Statistica Sinica **19** (2009), 611-629

NONPARAMETRIC BAYES KERNEL-BASED PRIORS FOR FUNCTIONAL DATA ANALYSIS

Richard F. MacLehose and David B. Dunson

University of Minnesota and Duke University

Abstract: We focus on developing nonparametric Bayes methods for collections of dependent random functions, allowing individual curves to vary flexibly while adaptively borrowing information. A prior is proposed, which is expressed as a hierarchical mixture of weighted kernels placed at unknown locations. The induced prior for any individual function is shown to fall within a reproducing kernel Hilbert space. We allow flexible borrowing of information through the use of a hierarchical Dirichlet process prior for the random locations, along with a functional Dirichlet process for the weights. Theoretical properties are considered and an efficient MCMC algorithm is developed, relying on stick-breaking truncations. The methods are illustrated using simulation examples and an application to reproductive hormone data.

Key words and phrases: Dirichlet process, functional data analysis, kernel smoothing, mixture model, random curve, RKHS.

1. Introduction

In functional data analysis (FDA), interest focuses on studying random curves for different subjects. There has been rapidly increasing interest in FDA in the statistical literature, with much of this literature focusing on developing more flexible methods for longitudinal data (Zhao, Marron and Wells (2004), Müller (2005) and Morris and Carroll (2006), among others). In this article, the focus is on nonparametric Bayesian methods, defining random probability measures for collections of dependent functions. Our goal is to treat the individual curves nonparametrically, while also borrowing information across subjects flexibly.

In contrast, much of the Bayesian literature on FDA treats the mean curve nonparametrically, but requires parametric assumptions on the distribution about the mean. For example, Bigelow and Dunson (2007) model the basis coefficients in a multivariate adaptive spline model as normally distributed. A related approach was independently developed by Thompson and Rosen (2006) in the setting of a univariate spline model, with variable selection used to select the basis functions. Morris and Carroll (2006) propose a wavelet-based functional mixed model, placing a normal distribution of the random wavelet coefficients.

Behseta, Kass and Wallstrom (2005) avoid choosing an explicit set of basis functions through use of a hierarchical Gaussian process.

In order to allow a distribution function to be unknown, a common approach is to use a Dirichlet process (DP) prior (Ferguson (1973, 1974)) or DP mixture (DPM) (Berry and Christensen (1979), Lo (1984), Escobar (1994) and Escobar and West (1995)). To model a collection of random functions, one possible strategy is to define a DP with support on a function space. Using a closely-related formulation to the dependent Dirichlet process (DDP) of MacEachern (1999, 2000), De Iorio, Müller, Rosner and MacEachern (2004) and Gelfand, Kottas and MacEachern (2005) proposed a functional DP (FDP) for spatial data. The FDP approach relies on replacing the atoms in the Sethuraman (1994) representation of the DP with random functions drawn from a Gaussian process (GP). Hence, the functional distribution is formulated as a mixture across infinitely-many GP realizations. A related FDP approach has been to approximate the function through a basis series expansion with a DP prior on the basis coefficients. Ray and Mallick (2006) placed a DP prior on the distribution of the coefficients in a hierarchical wavelet model to induce functional clustering. Bigelow and Dunson (2006) instead allow both the basis functions and their distribution to be unknown. Although these FDP models can be effective, they focus on global clustering of curves. Curves that are clustered together are assumed to take the same values along their range and no possibility exists for local deviations. The approach we propose allows curves to be clustered globally but to deviate from one another locally.

In the frequentist literature, kernel-based methods have been widely used for function estimation, due largely to practical advantages in high-dimensional problems (Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004)). There has also been some recent focus on kernel methods in the Bayes literature on estimation of a single function (Tipping (2001), Sollich (2002) and Chakraborty, Ghosh and Mallick (2005)). However, to our knowledge, kernel methods have not been used for Bayesian functional data analysis. Our proposed approach builds on recent work by Liang, Liao, Mukherjee and West (2006) and Pillai, Liang, Mukerjee, Wolpert and Wu (2006), who considered formal Bayes kernel methods for posterior inference on a single curve.

In Section 2, we first provide background on the functional Dirichlet process and recent work on Bayes kernel methods, proposing modifications to allow kernel selection in the single function case. Section 3 generalizes these methods to the hierarchical case. Section 4 develops methods for posterior computation.

Section 5 considers simulation examples. Section 6 presents an application to an epidemiologic data set, and Section 7 discusses the results.

2. Priors for Random Curves

2.1. Functional Dirichlet process

We propose a nonparametric prior for dependent random functions. Let the random function for subject i be denoted η_i, with $\{\eta_i(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ a stochastic process over $\mathcal{X}$, for $i = 1, \ldots, n$. Formally, η_i is a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is a function space, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $\mathcal{P}$ is a probability measure over $(\Omega, \mathcal{F})$. Following a Bayesian approach, we treat $\mathcal{P}$ as random to assign a prior over the collection of dependent random functions $\{\eta_i, i = 1, \ldots, n\}$.

One possibility for $\mathcal{P}$ is to use a DP prior (Ferguson (1973, 1974)), with a base measure chosen to have support on a function space. Specifically, using the Sethuraman (1994) constructive stick-breaking representation of the DP, one can let

$$\eta_i \sim G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}, \quad \theta_h \overset{i.i.d.}{\sim} GP(\mu, \mathcal{C}), \tag{2.1}$$

where $p_h = V_h \prod_{l<h}(1 - V_l)$, $V_h \overset{i.i.d.}{\sim}$ beta$(1, \alpha)$, $h = 1, \ldots, \infty$, and the functional atoms, $\{\theta_h\}$, are drawn independently from a Gaussian process (GP) centered on μ with covariance function $\mathcal{C}$. Note that (2.1) generates a random probability measure with support on a function space by assigning random weights to functional atoms generated from a Gaussian process. Due to the almost sure discreteness of the DP, there will be a positive probability that $\eta_i = \eta_{i'}$, so that the approach allows clustering of curves.

It is worth commenting on some properties of (2.1). The prior borrows information between η_i and $\eta_{i'}$ by allowing global clustering of functions allocated to the same functional atom and by generating the functional atoms from the same GP. Borrowing information through global clustering can be quite restrictive in that two functions, η_i and $\eta_{i'}$, may be very similar or even identical in certain subregions of $\mathcal{X}$ without being identical everywhere. The dependence structure between different functional atoms drawn from the same GP is also restrictive, being driven by the covariance function $\mathcal{C}$, which is typically parameterized by two or three unknowns. In addition, computation becomes increasingly difficult as the number of observations along the curve, accrued across all subjects in the sample, increases. The computational burden is driven by the need to calculate inverses of large matrices. Motivated by this problem in the setting of spatial data, Xia and Gelfand (2005) proposed a kernel-based method for fast approximate computation in Gaussian process models.

2.2. Kernel methods

Kernel-based methods have been widely used to estimate functions in the frequentist literature, due largely to practical advantages in high-dimensional problems (Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004)). From computational, theoretical and applied perspectives, it is often appealing to estimate the function η subject to the constraint that $\eta \in \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). As noted by Wahba (1990), functions generated from a GP are almost surely outside an RKHS. A number of authors (Tipping (2001), Sollich (2002) and Chakraborty et al. (2005)) have considered Bayesian kernel-based methods using the additive model (Hastie, Tibshirani and Friedman (2001)):

$$\eta(\mathbf{x}) = \sum_{i=1}^{n} w_i \, K(\mathbf{x}, \mathbf{x}_i), \tag{2.2}$$

where η is the function of interest (dropping the i subscript in focusing on estimation of a single function), $K : \mathcal{X} \times \mathcal{X} \to \Re^{+}$ is a uniformly bounded Mercer kernel, and $\mathbf{w} = (w_1, \ldots, w_n)'$ is a vector of unknown coefficients.

Expression (2.2) is motivated by the representer theorem of Kimeldorf and Wahba (1971), which states that the solution to the problem of minimizing a goodness-of-fit loss function subject to an RKHS norm penalty lies in a subspace of $\mathcal{H}$, represented as in (2.2). This allows one to use (2.2) to obtain an estimator of η, which can be interpreted in a Bayesian manner as a maximum a posteriori (MAP) estimator (Wahba (1999) and Poggio and Girosi (1990)). However, as noted by Liang et al. (2006), a Bayesian would typically be interested not just in a single function estimator, such as the MAP, but more generally in the posterior of η. The full posterior is very useful in assessing uncertainty in estimation, in performing inferences about features of the function, and in making predictions.

Pillai et al. (2006) and Liang et al. (2006) note problems in obtaining the posterior directly through use of the finite representation (2.2). One issue is that the prior would need to be defined in a sample-dependent manner, as the sample size n is a component of (2.2). Another important issue is that (2.2) was derived in solving an optimization problem. Using this representation for full posterior inference causes not just the MAP estimator but all samples from the posterior to lie in a subspace of $\mathcal{H}$ characterized by (2.2). Unless one has *a priori* reason to believe that this subspace is rich enough to characterize all uncertainty in the function, it seems more appealing to define a prior with large support in $\mathcal{H}$.

Pillai et al. (2006) instead induce a prior on a function space, $\mathcal{G}$, through a prior on a space of signed Borel measures, Γ, by using the integral operator $\mathcal{L}_K : \Gamma \rightarrow G$ defined as

$$\mathcal{L}_K[\gamma](\mathbf{x}) := \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{u}) d\gamma(\mathbf{u}) = \eta(\mathbf{x}), \tag{2.3}$$

where $\gamma \in \Gamma$ and $\eta \in \mathcal{G}$. When $\Gamma = \mathcal{B}(\mathcal{X})$, the space of all signed Borel measures, then $\mathcal{G} = \mathcal{H}_K$, with $\mathcal{H}_K$ the RKHS associated with kernel K. Pillai et al. (2006) focus on Lévy process priors, that form a general class including Brownian motion, Poisson processes, and Dirichlet processes as special cases. Liang et al. (2006) instead apply the decomposition $d\gamma(\mathbf{u}) = \phi(\mathbf{u})\, d\pi(\mathbf{u})$, with the coefficient function ϕ assigned a GP prior and the probability measure π assigned a DP prior on $\mathcal{X}$.

3. Priors for Dependent Random Curves

3.1. Hierarchical kernel priors

In order to generalize (2.3) to define a prior for dependent random functions, we let

$$\eta_i(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{u})\, d\gamma_i(\mathbf{u}), \quad i = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_n\}$ is a collection of dependent random signed measures, with $\gamma_i \in \Gamma \subset \mathcal{B}(\mathcal{X})$, for $i = 1, \ldots, n$. In particular, we focus on the case in which $\Gamma = \mathcal{M}$, with $\mathcal{M}$ the space of finite discrete measures expressed as

$$\begin{aligned} \mathcal{M} = \Big\{ \mu = \sum_h \phi_h \pi_h \delta_{\tau_h} : \{\phi_h\} \subset \Re, \{\pi_h\} \subset (0,1), \{\tau_h\} \subset \mathcal{X}, \\ \sum_h \pi_h = 1, \sum_h |\phi_h| \pi_h < \infty \Big\}. \end{aligned} \tag{3.2}$$

In this case, $\eta_i \in \mathcal{G} = \mathcal{L}_K[\mathcal{M}]$, the range of the integral operator $\mathcal{L}_K$ over the space $\mathcal{M}$. From Pillai et al. (2006) , for every $\mu \in \mathcal{M}$, $\mathcal{L}_K[\mu] \in \mathcal{H}_K$ and $\mathcal{L}_K(\mathcal{M})$ is dense in $\mathcal{H}_K$ with respect to the RKHS norm.

By choosing a prior for γ_i with support in $\mathcal{M}$, we induce a prior on η_i with support in $\mathcal{H}_K$. We modify the Liang et al. (2006) specification to favor a sparse representation through the use of kernel function selection. Bayesian stochastic search variable selection (SSVS) methods (George and McCulloch (1993)) have been widely used for basis function selection (Smith and Kohn (1996)). Using the decomposition $d\gamma_i(\mathbf{u}) = \phi_{0i}(\mathbf{u})\,\{1 - d\pi_i(\mathbf{u})\} + \phi_{1i}(\mathbf{u})\, d\pi_i(\mathbf{u})$, we propose the hierarchical specification

$$\pi_i \sim F, \qquad \phi_{0i} \sim G_0 \quad \text{and} \quad \phi_{1i} \sim G_1, \tag{3.3}$$

where $\pi_i \in \Pi$ is a random probability measure on $(\mathcal{X}, \mathcal{S})$, with Π the space of probability measures on $(\mathcal{X}, \mathcal{S})$, F is a random probability measure on $(\Pi, \mathcal{T})$, $\phi_{ji} : \mathcal{X} \to \Re$, $j = 0, 1$, are random functions, and G_j, $j = 0, 1$, are random probability measures on $(\Psi, \mathcal{U})$, with Ψ a function space. Here, $\mathcal{S}, \mathcal{T}, \mathcal{U}$ correspond to σ-algebras for each of the above mentioned spaces. The function ϕ_{0i} is assigned a GP prior (which we refer to as the GP spike) whose covariance function, $\mathcal{C}_0$, allows little deviation from the function's prior mean. The function ϕ_{1i} is also assigned a GP prior (which we refer to as the GP slab) but with covariance function $\mathcal{C}_1$ allowing large deviations from the prior mean.

Note that we now have a distribution of random probability measures, F, and a distribution of random functions, G_0 and G_1. To facilitate functional clustering, we allow F, G_0 and G_1 to be unknown through the nonparametric prior

$$\begin{array}{rl} F \sim DP(\alpha F_0), & F_0 \sim DP(\kappa F_0^*), \\ G_0 \sim DP(\beta_0 G_0^*), & G_0^* \equiv GP(\mu, \mathcal{C}_0), \\ G_1 \sim DP(\beta_1 G_1^*), & G_1^* \equiv GP(\mu, \mathcal{C}_1), \end{array} \qquad (3.4)$$

where $\alpha, \kappa, \beta_0, \beta_1$ are DP precision parameters and F_0^* is a known probability measure on $(\mathcal{X}, \mathcal{S})$. Here, F is assigned a DP prior with the base measure further assigned a DP, so that realizations from F will correspond to random probability measures. This prior for F is equivalent to the hierarchical DP (HDP) of Teh, Jordan, Beal and Blei (2006) and similar to that proposed by Tomlinson (1998), so a more concise notation would be $F \sim HDP(\alpha, \kappa, F_0^*)$. In addition, $G_j, j = 0, 1$, are assigned DP priors with the base measure chosen to correspond to a GP, so that realizations from G_j will correspond to random functions. The prior for G_j is equivalent to the FDP described in Section 2.1. However, we do not use the FDP as a prior for the distribution of η_i directly.

3.2. Properties

Using the Sethuraman (1994) representation of the DP components in formulation (3.1)$-$(3.4), with the Teh et al. (2006) HDP generalization, we obtain the specification

$$\begin{aligned} \eta_i(\mathbf{x}) &= \sum_{h=1}^{\infty} K(\mathbf{x}, \tau_h) \{(1 - \pi_{ih})\Phi_{0,Z_{0i}}(\tau_h) + \pi_{ih}\Phi_{1,Z_{1i}}(\tau_h)\}, \\ \Pr(Z_{0i} = k) &= \nu_{0k} = \nu'_{0k} \prod_{l<k} (1 - \nu'_{0l}), \quad \nu'_{0k} \sim \text{beta}(1, \beta_0), \\ \Pr(Z_{1i} = k) &= \nu_{1k} = \nu'_{1k} \prod_{l<k} (1 - \nu'_{1l}), \quad \nu'_{1k} \sim \text{beta}(1, \beta_1), \end{aligned} \qquad (3.5)$$

$$\begin{aligned}\pi_{ih} &= \pi'_{ih}\prod_{l<h}(1-\pi'_{il}), \quad \pi'_{ih} \sim \text{beta}\Big(\alpha\pi_{0h}, \alpha\Big(1-\sum_{l\leq h}\pi_{0l}\Big)\Big),\\ \pi_{0h} &= \pi'_{0h}\prod_{l<h}(1-\pi'_{0l}), \quad \pi'_{0h} \sim \text{beta}(1,\kappa),\\ \tau_h &\overset{i.i.d.}{\sim} F_0^*, \quad \Phi_{0h} \overset{i.i.d.}{\sim} GP(\mu,\mathcal{C}_0), \quad \Phi_{1h} \overset{i.i.d.}{\sim} GP(\mu,\mathcal{C}_1).\end{aligned}$$

To clarify, one generates an infinite sequence of kernel locations, $\boldsymbol{\tau} = (\tau_h, h = 1,\ldots,\infty)'$, by sampling independently from F_0^*. These locations are assigned *global* probability weights, $\boldsymbol{\pi}_0 = (\pi_{0h}, h = 1,\ldots,\infty)'$, generated from a stick-breaking process. To allow individuals to vary in the probabilities, while accommodating dependence, subject-specific weights, $\boldsymbol{\pi}_i = (\pi_{ih}, h = 1,\ldots,\infty)'$, are generated from a stick-breaking process centered on the global process. Finally, coefficient functions $\{\Phi_{0h}(\cdot), h = 1,\ldots,\infty\}$ and $\{\Phi_{1h}(\cdot), h = 1,\ldots,\infty\}$ are generated by i.i.d. sampling from Gaussian processes with covariance functions $\mathcal{C}_0$ and $\mathcal{C}_1$, respectively. Individual i is assigned to the hth GP spike function with stick-breaking probability ν_{0h} and to the hth GP slab function with probability ν_{1h}.

Under the hierarchical prior (3.5), the functions for subjects i and j are equal with probability

$$\Pr(\eta_i = \eta_j \mid \alpha, \beta_0, \beta_1) = \frac{1}{(1+\alpha)(1+\beta_0)(1+\beta_1)}. \tag{3.6}$$

This property follows from (3.1) - (3.4) using the Blackwell and MacQueen (1973) Pólya urn scheme. To clarify, note that, under specification (3.1), the functions η_i and η_j are equivalent if and only if $\phi_{0i} = \phi_{0j}$, $\phi_{1i} = \phi_{1j}$ and $\pi_i = \pi_j$. Following the hierarchical prior (3.3), this occurs when two draws from F are equal, two draws from G_0 are equal and two draws from G_1 are equal. Marginalizing out the random components in F, G_0 and G_1, we have $\Pr(\phi_{0i} = \phi_{0j}) = 1/(1+\beta_0)$, $\Pr(\phi_{1i} = \phi_{1j}) = 1/(1+\beta_1)$ and $\Pr(\pi_i = \pi_j) = 1/(1+\alpha)$. Since the priors are independent, (3.6) follows directly.

Note that, although the prior does allow functions to be identical, global clustering of functions is not the only mechanism for borrowing information. The component $f_{0ih}(\mathbf{x}) = K(\mathbf{x},\tau_h)(1-\pi_{ih})$ can be viewed as the hth spike basis function, with $\theta_{0ih} = \Phi_{0,Z_i}(\tau_h)$ the basis coefficient. Similarly, $f_{1ih}(\mathbf{x}) = K(\mathbf{x},\tau_h)\pi_{ih}$ is the hth slab basis function with coefficient $\theta_{1ih} = \Phi_{1,Z_i}(\tau_h)$. Then we have $\eta_i(\mathbf{x}) = \sum_{h=1}^{\infty}\theta_{0ih}f_{0ih}(\mathbf{x}) + \theta_{1ih}f_{1ih}(\mathbf{x})$, a linear combination of random spike and slab functions. Because the spike coefficient functions are concentrated in a neighborhood of 0, the hth component in this specification for $\eta_i(\mathbf{x})$ is close to 0 if $\pi_{ih} \approx 0$. Due to the structure of our specification, for reasonable values of the hyperparameters, π_{ih} will tend to be close to zero except when the index h

is small, so that a few basis functions will dominate for each individual, with the remaining effectively dropping out. Subjects having a dominant basis function at the same location will be clustered together *locally* in a neighborhood of that basis function. This in turn allows subjects to have locally similar curves, while allowing these curves to deviate in other regions of $\mathcal{X}$.

3.3. Finite approximation

To gain additional intuition and facilitate efficient posterior computation, it is useful to consider a finite approximation. Focusing on the case in which F_0^* is the uniform probability measure on the bounded interval, $[a, b]$, we consider the following approximation to F_0^*:

$$\begin{aligned} \widetilde{F}_0 &= \sum_{j=1}^{J} \pi_{0j}\delta_{\tau_j}, \quad \tau_j = a + \frac{(b-a)j}{J+1}, \\ \boldsymbol{\pi}_0 &\sim D_J\biggl(\frac{\kappa}{J}, \ldots, \frac{\kappa}{J}\biggr), \end{aligned}$$

where $\boldsymbol{\pi}_0 = (\pi_{01}, \ldots, \pi_{0J})'$, and D_J is the finite J-dimensional Dirichlet distribution. This form is obtained by first using a discrete uniform approximation to F_0^*, and then relying on the result of Ishwaran and Zarepour (2002) to obtain a finite Dirichlet approximation to the weights.

Treating $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)'$ as a prespecified vector of potential kernel locations, we then obtain the approximation

$$\begin{aligned} \eta_i(\mathbf{x}_i) &= \sum_{h=1}^{J} K(\mathbf{x}_i, \tau_h)\{(1-\pi_{ih})\phi_{i0}(\tau_h) + \pi_{ih}\phi_{i1}(\tau_h)\}, \\ \boldsymbol{\pi}_i &\sim F, \qquad F \sim DP\bigl(\alpha\widetilde{F}_0\bigr), \\ \phi_{i0} &\sim G_0, \quad G_0 \sim DP\bigl(\beta_0 G_0^*\bigr), \\ \phi_{i1} &\sim G_1, \quad G_1 \sim DP\bigl(\beta_1 G_1^*\bigr), \end{aligned} \tag{3.7}$$

where $\mathbf{x}_i$ is an $n_i \times 1$ vector of locations at which η_i is observed, and G_0^* and G_1^* are spike and slab Gaussian Processes as defined at (3.4).

To illustrate the approach, we plot samples from the prior in Figure 1, focusing on the case in which $K(t, \tau_j) = \exp(-\psi||t-\tau_j||^2)$, $\boldsymbol{\tau}$ is a grid of equally-spaced values between 1 and 35, $J = 100$, $\alpha = 1$, and G_{0j} corresponds to the standard normal distribution. For small values of ψ, the curves are very smooth, while for larger values they fluctuate quite rapidly. In the bottom row of Figure 1, we vary $\beta_0 = \beta_1$ while fixing $\psi = 0.05$. Larger values of β decrease the probability of clustering, with each observation having its own curve as $\beta \rightarrow \infty$. Smaller values of β increase the probability of clustering the spike and slab coefficients across

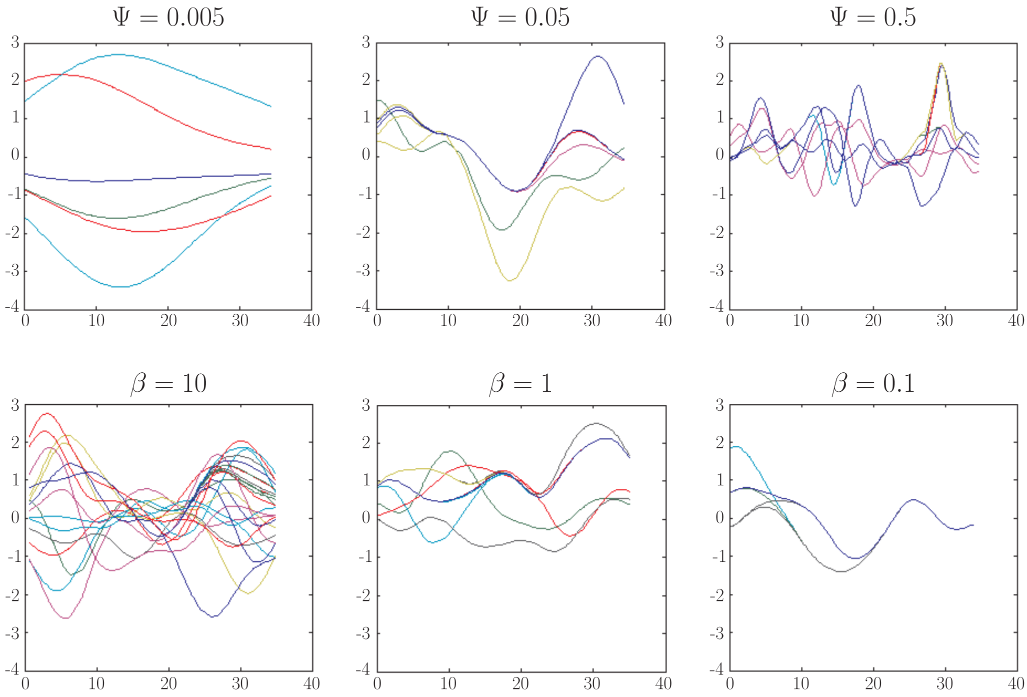

Figure 1. Random draws from the prior distribution (3.7) for different values of Ψ and β.

individuals. However, even when coefficients are clustered together, the HDP specification of the spike and slab mixture allows local deviations of individual curves (as in the $\beta = 0.1$ panel of Figure 1). The 1-35 range for the data is motivated by applications to modeling of hormone curves in the menstrual cycle.

4. Posterior Computation

Focusing on (3.7), we propose a Metropolis within Gibbs algorithm for posterior computation. To avoid computations for the infinite dimensional GP's we focus on the value of the GP at J finite realizations $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)$, where $\mathbf{x}_i \subset \boldsymbol{\tau}, \forall i$. We assume as in (3.7) that $\phi_{ij} \sim G_j$ with $G_j \sim DP(\beta_j G_j^*)$ and $G_j^* = GP(\mathbf{0}, \mathcal{C}_j)$ for $j \in \{0, 1\}$. Therefore, the function ϕ_{ij} evaluated at points $\boldsymbol{\tau}$ is a random variable $\phi_{ij}(\boldsymbol{\tau}) \sim G_j(\boldsymbol{\tau})$ with $G_j(\boldsymbol{\tau}) \sim DP(\beta_j G_j^*(\boldsymbol{\tau}))$ and $G_j^*(\boldsymbol{\tau}) \equiv N_J(\mathbf{0}, \mathcal{C}_j(\boldsymbol{\tau}))$. For posterior computation, consider that:

$$\begin{aligned}
\boldsymbol{y}_i &\sim N_{n_i}(\eta_i(\mathbf{x_i}), \boldsymbol{\Sigma}_i),\\
\eta_i(\mathbf{x}_i) &= \mathbf{K}_i\{(\mathbf{I}_J - \mathbf{S}_i)\phi_{i0}(\boldsymbol{\tau}) + (\mathbf{S}_i)\phi_{i1}(\boldsymbol{\tau})\},\\
z_{ij} &\sim \text{Bernoulli}(\pi_{ij}),\\
\phi_{i0}(\boldsymbol{\tau}) &\sim G_0(\boldsymbol{\tau}), \qquad \phi_{i1}(\boldsymbol{\tau}) \sim G_1(\boldsymbol{\tau}), \qquad (4.1)
\end{aligned}$$

$$
\begin{aligned}
G_0(\boldsymbol{\tau}) &\sim DP(\beta_0 G_0^*(\boldsymbol{\tau})), & G_1(\boldsymbol{\tau}) &\sim DP(\beta_1 G_1^*(\boldsymbol{\tau})),\\
G_0^*(\boldsymbol{\tau}) &\equiv N_J(\mathbf{0}, \mathcal{C}_0(\boldsymbol{\tau})), & G_1^*(\boldsymbol{\tau}) &\equiv N_J(\mathbf{0}, \mathcal{C}_1(\boldsymbol{\tau})),\\
\boldsymbol{\pi}_i &\sim \text{Dirichlet}_J(\alpha\boldsymbol{\pi}_0), & \boldsymbol{\pi}_0 &\sim \text{Dirichlet}_J(\frac{\kappa}{J}),
\end{aligned}
$$

where $\mathbf{K}_i$ is an $n_i \times J$ matrix with elements $K(x_i, \tau_j)$, $\mathbf{I}_J$ is a $J \times J$ identity matrix, z_{ij} is a latent indicator for whether the j^{th} basis coefficient for the i^{th} curve comes from the GP spike ($z_{ij} = 0$) or from the GP slab ($z_{ij} = 1$), $\mathbf{S}_i$ is a $J \times J$ matrix with diagonal elements $\mathbf{z}_i = (z_{i1} \ldots z_{iJ})$, $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$ and the vector $\boldsymbol{\pi}_i = (\pi_{i1} \ldots \pi_{iJ})$.

The likelihood contribution for individual i is $L_i = N_{n_i}(\boldsymbol{y}_i \mid \eta_i(\mathbf{x}_i), \boldsymbol{\Sigma}_i)$. Given the observed data, our sampling algorithm proceeds by first allocating individual curves to clusters and then sampling those cluster-specific functions from their full conditional posterior using the methods of Escobar and West (1995) and MacEachern and Müller (1998). The conditional prior distributions for $\phi_{i0}(\boldsymbol{\tau})$ and $\phi_{i1}(\boldsymbol{\tau})$ are

$$
\phi_{i0}(\boldsymbol{\tau}) \mid \Psi_{j0}^{(i)}(\boldsymbol{\tau}) \sim \frac{\beta_0}{\beta_0 + n - 1} N_J(\phi_{i0}(\boldsymbol{\tau}) \mid \mathbf{0}, \mathcal{C}_0(\boldsymbol{\tau})) + \sum_j \frac{p_{j0}^{(i)}}{\beta_0 + n - 1} \delta_{\Psi_{j0}(\boldsymbol{\tau})}, \quad (4.2)
$$

$$
\phi_{i1}(\boldsymbol{\tau}) \mid \Psi_{j1}^{(i)}(\boldsymbol{\tau}) \sim \frac{\beta_1}{\beta_1 + n - 1} N_J(\phi_{i1}(\boldsymbol{\tau}) \mid \mathbf{0}, \mathcal{C}_1(\boldsymbol{\tau})) + \sum_j \frac{p_{j1}^{(i)}}{\beta_1 + n - 1} \delta_{\Psi_{j1}(\boldsymbol{\tau})}, \quad (4.3)
$$

where $\Psi_{j0}(\boldsymbol{\tau})$ and $\Psi_{j1}(\boldsymbol{\tau})$ are the common basis coefficients for all curves falling in the j^{th} cluster for the spike and slab curves, respectively. Take p_{j0} and p_{j1} to be the number of curves that fall in the j^{th} cluster for the spike and slab components, respectively. The superscript (i) signifies the vector obtained without the ith observation. To update (4.2) and (4.3) with L_i, we define $\boldsymbol{y}_{i0} = \boldsymbol{y}_i - \mathbf{K}_i \mathbf{S}_i \phi_{i1}(\boldsymbol{\tau})$ and $\boldsymbol{y}_{i1} = \boldsymbol{y}_i - \mathbf{K}_i(\mathbf{I} - \mathbf{S}_i)\phi_{i0}(\boldsymbol{\tau})$. Then we have

$$
\phi_{i0}(\boldsymbol{\tau}) | \boldsymbol{y}_i, \Psi_{j0}^{(i)}(\boldsymbol{\tau}) \sim q_{i0} N(\phi_{i0}(\boldsymbol{\tau}) \mid E_{i0}, V_{i0}) + \sum_j q_{ij} \delta_{\Psi_{j0}^{(i)}(\boldsymbol{\tau})},
$$

$$
\phi_{i1}(\boldsymbol{\tau}) | \boldsymbol{y}_i, \Psi_{j1}^{(i)}(\boldsymbol{\tau}) \sim w_{i0} N(\phi_{i1}(\boldsymbol{\tau}) | E_{i1}, V_{i1}) + \sum_j w_{ij} \delta_{\Psi_{j1}^{(i)}(\boldsymbol{\tau})},
$$

where $V_{i0} = \{(\mathbf{I}_J - \mathbf{S}_i)' \mathbf{K}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{K}_i (\mathbf{I}_J - \mathbf{S}_i) + \mathcal{C}_{\mathbf{0}}(\boldsymbol{\tau})^{-1}\}^{-1}$, $E_{i0} = V_{i0}(\mathbf{I}_J - \mathbf{S}_i)' \mathbf{K}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_{i0}$, $V_{i1} = \{\mathbf{S}_i' \mathbf{K}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{K}_i \mathbf{S}_i + \mathcal{C}_{\mathbf{1}}(\boldsymbol{\tau})^{-1}\}^{-1}$, and $E_{i1} = V_{i1} \mathbf{S}_i' \mathbf{K}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_{i1}$. The weights are

$$
\begin{aligned}
q_{i0} &\propto \beta_0 \int N(\boldsymbol{y}_{i0} | \mathbf{K}_i (\mathbf{I}_J - \mathbf{S}_i) \phi_{i0}(\boldsymbol{\tau}), \boldsymbol{\Sigma}_i) dN(\phi_{i0}(\boldsymbol{\tau}) | \mathbf{0}, \mathcal{C}_0(\boldsymbol{\tau}))\\
&= \beta_0 \frac{N(\boldsymbol{y}_{i0} | \mathbf{0}, \boldsymbol{\Sigma}_i) N(\mathbf{0} \mid \mathbf{0}, \mathcal{C}_0(\boldsymbol{\tau}))}{N(E_{i0} | \mathbf{0}, V_{i0})},
\end{aligned}
$$

$$\begin{aligned} q_{ij} &\propto p_{0j}^{(i)} N(\boldsymbol{y}_{i0}|\mathbf{K}_i(\mathbf{I}_J - \mathbf{S}_i)\Psi_{j0}^{(i)}(\boldsymbol{\tau}), \boldsymbol{\Sigma}_i), \\ w_{i0} &\propto \beta_1 \int N(\boldsymbol{y}_{i1}|\mathbf{K}_i\mathbf{S}_i\phi_{i1}(\boldsymbol{\tau}), \boldsymbol{\Sigma}_i) dN(\phi_{i1}(\boldsymbol{\tau})|\mathbf{0}, \mathcal{C}_1(\boldsymbol{\tau})) \\ &= \beta_1 \frac{N(\boldsymbol{y}_{i1}|\mathbf{0}, \boldsymbol{\Sigma}_i) N(\mathbf{0}\,|\,\mathbf{0}, \mathcal{C}_1(\boldsymbol{\tau}))}{N(E_{i1}|\mathbf{0}, V_{i1})}, \\ w_{ij} &\propto\ p_{1j}^{(i)} N(\boldsymbol{y}_{i1}|\mathbf{K}_i\mathbf{S}_i\Psi_{j1}^{(i)}(\boldsymbol{\tau}), \boldsymbol{\Sigma}_i). \end{aligned}$$

The Gibbs sampling proceeds by allocating basis coefficients from the spike and the slab. Basis coefficients from the GP spike are clustered with other coefficients with probability q_{ij} or are sampled from the posterior base distribution with probability q_{i0}. Similarly GP slab coefficients are clustered with other coefficients with probability w_{ij}, or sampled from the posterior base with probability w_{i0}. The vector $\mathbf{g}_0 = (g_{01} \ldots g_{0n})'$ indicates which cluster each of the spike functions falls into, and the vector $\mathbf{g}_1 = (g_{11} \ldots g_{1n})'$ indicates which cluster each of the slab functions falls into. To improve mixing, we update the cluster-specific coefficients using the approach of Bush and MacEachern (1996): $\Psi_{j0} \sim N_J(E_{j0}^*, V_{j0}^*)$ and $\Psi_{j1} \sim N_J(E_{j1}^*, V_{j1}^*)$, where $V_{j0}^* = \{\mathcal{C}_0(\boldsymbol{\tau})^{-1} + \sum_{h:g_{0h}=j}(\mathbf{I}_J - \mathbf{S}_h)'\mathbf{K}_h'\Sigma_h^{-1}\mathbf{K}_h(\mathbf{I}_J - \mathbf{S}_h)\}^{-1}$, $E_{j0}^* = V_{j0}^* \sum_{h:g_{0h}=j}(\mathbf{I}_J - \mathbf{S}_h)'\mathbf{K}_h'\Sigma_h^{-1}\boldsymbol{y}_{0h}$, $V_{j1}^* = \{\mathcal{C}_1(\boldsymbol{\tau})^{-1} + \sum_{h:g_{1h}=j}\mathbf{S}_h'\mathbf{K}_h'\Sigma_h^{-1}\mathbf{K}_h\mathbf{S}_h\}^{-1}$, $E_{j1}^* = V_{j1}^* \sum_{h:g_{1h}=j}\mathbf{S}_h'\mathbf{K}_h'\Sigma_h^{-1}\boldsymbol{y}_{1h}$. We update the precision parameters, β_0 and β_1, using the data augmentation approach of Escobar and West (1995).

The J-dimensional Dirichlet prior for $\boldsymbol{\pi}_i$ is not conjugate with Bernoulli z_{ij}, so we introduce a Metropolis-Hastings step in which we draw a proposal value, $\boldsymbol{\pi}_i^*$, from the density:

$$q(\boldsymbol{\pi}_i^*\,|\,\boldsymbol{\pi}_0, \boldsymbol{Z}_i = 1) = \text{Dirichlet}_J\big(\boldsymbol{\pi}_i^*\,|\,\alpha\pi_{01} + I(z_{i1} = 1), \ldots, \alpha\pi_{0J} + I(z_{iJ} = 1)\big),$$

where $I(\cdot)$ is an indicator function equal to one if the statement $\cdot$ is true. We accept $\boldsymbol{\pi}_i^*$ with probability r_1; otherwise, we keep the value from the current (g^{th}) iteration, $\boldsymbol{\pi}_i^g$. The acceptance probability is:

$$r_1 = min\bigg(1, \frac{p(\boldsymbol{\pi}_i^*\,|\,\boldsymbol{\pi}_0, \boldsymbol{Z}_i) q(\boldsymbol{\pi}_i^g\,|\,\boldsymbol{\pi}_0, \boldsymbol{Z}_i = 1)}{p(\boldsymbol{\pi}_i^g\,|\,\boldsymbol{\pi}_0, \boldsymbol{Z}_i) q(\boldsymbol{\pi}_i^*\,|\,\boldsymbol{\pi}_0, \boldsymbol{Z}_i = 1)}\bigg),$$

where $p(\boldsymbol{\pi}_i\,|\,\boldsymbol{\pi}_0, \boldsymbol{Z}_i) = \text{Dirichlet}(\boldsymbol{\pi}_i\,|\,\alpha\boldsymbol{\pi}_0) \prod_{j=1}^J \text{Bernoulli}(z_{ij}\,|\,\pi_{ij})$.

The indicator variable, z_{ij}, determining whether the j^{th} basis coefficient for function i falls in the spike or slab is sampled from a Bernoulli distribution with parameter $\widehat{\pi_{ij}} = a/(a + b)$, where

$$\begin{aligned} a &= \pi_{ij} N(\boldsymbol{y}_i\,|\,\mathbf{K}_i\{(\mathbf{I_J} - \mathbf{S}_i^*)\phi_{i0}(\boldsymbol{\tau}) + (\mathbf{S}_i^*)\phi_{i1}(\boldsymbol{\tau})\}, \Sigma_i) \\ b &= (1 - \pi_{ij}) N(\boldsymbol{y}_i\,|\,\mathbf{K}_i\{(\mathbf{I_J} - \mathbf{S}_i^{**})\phi_{i0}(\boldsymbol{\tau}) + (\mathbf{S}_i^{**})\phi_{i1}(\boldsymbol{\tau})\}, \Sigma_i), \end{aligned}$$

where $\mathbf{S}_i^* = \mathrm{diag}(\mathbf{z}_i^*)$, $\mathbf{z}_i^* = (z_{i1}, \ldots, z_{ij-1}, 1, z_{ij+1}, \ldots, z_J)$, $\mathbf{S}_i^{**} = \mathrm{diag}(\mathbf{z}_i^{**})$, and $\mathbf{z}_i^{**} = (z_{i1}, \ldots, z_{ij-1}, 0, z_{ij+1}, \ldots, z_J)$.

A Gibbs step is not possible for updating $\boldsymbol{\pi}_0$ and instead we propose another Metropolis-Hastings step. We sample the proposal values $\boldsymbol{\lambda}^* = (\lambda_1^* \ldots \lambda_{J-1}^*)'$ from the proposal density $N_{J-1}(\boldsymbol{\lambda}^g, \Sigma^*)$, where $\boldsymbol{\lambda}^g$ are the values of $\boldsymbol{\lambda}$ from the previous Gibbs iteration, $\Sigma^* = s\mathbf{I}_{J-1}$ and $\mathbf{I}_{J-1}$ is an identity matrix. Sampled parameters are transformed to $\boldsymbol{\pi}_0^*$ through the function $\pi_{0j}^* = h_j(\lambda_j^*) = \exp(\lambda_j^*)/(1+\exp(\lambda_j^*))(1-\sum_{h=1}^{j-1} \pi_{0h}^*)$, and $q(\boldsymbol{\pi}_0^* \,|\, \boldsymbol{\lambda}^g, \Sigma^*)$ is a modified logit-normal distribution.

The acceptance probability for this Metropolis-Hastings step is given by:

$$r_2 = \min\left(1, \frac{p(\boldsymbol{\pi}_0^* \,|\, \boldsymbol{\pi}_i) q(\boldsymbol{\pi}_0^g \,|\, \boldsymbol{\lambda}^*, \Sigma^*)}{p(\boldsymbol{\pi}_0^g \,|\, \boldsymbol{\pi}_i) q(\boldsymbol{\pi}_0^* \,|\, \boldsymbol{\lambda}^g, \Sigma^*)}\right),$$

where $p(\boldsymbol{\pi}_0 \,|\, \boldsymbol{\pi}_i) = \mathrm{Dirichlet}(\boldsymbol{\pi}_0 \,|\, \kappa/J) \prod_{i=1}^n \mathrm{Dirichlet}(\boldsymbol{\pi}_i \,|\, \alpha\boldsymbol{\pi}_0)$. In implementing the Metropolis step, we run the Gibbs-Metropolis algorithm for a period during which we allow the proposal density variance, s, to vary in order to obtain an acceptance probability of approximately 0.20. After this period, we fix the value of s and implement our Metropolis within Gibbs algorithm.

5. Simulation Examples

We evaluated the approach using two small simulation studies. For the first simulation, we generated curves for 50 individuals with half of the curves following the function $f(x) = 0$ and half following the function $f(x) = \exp(0.38x)/1000$. The second simulation generated 15 curves from each of the three true mean functions:

$$f_1(x) = 0, \quad f_2(x) = -\cos\Big(\frac{5x}{n+1}\Big), \quad f_3(x) = 10\Big(\frac{x-10}{n+1}\Big)^2$$

The outcome was simulated at the locations $\mathbf{x} = 1, \ldots, 20$ with Gaussian random noise added to the functions at each of the points. We let $\alpha = 1, \beta_0 = 1, \beta_1 = 1$, and $\kappa = 1$. The covariance functions were taken to be squared exponential: $\mathcal{C}_1(t,t') = \exp(-\psi_c(t-t')^2)$, $\mathcal{C}_0(t,t') = 0.01 * \mathcal{C}_1(t,t')$, and $K(t,t') = \exp(-\psi_k(t-t')^2)$. We took $\psi_c = 0.1$ and $\psi_k = 0.05$ to allow the coefficient functions, $\phi_{i0}(\boldsymbol{\tau})$ and $\phi_{i1}(\boldsymbol{\tau})$, to vary considerably with location, while producing a smooth η_i. We tuned the variance parameter for the Metropolis step for 1,000 iterations. We then ran the algorithm for an additional 50,000 iterations, retaining every 50^{th} iteration for inference and discarding the initial 200 iterations as a burn-in period. The MCMC algorithm converged rapidly after the tuning parameter was fixed. However, we note that some label shifting did occur. Coefficients for points τ_s and τ_{s+h} may have nearly equal ability to define a curve when h is small and,

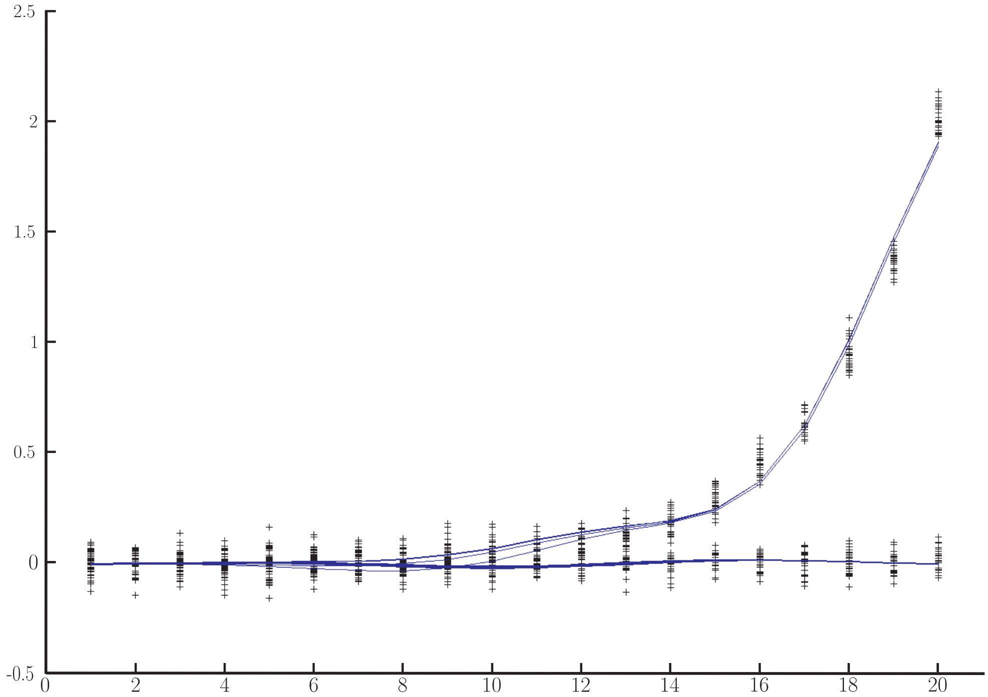

Figure 2. Simulated data (+'s) and posterior mean curves for the first simulation.

in this case, τ_s may be sampled from the GP slab for a period before τ_{s+h} is sampled from the slab. This did not affect the convergence of $\eta(\boldsymbol{\tau})$. Figure 2 shows the mean of each of the estimated functions for the first simulation, and Figure 3 shows the estimated functions for the second. In both simulations, the true trajectories were accurately characterized.

6. Application: Progesterone Trajectories

We applied our approach to a study of progesterone data previously analyzed by Brumback and Rice (1998). The data consist of progesterone levels ascertained through daily urine samples during the menstrual cycle. Fifty-one women provided samples over a total of 91 cycles, 22 of which were conceptive and 69 were non-conceptive. Measurements were not complete and progesterone values were missing for some days. Brumback and Rice (1998) use a mixed-effects model to fit flexible smoothing splines to these data. We allowed progesterone curves in conceptive and non-conceptive cycles to have separate nonparametric kernel priors, by fitting the model in (4.1) separately for conceptive and non-conceptive cycles.

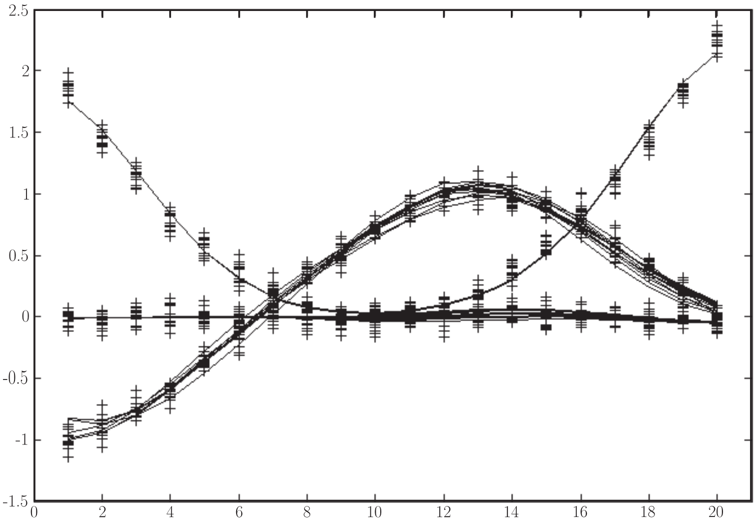

Figure 3. Simulated data (+'s) and posterior mean curves for the second simulation.

The analysis was implemented as in the simulation examples of Section 4, with $\boldsymbol{\tau}$ chosen to include each day of the menstrual cycle relative to the day of ovulation (days -8 to 15) and 10 additional equally space points over the range [-7, 16]. We note that by borrowing information between functions, our method is very useful for imputing functional form in the presence of missing data. Figure 4 plots posterior mean functions for all 91 cycles and Figure 5 shows the posterior mean functions and 95% credible intervals for conceptive and non-conceptive cycles alone. We demonstrate the plausibility of borrowing information between curves while allowing local deviations in Figure 6, which shows the posterior mean curve and observed data for four individuals. The data for the four observations show little deviation between days -8 and -2, so the three curves are locally clustered with high probability during these days. However, between days 6 and 15 two of the curves exhibit local deviation from the other two curves, in order to follow the observed data more closely.

Scientific inference often centers on assessing differences between posterior functions during some duration. Such inference is difficult in Brumback and Rice's approach, relying on bootstrap re-sampling. The method proposed in this paper is much more straightforward, and consists of comparing the proportion of MCMC samples for one function that fall above or below the other function. We compare the mean conceptive and non-conceptive curves at the beginning of the cycle by comparing the functional value of these curves at days in the interval [-8,0] and find the conceptive curve is uniformly higher than the non-conceptive curve with posterior probability of 89.3%. Not surprisingly, at the end of the

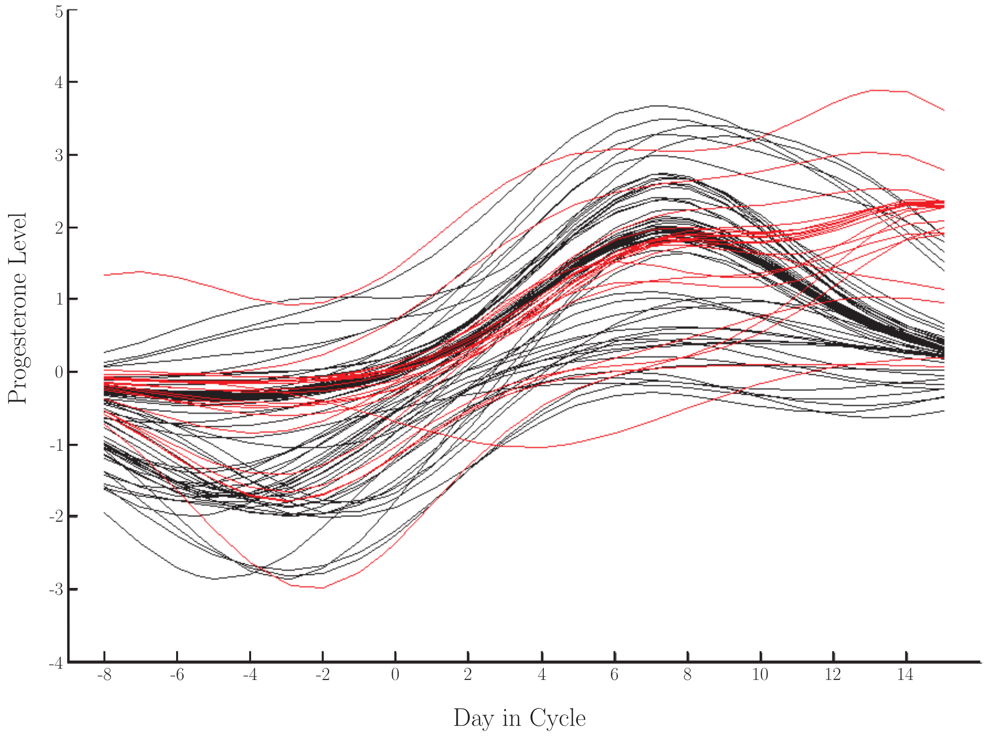

Figure 4. Individual posterior mean progesterone curves for conceptive (red) and non-conceptive (black) cycles.

cycle in the interval [10, 15] the mean conceptive curve is higher with probability 94.4%.

7. Discussion

This article has proposed nonparametric Bayes methods for modeling random functions, allowing individual curves to vary flexibly while adaptively borrowing information between functions. We have implemented this method in simulated data and to predict progesterone levels during the menstrual cycle. However, it should prove useful in many applications involving functional data on multiple individuals, a common problem in reproductive and environmental research.

The proposed kernel-based nonparametric approach extends earlier approaches that allowed only global clustering of curves. By incorporating an HDP for basis function selection, we allow global clustering of curves while retaining the flexibility for local deviations between clustered curves. Further, by specifying functions as a kernel mixture of the latent GP spike and slab functions, we lessen our dependence on the covariance specification and diminish the tendency of the

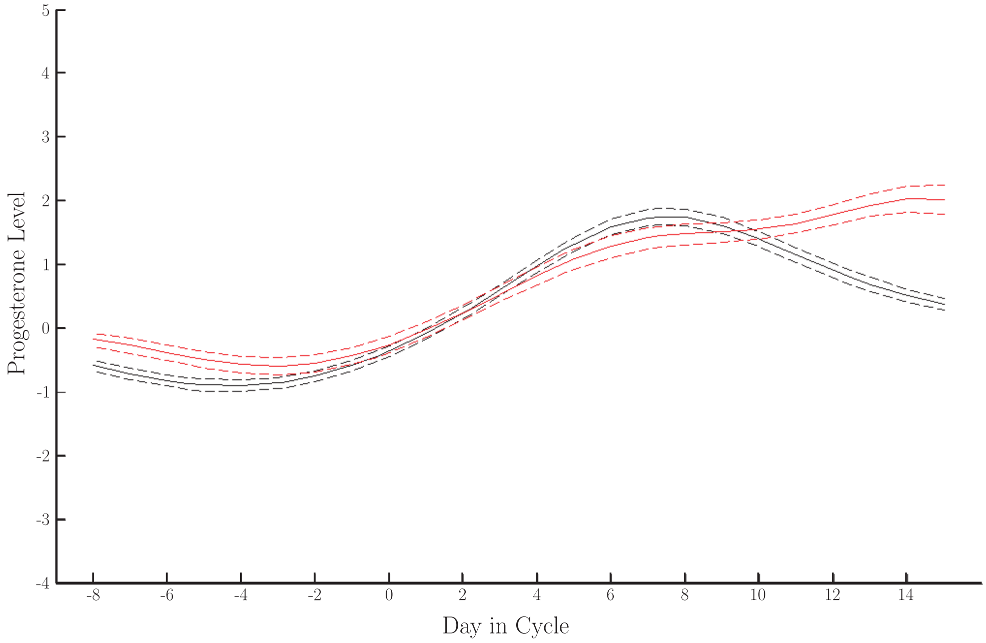

Figure 5. Posterior mean progesterone curve (solid line) and 95% credible intervals (dashed lines) during conceptive cycles (black) and non-conceptive cycles (red).

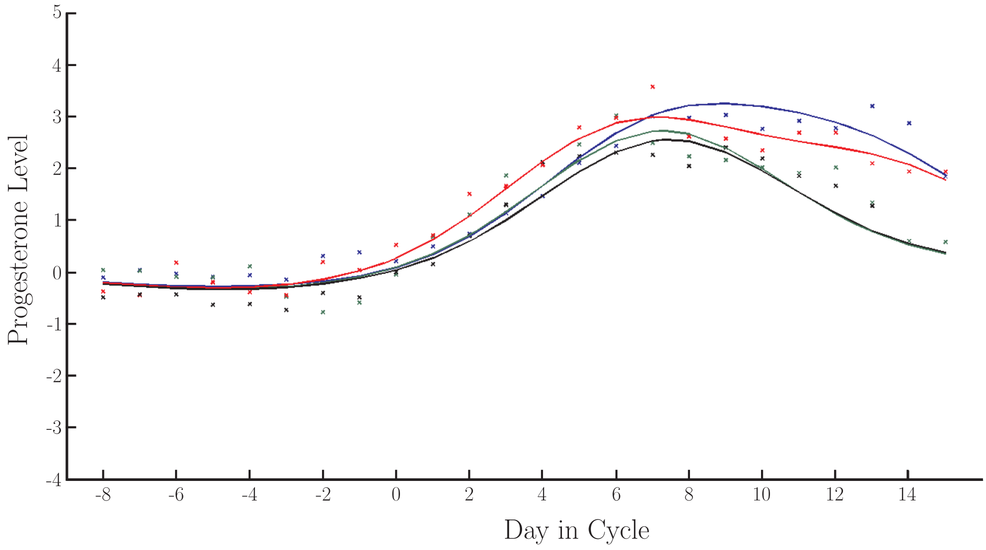

Figure 6. Posterior mean progesterone curves (solid lines) and observed progesterone levels for four selected non-conceptive cycles. Estimated progesterone curves and observed data are linked by color.

GPs to 'chase' data points. In future research it would be interesting to extend the method to allow joint nonparametric modeling of functional predictors with an outcome variable.

Acknowledgement

This research was supported, in part, by the Intramural Research Program of the NIH, and NIEHS.

References

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

Berry, D. A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7**, 558-568.

Bigelow, J. L. and Dunson, D.B. (2006). Posterior simulation across nonparametric models for functional clustering. ISDS Discussion Paper.

Bigelow, J. L. and Dunson, D. B. (2007). Bayesian adaptive regression splines for hierarchical data. *Biometrics* **63**, 724-732.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353-355.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**, 961-976.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275-285.

Chakraborty, S., Ghosh, M., and Mallick, B. (2005). Bayesian non-linear regression for large p, small n problems. Technical Report. University of Florida.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An Anova model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205-215.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268-277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100**, 1021-1035.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Springer, New York.

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12**, 941-963.

Kimeldorf, G. S. and Wahba, G. (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41**, 495-502.

Liang, F., Liao, M., Mukherjee, S. and West, M. (2006). Nonparametric Bayesian kernel models. *ISDS Discussion Paper*, Duke University.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351-357.

MacEachern, S. N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria, VA.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223-238.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *J. Roy. Statist. Soc. Ser. B* **68**, 179-199.

Müller, H. G. (2005). Functional modelling and classification of longitudinal data. *Scand. J. Statist.* **32**, 223-240.

Pillai, N., Liang, F., Mukerjee, S., Wolpert, R. and Wu, Q. (2006). Characterizing the function space for Bayesian kernel models.

Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247**, 978-982.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *J. Roy. Statist. Soc. Ser. B* **68**, 305-332.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statist. Sinica* **2**, 639-650.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317-343.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning*, 21-52.

Thompson, W. K. and Rosen, O. (2006). A Bayesian model for sparse functional data. Technical Report.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarhcical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566-1581.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211-244.

Tomlinson, G. (1998). Analysis of Densities. Unpublished dissertation. University of Toronto.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Society for Industrial and Applied Mathematics, Philadelphia.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. *Advances in Kernel Methods: Support Vector Learning*, 69-88.

Xia, G. and Gelfand, A. E. (2005). Stationary process approximation for the analysis of large spatial datasets. *ISDS DIscussion Paper* **2005-24**, Duke University, Durham, NC.

Zhao, X., Marron, J. S. and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statist. Sinica* **14**, 789-808.

Division of Biostatistics, University of Minnesota, A452 Mayo Building, 420 Delaware St SE, Box 303, Minneapolis, MN 55455, U.S.A.

E-mail: macl0029@umn.edu

Department of Statistical Science, Duke University, 219A Old Chemistry Building, Box 90251, Durham, NC 27708, U.S.A.

E-mail: dunson@stat.duke.edu

(Received March 2007; accepted December 2007)