and 0.5592 for subjects 8 and 46, respectively.

[Figure 2 about here.]

The analysis was repeated under different choices of the hyperparameter values and initial number of factors for the LFRM. The results were robust, with no noticeable differences in the conclusions.

5. Application to the Healthy Pregnancy, Healthy Baby Study

In this section, we return to the HPHB study that was briefly described in Section 1. Women with high blood pressure are more likely to have certain complications during pregnancy than normotensive women. In particular, hypertension during pregnancy is associated with low birth weight and early delivery, and in the most serious cases the mother develops preeclampsia which can threaten the lives of both the mother and the fetus. Typically, blood pressure declines steadily until mid-gestation and then rises until delivery in normotensive women, whereas women who develop preeclampsia tend to have higher blood pressure levels at early pregnancy, usually remaining constant until mid-gestation, then the levels rise until delivery. Preeclampsia rarely develops before the $20th$ week, and typically shows up after 37 weeks of gestation and more generally at any time in the second half of the pregnancy, including labor and even after delivery, usually within 48 hours. Monitoring the blood pressure trajectory during pregnancy can help identify women at risk of adverse birth outcomes, and point to appropriate treatments.

Data were available for 1,027 English-literate women at least 18 years old, for a total of 10,290 measurements. Women with twin gestation or with known congenital anomalies were not included in our analysis. Women with pre-gestational chronic hypertension were also excluded from the analysis since their blood pressure was artificially lowered by medical

treatment. Moreover, our analysis only included non-Hispanic black and non-Hispanic white women due to the limited number of Hispanics and other ethnic groups in the study.

It is of interest to examine the differences in blood pressure trajectories and analyze the effect of covariates on average MAP curves for specific groups. Let y_{ij} denote the jth measurement of MAP for woman i occurring t_{ij} days after the estimated day of ovulation; we assume $y_{ij} \sim \mathrm{N}(f_i(t_{ij}), \psi^2)$, where f_i is a smooth trajectory in MAP for woman i and ψ^2 is the measurement error variance. The data contain between $n_i = 1$ and $n_i = 25$ observations per woman, with an average of $\bar{n} = 10$.

The sampler described in the Appendix was run for 25,000 iterations, with the first 5,000 samples discarded as a burn-in and collecting every fifth sample to thin the chain. The update took 71 seconds per hundred iterations for the LFRM versus 64 seconds for the mCG approach in Matlab on an Intel(R) Core(TM)2 Duo machine. Traceplots of the subject-specific basis coefficients, the factor loadings and the latent factors showed slow mixing in the MCMC implementation, with high autocorrelation. However, because traceplots of the functions estimates $f_i(t_{ij})$ at a variety of time locations j and for a variety of subjects i were well behaved, the mixing problem above did not appear to impact our inferences. The estimated number of factors was 11, with a 95% credible interval of [9, 13].

Figure 3 shows the results for 6 randomly selected women, and we can see that the MAP estimates followed the typical U-shaped trajectory.

[Figure 3 about here.]

[Figure 4 about here.]

Repeating the analysis for the mCG approach (Figure 4), we observed accurate estimates at locations close to data points (although less smoothed than those induced by the LFRM), but the estimates were inferior when no or few measurements were recorded, with much wider 95% credible intervals and unrealistic estimates up to the $5th$ week of gestation for subjects

in the last 5 panels. Because it keeps the factor loadings matrix $\mathbf{\Lambda}$ fixed, the mCG approach does not inherit the shrinkage property induced on the basis coefficients by the multiplicative gamma shrinkage prior adopted for $\mathbf{\Lambda}$ in the LFRM. Therefore, no basis selection is induced, and the adoption of a pre-specified over-complete set of basis functions may lead to overly-spiky curves.

To assess the predictive performance, we repeated the analysis holding out the MAP measurements collected after the $30th$ week for 300 randomly selected women who had at least one measurement in the first 30 weeks. In general, the mCG approach lead to worse predictive accuracy than the LFRM. For example, the mean square predictive error was 90.04 under the LFRM and 100.21 under the mCG approach.

It is also of interest to study the impact of predictors on the distribution of the curves. In particular, we considered 12 predictors: pre-gestational diabetes, renal disease, insurance status, sex of the infant, maternal race, maternal education, serum cotinine in ng/mL, serum cadmium in ng/mL, serum lead in ug/dL, parity, pre-gestational body mass index (BMI) and weight gain. Figure 5 shows how average MAP trajectories change across six different covariate groups. Our findings are confirmed by the literature results on blood pressure during pregnancy, with older and primiparous women having higher blood pressure, although discrepancies are small. Neither the maternal race nor the sex of the infant seem to affect the gestational blood pressure, as shown in panels (2,1) and (2,2). No differences are found in the gestational blood pressure of women with high lead levels with respect to their counterpart (panel (3,2)). Finally, women with diabetes have higher gestational blood pressure than healthy women, with non-overlapping 95% credible intervals after mid-gestation until the $35th$ week. Women with diabetes include women on treatment (either oral or insulin) or on diet. Our findings on the impact of predictors on the distribution of the curves were confirmed by the results obtained with the mCG approach.

[Figure 5 about here.]

The estimated posterior means of the two Gaussian components in the joint model of Section 3.3 were $(36.297, 2.888)$ and $(39.375, 3.192)$, respectively for the first and second bivariate component, with 95% credible intervals equal to $([35.660, 36.875], [2.693, 3.060])$ for the first component and $([39.268, 39.480], [3.152, 3.234])$ for the second component, respectively.

[Table 1 about here.]

Table 1 reports the posterior mean estimates of the marginal probabilities of preeclampsia and LBW (with Monte Carlo standard errors) computed at the $20th$, $25th$, $30th$ and $35th$ week of gestation for four women in the test set. The standard errors were set to be the ratio between the standard deviation of the marginal probabilities across MCMC draws and the square root of the number of MCMC samples used to compute the average marginal probabilities. z_p^i and z_{lbw}^i are indicator variables equal to 1 if woman i developed preeclampsia and delivered a LBW infant, respectively. All women in the test set had at least one MAP measurement collected before the $20th$ week, and at least one measurement collected after the $35th$ week. Woman 1 was preeclamptic and delivered a LBW infant; woman 2 was preeclamptic but did not deliver a LBW infant; woman 3 was not preeclamptic but delivered a LBW infant; woman 4 was neither preeclamptic nor delivered a LBW baby. Even as early as 20 weeks of gestation the LFRM detects the risk of preeclampsia and LBW for woman 1 and the risk of LBW for woman 3, with estimated probabilities over three times higher than the 8.2% national rate for LBW and the typical 5-8% incidence range of preeclampsia over all pregnancies. As for subject 2, the probability of LBW remains very low throughout the entire course of the pregnancy. However, the probability of preeclampsia increases from the $25th$ week to the $30th$ week of gestation, but then decreases to 11.41% at the $35th$ week. One potential explanation can be seen in Figure 6, which shows the MAP trajectories for the 4 women at the $35th$ week. In particular, the trajectory and the MAP measurements

for woman 2 are similar to those of normotensive woman 4. Consequently, it is possible that woman 2 had normal blood pressure during the prenatal visits, but is still preeclamptic because she had very high blood pressure (and urine proteins) at time of delivery.

[Figure 6 about here.]

Therefore, even as early as 20 weeks of gestation, when typically very few MAP measurements have been collected per woman, the LFRM is able to identify women at high risk for adverse birth outcomes likely to happen later in pregnancy. Predictions get more accurate around the $30th$ to $35th$ week of gestation, although in some cases the information on the maternal blood pressure carried by latent factors may not be enough to detect the risk of adverse outcomes, as for woman 2, and additional information on the health of the woman should be incorporated.

6. Discussion

The article has proposed a Bayesian latent factor regression model for functional data. The basic formulation generalizes the sparse Bayesian infinite factor model of Bhattacharya and Dunson (2011), which was developed for estimation of high-dimensional covariance matrices for vector data, to the functional data case. This allows one to include a high-dimensional set of pre-specified basis functions, while allowing automatic shrinkage and effective removal of basis coefficients that are not needed to characterize any of the curves under study. In addition, we consider several additional generalizations allowing predictors to impact the latent factor scores and accommodating joint modeling of functional predictors with scalar responses that are modeled parametrically or via mixture models. Along the same lines, we can consider joint modeling of multiple related functions easily within the proposed framework, but our emphasis was on developing methods motivated by the application to the study of blood pressure and pregnancy outcomes.

The proposed framework has the advantage of straightforward computation via a simple Gibbs sampler and easy modifications for joint modeling of disparate data of many different types. In particular, the $\boldsymbol{\theta}_i$ vector of basis coefficients in the functional data model can instead be replaced with concatenated coefficients within component models for different types of objects, including not only time trajectories but also images, movies, text, vectors, etc. This leads to a general shared latent factor framework for modeling high-dimensional mixed domain data that should have broad utility to be explored in future research. An interesting modification would be a semiparametric case that allows the latent variables densities to be unknown via nonparametric Bayes priors.

Acknowledgements

The authors thank Marie Lynn Miranda for access to the data and for helpful discussions. This work was partially funded by Award Number R01ES17240 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business & Economic Statistics* **18,** 338–357.

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

Arminguer, G. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika* **63,** 271–300.

Behseta, S., Kass, R. E. and Wallstrom, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika* **92,** 419–434.

Besse, P. and Ramsey, J.O (1986). Principal component analysis of sampled functions. *Psychometrika* **51,** 285–311.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98,** 291–306.

Bosq, D. (2000). *Linear Processes in Function Spaces. Theory and Applications. Lecture Notes in Statististics.* **149,** Springer, New York.

Cardot, H. (2006). Conditional functional principal components analysis. *Scandinavian journal of statistics* **34,** 317–335.

Carvalho, C. et al. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103,** 1438–1456.

Chiou, J.-M., Müller, H.-G. and Wang, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B* **65,** 405–423.

Crainiceanu C. and Goldsmith J. (2010). Bayesian functional data analysis using WinBUGS. *Journal of Statistical Software* **32,** 1–33.

Cunningham, F. G., Gant, N. F., Leveno, K. J., Gilstrap, L. C., Hauth, J. C., and Wenstrom, K. D. (2001). *Hypertensive disorders in pregnancy.* In: *Williams Obstetrics*, 567–618. McGraw-Hill, New York, 21st edition.

Geweke, J. F. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies* **9,** 557–587.

Hamilton, B. E., Martin, J. A. and Ventura, S. J. (2010) Births: preliminary data for 2008. *National Vital Statistic Reports* **58,** 1–17.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal components models for sparse

functional data. *Biometrika* **87,** 587–602.

James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98,** 397–408.

Jiang, C.-R. and Wang, J.-L. (2010). Covariate adjusted functional principal components analysis for longitudinal data. *Annals of Statistics* **38,** 1194–1226.

Jones, B.L., Nagin, D.S. and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods and Research* **29** 374–393.

Lin H. et al. (2000). A latent class mixed model for analyzing biormarker trajectories in longitudinal data with irregularly scheduled observations. *Statistics in Medicine* **19,** 1303–1318.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14,** 41–67.

Marin, J.M., Mengenrsen, K and Casella, R. (2005). *Handbook of Statistics 25: Bayesian thinking, modeling and computation.* Elsevier: New York.

McLachlan, G. and Peel, D. (2000). *Finite mixture models.* Wiley-Interscience: New York.

Muthén, B. O., and Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and experimental research* **24,** 882–891.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric group-based approach. *Psychological Methods* **4,** 139–157.

O'Brien, S.M. and Dunson, D. B. (2004) Bayesian Multivariate Logistic Regression. *Biometrics* **60,** 739–746.

Ramsey, J.O. and Silverman, B.W. (2005). *Functional Data Analysis,* 2nd edition. New York: Springer - Verlag.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* **14,** 631–647.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B* **53,** 233–243.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57,** 253–259.

Spiegelhalter, D. J. (1998). Bayesian graphical modelling: a case study in monitoring health outcomes. *Applied Statistics* **47,** 115–133.

Staniswalis, J.G., and Lee, J.J., (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **93,** 1403–1418.

Yang, C.-C., Muthén, B. O. and Yang, C.-C. (1999). Finite mixture multivariate generalized linear models using Gibbs sampling and E-M algorithms. *Proceedings of the National Science Council ROC(A)* **23,** 695–702.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100,** 577–590.

Received June 2011. *Revised XXXXXX* 2011.

Accepted XXXXX 20XX.

Appendix

MCMC algorithm for the latent factor regression model

This section contains a description of the MCMC algorithm used to update from the posterior distributions of the parameters based on the priors given in Section 2. The sampler cycles through the following steps:

- Step 1 Update of $\mathbf{\Lambda}$: sample $\lambda_{jh}, \delta_1, \delta_h, \phi_{jh}$ from the following posteriors:

(1) Denote the jth row of $\mathbf{\Lambda}_{k^*}$ (the loading matrix $\mathbf{\Lambda}$ truncated to $k^* << p$) by $\boldsymbol{\lambda}_j$; then the $\boldsymbol{\lambda}_j$'s have independent conditionally conjugate posteriors given by

$$\pi(\boldsymbol{\lambda}_j|-) \sim N_{k^*}((\mathbf{D}_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}'\boldsymbol{\eta})^{-1}\boldsymbol{\eta}'\sigma_j^{-2}\boldsymbol{\theta}^{(j)}, (\mathbf{D}_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}'\boldsymbol{\eta})^{-1})$$

with $\mathbf{D}_j^{-1} = \text{diag}(\phi_{j1}\tau_1, \ldots, \phi_{jk}\tau_{k^*})$, $\boldsymbol{\eta}' = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{k^*}]$ and $\boldsymbol{\theta}^{(j)} = (\theta_{j1}, \ldots, \theta_{jn})$, for $j = 1, \ldots, p$.

(2) Sample ϕ_{jh} from

$$\pi(\phi_{jh}|-) \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{\tau_h\lambda_{jh}^2}{2}\right)$$

(3) Sample δ_1 from

$$\pi(\delta_1|-) \sim \text{Gamma}\left(a_1 + \frac{pk^*}{2}, 1 + \frac{1}{2}\sum_{l=h}^{k^*}\tau_l^{(1)}\sum_{j=1}^{p}\phi_{jl}\lambda_{jl}^2\right)$$

(4) Sample δ_h from

$$\pi(\delta_h|-) \sim \text{Gamma}\left(a_2 + \frac{p^*}{2}(k-h+1), 1 + \frac{1}{2}\sum_{l=1}^{k^*}\tau_l^{(h)}\sum_{j=1}^{p}\phi_{jl}\lambda_{jl}^2\right)$$

for $h \geqslant 2$, where $\tau_l^{(h)} = \prod_{t=1, t\neq h}^{l}\delta_t$ for $h = 1, \ldots, p$.

The sampling begins with a very conservative choice of k^* which is then automatically selected within the adaptive Gibbs sampler as described in Bhattacharya and Dunson (2011).

- Step 2 Update of $\sigma_j^2, j = 1, \ldots, p$: denoting as σ_j^{-2} the diagonal elements of $\mathbf{\Sigma}^{-1}$, sample σ_j^{-2} from conditionally independent posteriors

$$\pi(\sigma_j^{-2}|-) \sim \text{Gamma}\left(\frac{n}{2} + a_\sigma, b_\sigma + \frac{\sum_{i=1}^{n}(\boldsymbol{\theta}_i - \mathbf{\Lambda}\boldsymbol{\eta}_i)^2}{2}\right)$$

- Step 3 Update of ψ^{-2}: sample ψ^{-2} from

$$\pi(\psi^{-2}|-) \sim \text{Gamma}\left(\frac{N}{2} + a_\psi, b_\psi + \frac{\sum_{j=1}^{N}(y_j - \mathbf{\Theta}_j)^2}{2}\right)$$

where N denotes the total number of observations, $\mathbf{y}$ is a column vector which stacks the measurements for all women, $\mathbf{y} = (y_{1,t_{1,1}}, \ldots, y_{n,t_{n,n_n}})'$, and $\mathbf{\Theta}$ is a $N \times 1$ column vector which stacks the scores for all subjects, $\mathbf{\Theta} = \{\mathbf{B}_i\boldsymbol{\theta}_i, \ldots, \mathbf{B}_n\boldsymbol{\theta}_n\}'$, where each $\mathbf{B}_i\boldsymbol{\theta}_i$ has dimension $n_i \times 1$ with n_i the number of measurements for subject i.

- Step 4: Update of $\boldsymbol{\beta}$ and ω elements:

4-a) Given the prior $\omega_{lj} \sim \text{Gamma}(1/2, 1/2)$, $l = 1, \ldots, r$ and $j = 1, \ldots, k$, sample ω_{lj} from the full conditional posterior

$$\pi(\omega_{lj}|-) \sim \text{Gamma}\left(1, \frac{1}{2}\left(1 + \beta_{lj}^2\right)\right)$$

4-b) Sample the jth column of the matrix of coefficients $\boldsymbol{\beta}$ from the full conditional posterior

$$\pi(\boldsymbol{\beta}_j|-) \sim N\left(\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}' + \mathbf{E}^{-1}\right)^{-1}\tilde{\mathbf{X}}\boldsymbol{\eta}'_{\cdot j}, \left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}' + \mathbf{E}^{-1}\right)^{-1}\right)$$

with matrix $\mathbf{E}$ corresponding to $\mathbf{E} = \text{Diag}(\omega_{lj}^{-1})$, $l = 1, \ldots, r$ and $j = 1, \ldots, k$.

- Step 5 Update of $\boldsymbol{\eta}_i$: marginalizing out $\boldsymbol{\theta}_i$, the model can be rewritten as

$$\begin{aligned}\mathbf{y}_i &= \mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{B}_i\boldsymbol{\epsilon}_i + \boldsymbol{\epsilon}_i^*, \quad \boldsymbol{\epsilon}_i^* \sim N(0, \psi^2\mathbf{I}_{n_i}), \boldsymbol{\epsilon}_i \sim N_p(0, \boldsymbol{\Sigma}) \\ &= \mathbf{B}_i\boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\alpha}_i^*, \quad \boldsymbol{\alpha}_i^* \sim N(0, \psi^2\boldsymbol{I}_{n_i} + \mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i')\end{aligned}$$

Thus, sample $\boldsymbol{\eta}_i$ from the full conditional posterior

$$\begin{aligned}\pi(\boldsymbol{\eta}_i|-) &\sim N(\mathbf{A}^{-1} \times \mathbf{C}, \mathbf{A}^{-1}) \\ \mathbf{A} &= \boldsymbol{\Lambda}'\mathbf{B}_i'(\psi^2\boldsymbol{I}_{n_i} + \mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i')^{-1}\mathbf{B}_i\boldsymbol{\Lambda} + \boldsymbol{I}_k \\ \mathbf{B} &= \boldsymbol{\beta}'\boldsymbol{x}_i + \boldsymbol{\Lambda}'\mathbf{B}_i'(\psi^2\mathrm{I}_{n_i} + \mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i')^{-1}\mathbf{y}_i\end{aligned}$$

- Step 6 Update of $\boldsymbol{\theta}_i$: sample $\boldsymbol{\theta}_i$ from conditionally independent posteriors

$$\begin{aligned}\pi(\boldsymbol{\theta}_i|-) \sim N_p((\psi^{-2}\mathbf{B}_i'\mathbf{B}_i + \boldsymbol{\Sigma}^{-1})^{-1}(\psi^{-2}\mathbf{B}_i'\mathbf{y}_i + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\eta}_i), \\ (\psi^{-2}\mathbf{B}_i'\mathbf{B}_i + \boldsymbol{\Sigma}^{-1})^{-1})\end{aligned}$$

1 Data and function estimates for 3 subjects in the simulation example under the LFRM (left panels) and the modified Crainiceanu-Goldsmith method (right panels). The true functions are represented with black lines, the posterior means with red lines, and the dotted lines are 95% pointwise credible intervals.

2 LFRM-estimated joint distribution of gestational age (weeks) and birth weight (Kg) and contour plot for subjects 8 and 46 in the simulation example.

3 MAP function estimates for selected women in the Healthy Pregnancy, Healthy Baby Study. The posterior means are solid lines and dotted lines are 95% pointwise credible intervals. The x-axis scale is time in weeks starting at the estimated day of ovulation.

4 MAP function estimates for selected women under the modified Crainiceanu-Goldsmith approach. The posterior means are solid lines and dotted lines are 95% pointwise credible intervals. The x-axis scale is time in weeks starting at the estimated day of ovulation.

5 MAP function estimates for 6 representative covariate groups.

6 MAP function estimates at the $35th$ week for subjects 1-4 in the test set.

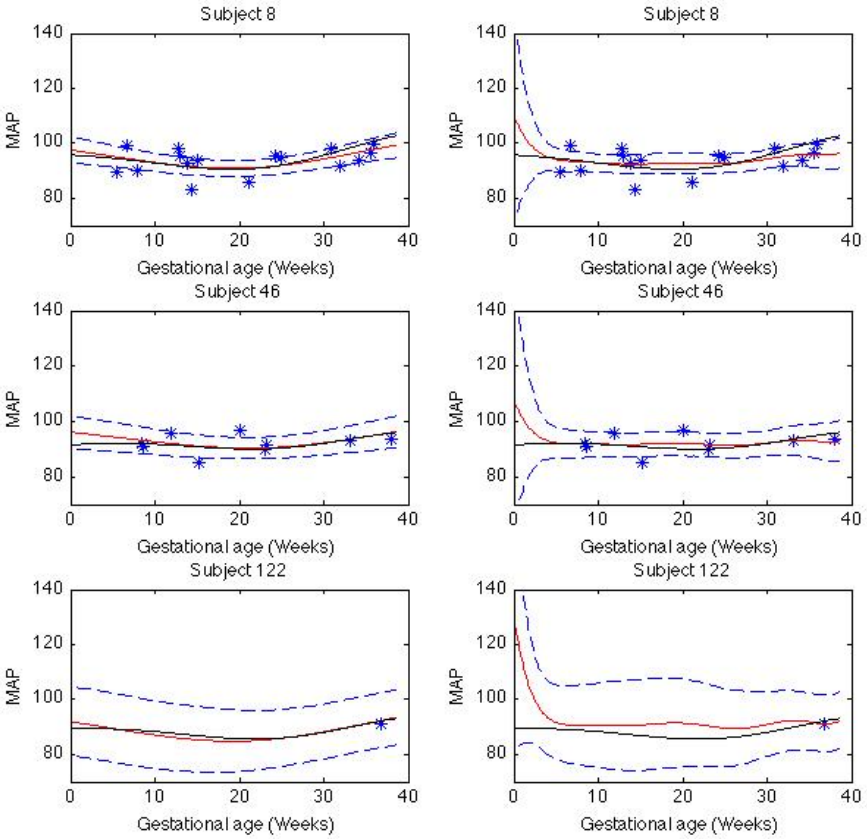

Figure 1: Data and function estimates for 3 subjects in the simulation example under the LFRM (left panels) and the modified Crainiceanu-Goldsmith method (right panels). The true functions are represented with black lines, the posterior means with red lines, and the dotted lines are 95% pointwise credible intervals.

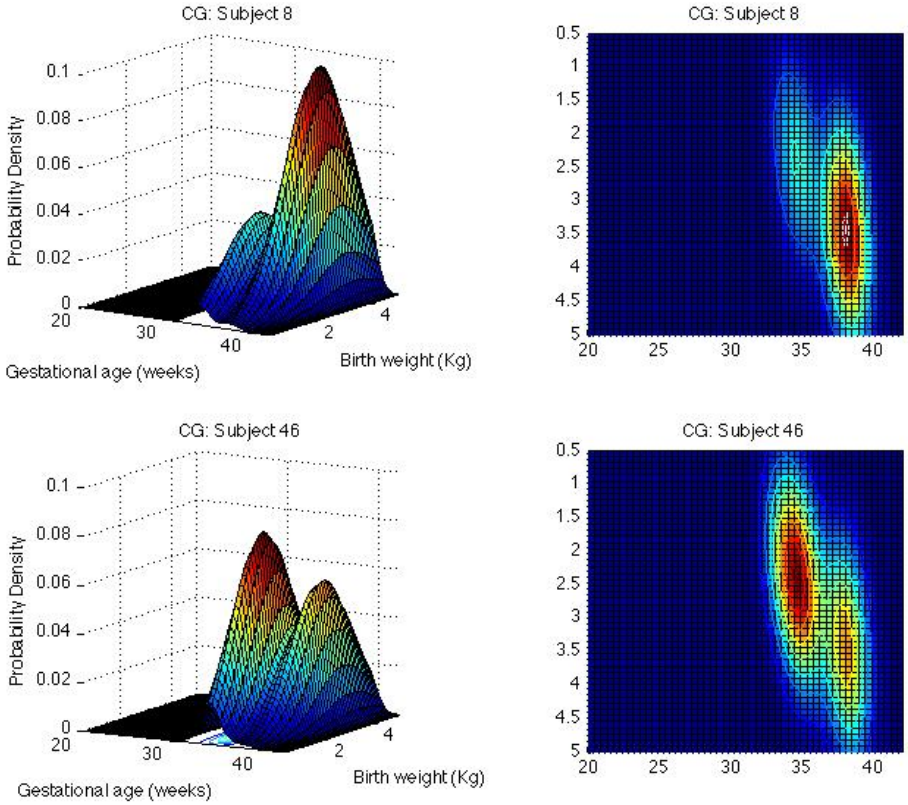

Figure 2: LFRM-estimated joint distribution of gestational age (weeks) and birth weight (Kg) and contour plot for subjects 8 and 46 in the simulation example.

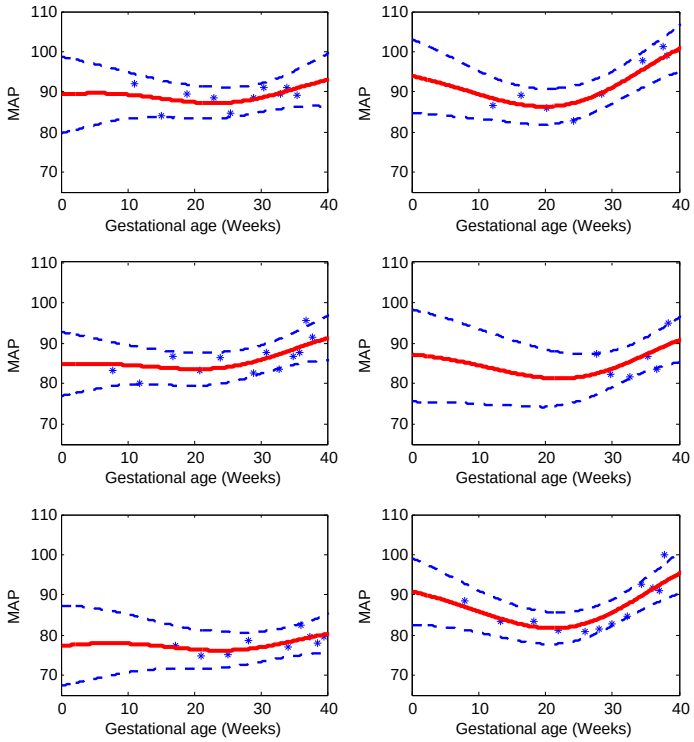

Figure 3: MAP function estimates for selected women in the Healthy Pregnancy, Healthy Baby Study. The posterior means are solid lines and dotted lines are 95% pointwise credible intervals. The x-axis scale is time in weeks starting at the estimated day of ovulation.

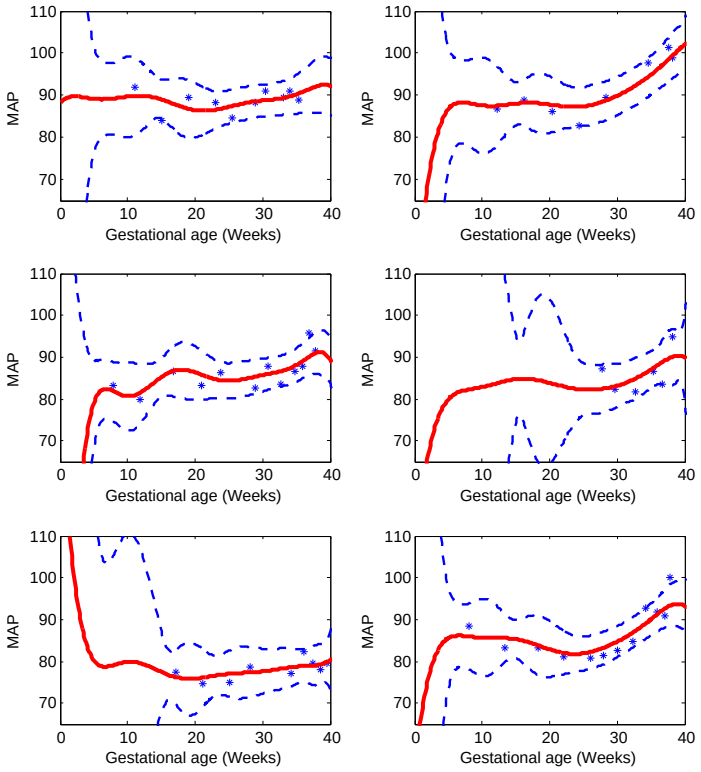

Figure 4: MAP function estimates for selected women under the modified Crainiceanu-Goldsmith approach. The posterior means are solid lines and dotted lines are 95% pointwise credible intervals. The x-axis scale is time in weeks starting at the estimated day of ovulation.

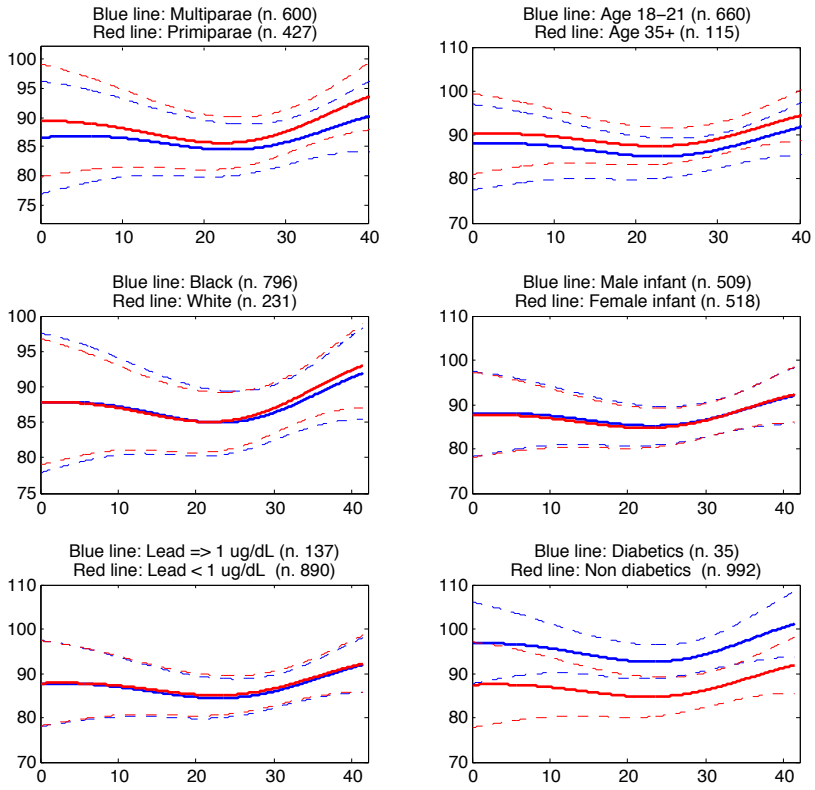

Figure 5: MAP function estimates for 6 representative covariate groups. The posterior means are solid lines and dotted lines represent 95% pointwise credible intervals.

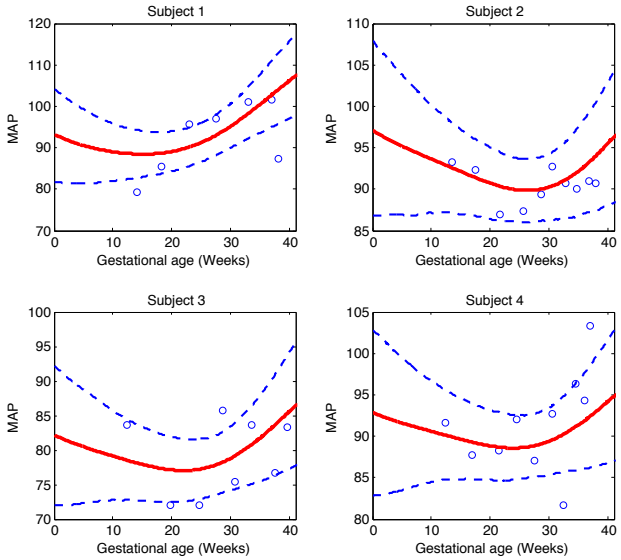

Figure 6: MAP function estimates at the 35*th* week for subjects 1-4 in the test set.

Table 1

Posterior mean estimates of the probabilities of preeclampsia and LBW (with Monte Carlo standard errors)

	Subjects			
$\Pr(z_p^i = 1)$	1	2	3	4
$20th$ week	0.2545 (0.0037)	0.2085 (0.0034)	0.0711 (0.0019)	0.1179 (0.0025)
$25th$ week	0.2819 (0.0047)	0.1314 (0.0031)	0.1148 (0.0038)	0.1046 (0.0027)
$30th$ week	0.3640 (0.0044)	0.1960 (0.0035)	0.0855 (0.0023)	0.0985 (0.0023)
$35th$ week	0.4185 (0.0042)	0.1141 (0.0023)	0.1128 (0.0024)	0.0983 (0.0021)
$\Pr(z_{lbw}^i = 1)$	1	2	3	4
$20th$ week	0.2582 (0.0054)	0.0858 (0.0032)	0.2544 (0.0053)	0.1144 (0.0037)
$25th$ week	0.2391 (0.0056)	0.0644 (0.0030)	0.3166 (0.0062)	0.0981 (0.0038)
$30th$ week	0.3193 (0.0058)	0.0986 (0.0035)	0.2865 (0.0057)	0.1056 (0.0036)
$35th$ week	0.3462 (0.0058)	0.0608 (0.0027)	0.3462 (0.0058)	0.0997 (0.0034)