Biometrics 65, 772–780
September 2009
DOI: 10.1111/j.1541-0420.2008.01163.x

Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors

Bruno Scarpa[1,*] and David B. Dunson[2]

[1]Department of Statistical Sciences, University of Padua, via Cesare Battisti, 241 Padua, Italy
[2]Department of Statistical Sciences, 218 Old Chemistry Building, Duke University, Durham, North Carolina 27708, U.S.A.
[*]*email:* scarpa@stat.unipd.it

Summary. A variety of flexible approaches have been proposed for functional data analysis, allowing both the mean curve and the distribution about the mean to be unknown. Such methods are most useful when there is limited prior information. Motivated by applications to modeling of temperature curves in the menstrual cycle, this article proposes a flexible approach for incorporating prior information in semiparametric Bayesian analyses of hierarchical functional data. The proposed approach is based on specifying the distribution of functions as a mixture of a parametric hierarchical model and a nonparametric contamination. The parametric component is chosen based on prior knowledge, while the contamination is characterized as a functional Dirichlet process. In the motivating application, the contamination component allows unanticipated curve shapes in unhealthy menstrual cycles. Methods are developed for posterior computation, and the approach is applied to data from a European fecundability study.

Key words: Clustering; Functional Dirichlet process; Latent trajectory curves; Mixture model; Nonparametric Bayes.

1. Introduction

In many applications, interest focuses on studying the distribution of a random curve. For example, in studies of female reproductive functioning, one may collect data on the level of a hormone, or of basal body temperature (bbt), across the menstrual cycle. As shown in Figure 1, for healthy women of reproductive age, bbt curves follow a characteristic trajectory. During the follicular phase of the cycle leading up to ovulation, the bbt values tend to be low, with the nadir occurring close to the time of ovulation. After ovulation, bbt rises progressively before dropping prior to the next cycle. This characteristic pattern has been used as a basis for algorithms for identifying the fertile days of the menstrual cycle, with the last day of hypothermia prior to the postovulatory rise in bbt a marker of the day of ovulation and end of the fertile window (e.g., Vincent, 1964; Marshall, 1979).

Potentially, one can characterize the bbt curve data for menstrual cycles from different women using a hierarchical model of the form:

$$\begin{aligned} y_{ij}(t) &= \eta_{ij}(t) + \epsilon_{ij}(t), \quad \epsilon_{ij}(t) \sim N(0, \sigma^2) \\ \eta_{ij} &\sim G_i, \quad \mathcal{G} = \{G_i\}_{i=1}^n \sim P, \end{aligned} \tag{1}$$

where $y_{ij}(t)$ is the measured bbt level at time t of the jth cycle ($j = 1, \ldots, n_i$) for woman i, $\eta_{ij} : \mathcal{T} \to \Re$ is a smooth function, $\mathcal{T}$ is a subset of $\Re^+$, $\epsilon_{ij}(t)$ is a measurement error, G_i is a prior for the distribution of functions for woman i, and P is a prior for the collection $\mathcal{G}$ of distributions for the different women. Hence, G_i characterizes variability within woman i, while P characterizes variability between women.

Expression (1) provides a general framework for hierarchical functional data analysis, and a variety of approaches can be considered for specifying the components. A common strategy is to simplify the problem by considering a basis representation:

$$\eta_{ij}(t) = \sum_{h=1}^{k} \theta_{ijh} b_h(t), \quad \forall t \in \mathcal{T}, \tag{2}$$

where $\{b_h\}_{h=1}^k$ is a prespecified basis and $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \ldots, \theta_{ijk})'$ are cycle-specific basis coefficients. Then, one can potentially choose a hierarchical normal model for the basis coefficients as follows:

$$\boldsymbol{\theta}_{ij} \sim N_k(\boldsymbol{\alpha}_i, \boldsymbol{\Omega}), \quad \boldsymbol{\alpha}_i \sim N_k(\boldsymbol{\alpha}, \boldsymbol{\Omega}_1), \tag{3}$$

where $\boldsymbol{\alpha}_i$ provide woman-specific coefficients, $\boldsymbol{\Omega}$ measures within-woman variability, $\boldsymbol{\alpha}$ are global mean basis coefficients, and $\boldsymbol{\Omega}_1$ measures between woman variability.

Many papers have been published dealing with nested and hierarchical functional data, using classical or Bayesian approaches: Brumback and Rice (1998) present a flexible smoothing spline-based method treating individual-specific curves as fixed effects, Morris et al. (2003) developed a wavelet-based methodology for modeling functional data occurring within a nested hierarchy, and Morris and Carroll (2006) propose a unified Bayesian wavelet-based approach for the much more general functional mixed models framework; Baladandayuthapani et al. (2008) extend their methodology, by accommodating for more general between-curve covariance structures working with splines.

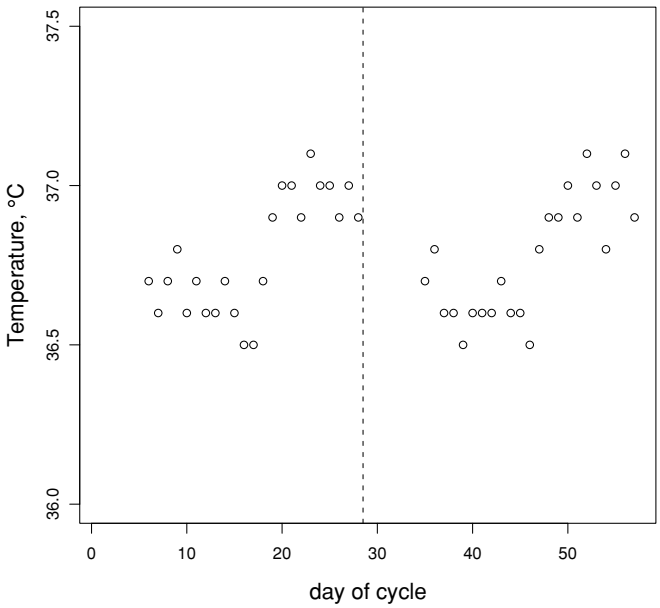

Figure 1. The bbt data for two cycles from a woman. These data exhibit the characteristic biphasic shape expected for a healthy, ovulating woman. The vertical line shows the beginning of the second cycle.

In the one-level simplification of equations (1)–(3), recent articles propose methods to allow uncertainty in basis function selection (Bigelow and Dunson, 2007; Thompson and Rosen, 2007). To avoid the assumption of normality of the basis coefficients within a wavelet model, Ray and Mallick (2006) proposed using a Dirichlet process (DP) prior (Ferguson, 1973, 1974). Bigelow and Dunson (2005) generalize this approach to allow uncertainty in basis function selection. Rodriguez, Dunson, and Gelfand (2008) avoid the basis function representation through the use of dependent DP mixtures. Because of the almost sure discreteness property of the DP, these approaches group subjects into functional clusters. To generalize these methods to the two-level setting, one can use a hierarchical DP (HDP; Teh et al., 2006) as a prior for $\mathcal{G}$. MacLehose and Dunson (2008) use the HDP in a kernel-based prior for functional data analysis. An alternative Bayesian approach for hierarchical functional clustering was proposed by Heard, Holmes, and Stephens (2006).

These approaches are useful for cases in which there is limited prior information about the shapes of the functions under study and one wants a highly flexible model. However, in many cases, such as in the menstrual cycle hormone and bbt curve applications motivating this article, there is abundant prior information available. It is not at all straightforward to include such information within the flexible semiparametric approaches currently available for modeling of distributions of curves. Potentially, one could sacrifice some of this flexibility in favor of a parametric model. However, although a known parametric model may provide a good characterization of many or even most of the curves, it is often the case that a subset of the curves are irregular. For example, in the menstrual cycle application, one may be able to choose a parametric model for bbt curves in healthy cycles, but it is much more difficult to parametrically characterize all the possible curves that may occur in unhealthy cycles.

To address this problem, we propose modeling $\mathcal{G}$ in specification (1) using a mixture of a parametric component and a nonparametric *contamination*. In particular, the nonparametric component will be characterized as a functional DP. For a well-chosen parametric model, curves drawn from the nonparametric component can be considered irregular and likely unhealthy. Hence, a useful approach for inferences on predictors of unhealthy curves is to allow the mixing proportion to depend on predictors.

In Section 2, we provide background motivation for the bbt curve applications, proposing a parametric hierarchical model. Section 3 proposes the semiparametric mixture specification and considers properties. Section 4 develops a Markov chain Monte Carlo (MCMC) approach for posterior computation. Section 5 applies the approach to simulated data. Section 6 contains an application to a large European fecundability study, and Section 7 discusses the results.

2. Parametric Modeling of Temperature Curves

As illustrated in Figure 1, bbt curves in ovulation cycles typically have a biphasic shape with a low plateau in the follicular phase, a temperature minimum at or just prior to ovulation, followed by a sharp rise after ovulation to a plateau (Dunlop, Schultz, and Frank, 2005). Sometimes bbt does not show the typical shape, indicating an unhealthy cycle. Royston and Abrams (1980) and Carter and Blight (1981) proposed approaches for estimating the shift in bbt as a marker of ovulation.

Let n_{ij} be the number of days in the jth cycle of woman i ($i = 1 \ldots, n$) and let $\mathbf{y}_{ij} = [y_{ij1}, \ldots, y_{ijn_{ij}}]^T$ denote the temperatures in cycle j of woman i. We use a piecewise linear function with three components as a simple shape that mimics the classical bbt pattern.

$$y_{ijt} = \begin{cases} \theta_{1ij} + \varepsilon_{ijt} & 1 \le t \le k_{ij} \\ \theta_{1ij} + \theta_{2ij}\left(\dfrac{t - k_{ij}}{r_{ij}}\right) + \varepsilon_{ijt} & k_{ij} < t \le k_{ij} + r_{ij} \\ \theta_{1ij} + \theta_{2ij} + \varepsilon_{ijt} & k_{ij} + r_{ij} < t \le n_{ij} \end{cases}$$

$$\varepsilon_{ijt} \sim N(0, \sigma^2),$$

where k_{ij} and r_{ij} indicate the last day of hypothermia and the number of days in which temperature rises, respectively; θ_{1ij} is the temperature for the follicular phase; θ_{2ij} is the rate of increase in temperature, the shift in bbt, following ovulation; and σ^2 is the measurement error variance. This model may be rewritten as

$$\mathbf{y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\theta}_{ij} + \boldsymbol{\varepsilon}_{ij} \qquad \boldsymbol{\varepsilon}_{ij} \sim N_{n_{ij}}\left(\mathbf{0}, I_{n_{ij}}\sigma^2\right), \tag{4}$$

where $\boldsymbol{\theta}_{ij} = [\theta_{1ij}, \theta_{2ij}]^T$ and $\mathbf{X}_{ij}$ is the $n_{ij} \times 2$ matrix

$$\mathbf{X}_{ij}^T = \begin{bmatrix} \underbrace{1 \ \ldots \ 1}_{k_{ij}} & \underbrace{1 \ \ 1 \ \ldots \ 1}_{r_{ij}} & \underbrace{1 \ \ 1 \ \ldots \ 1}_{n_{ij}-k_{ij}-r_{ij}} \\ \underbrace{0 \ \ldots \ 0}_{k_{ij}} & \underbrace{\frac{1}{r} \ \ \frac{2}{r} \ \ldots \ 1}_{r_{ij}} & 1 \ \ 1 \ \ldots \ 1 \end{bmatrix}.$$

The cycle-specific curve is characterized as $\eta_{ij}(t) = \mathbf{x}_{ij}(t)^T \boldsymbol{\theta}_{ij}$, with $\mathbf{x}_{ij}(t)$ a vector of basis functions evaluated at time t and $\boldsymbol{\theta}_{ij}$ a vector of cycle-specific basis coefficients. One can easily modify the choice of basis functions in other applications.

To model within- and between-woman heterogeneity, we use equation (3), with $\boldsymbol{\Omega} = \text{diag}\{\omega_1, \omega_2\}$ and $\boldsymbol{\Omega}_1 = \text{diag}\{\omega_{11}, \omega_{12}\}$. A Bayesian specification is completed with priors,

$$k_{ij} \sim \mathcal{U}(m_{ij}, m_{ij} + 20), \qquad r_{ij} \sim \mathcal{U}(1, 15),$$

where $\mathcal{U}(a, b)$ indicates the Discrete uniform distribution between a and b, m_{ij} is the first day after bleeding for the cycle j of woman i, and

$$\boldsymbol{\alpha} \sim N_{k+1}(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}_\alpha) \qquad \sigma^{-2} \sim \mathcal{G}(c, d),$$
$$\omega_h^{-1} \sim \mathcal{G}(a_h, b_h), \quad \omega_{1h}^{-1} \sim \mathcal{G}(a_{1h}, b_{1h}), \quad \text{for } h = 1, 2,$$

where $\mathcal{G}(a, b)$ denotes the Gamma distribution with mean a/b and variance a/b^2. Here $\boldsymbol{\alpha}_0$, $\boldsymbol{\Sigma}_\alpha$, c, d, a_h, b_h, a_{h1}, and b_{h1} are prespecified hyperparameters. Note that dependence in equation (4) is induced in marginalizing out the random effects.

3. Nonparametric Contamination for Functional Data

We propose to combine the parametric model of Section 2.2 with a functional Dirichlet process (FDP) through a mixture specification. Our prior expectation is that the majority of bbt trajectories can be well characterized by the parametric model of Section 2.2, but that there is a minority of unhealthy cycles for which this model does not apply. Our proposed hierarchical model is specified as

$$\begin{aligned} \eta_{ij} &\sim \pi_{ij} \delta_{\mathbf{X}_{ij}} \boldsymbol{\theta}_{ij} + (1 - \pi_{ij}) G, \\ G &= \sum_{h=1}^{\infty} p_h \delta_{\Theta_h}, \qquad \Theta_h \sim GP(\mu, \mathcal{C}), \end{aligned} \tag{5}$$

where $\mathbf{X}_{ij}$ and $\boldsymbol{\theta}_{ij}$ are as defined in Section 2, π_{ij} is the probability that cycle j from woman i falls in the parametric component, and $(1 - \pi_{ij})G$ is a nonparametric contamination, with G assigned an FDP prior. The FDP prior implies that G is discrete, with probability masses $\{p_h\}_{h=1}^{\infty}$ placed at functions $\{\Theta_h\}_{h=1}^{\infty}$. The probability masses can be expressed in stick-breaking form (Sethuraman, 1994) as $p_h = V_h \prod_{l<h}(1 - V_l)$, with $V_h \sim \mathcal{B}(1, \alpha)$ and $\mathcal{B}()$ denoting the beta density. In addition, the Θ_hs are generated independently of the weights from a Gaussian process having mean function μ and covariance function $\mathcal{C}$. The FDP has the same form as the dependent DP (DDP) proposed by MacEachern (1999). Refer to De Iorio et al. (2004) and Gelfand, Kottas, and MacEachern (2005) for other applications of the DDP.

Letting $\mathbf{w}_{ij} = (1, w_{ij1}, \ldots, w_{ijp})'$ denote a vector of woman- and cycle-specific predictors, we let $\pi_{ij} = \Phi(\mathbf{w}_{ij}' \boldsymbol{\beta})$, where $\Phi(\cdot)$ is the standard normal distribution function, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients. This formulation allows predictors to impact the probability a cycle is not well characterized by the parametric model, and hence is abnormal.

Although G is formally a probability measure with support on a functional space, one can conceptually view G as characterizing the distribution of curves through a discrete specification that places probability masses on different shapes, leading to functional clustering. In particular, if $\eta_{ij} \sim G$, then the prior probability of allocation of cycle i, j to functional cluster h is $\Pr(\eta_{ij} = \Theta_h) = p_h$. Variability among the curves is controlled by the weights $\{p_h\}_{h=1}^{\infty}$ and by the covariance $\mathcal{C}$. For example, when $p_1 + p_2 \approx 1$, almost all of the curves will have one of two shapes, with the deviations in these two shapes from μ and from each other being controlled by $\mathcal{C}$.

Samples from a Gaussian process (GP) can take a very wide variety of shapes that have limited sensitivity to the mean function. For this reason, by using Gaussian process priors for unknown functions, we allow an unknown, fixed mean, to avoid sensitivity to the scale of the temperature curves; note that this approach still allows a very wide variety of trajectory shapes. The covariance function $\mathcal{C}$ controls the types of shapes observed. As a simple default choice, we recommend the exponential covariance function, as it allows a wide variety of functional shapes, while squared exponential may overly favor smooth functions. More flexible choices, such as Matern, are characterized by more parameters that need to be estimated based on the data for the (often small) subset of functions placed in the nonparametric component. The concentration α is estimated by assuming a gamma hyperprior (Escobar and West, 1995).

4. Posterior Computation

Posterior computation can proceed via a straightforward modification of the Pólya urn Gibbs sampling algorithm widely used in DP mixture models (Bush and MacEachern, 1996). The algorithm alternates between (1) a data-augmentation step for allocating subjects to the parametric or nonparametric component; (2) standard Gibbs sampling steps for updating the parametric component model unknowns from their conditional distributions; and (3) Pólya urn Gibbs sampling steps for updating the nonparametric component model unknowns.

Before outlining the steps, we introduce some additional notation. For cycle j from woman i, let $S_{ij} = -1$ denote that the cycle belongs to the parametric component, and $S_{ij} = h$, for $h = 1, \ldots, k$, if the cycle is in cluster h of the nonparametric component, implying that $\eta_{ij} = \psi_h$, with $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)^T$ the k unique values of η_{ij} among those cycles with $S_{ij} \neq -1$. In addition, we use the (ij) superscript to denote a vector obtained excluding the observation from cycle i, j. For example, $\boldsymbol{\psi}^{(ij)}$ denotes the $k^{(ij)} \times 1$ vector of unique values of $\eta_{i'j'}$ for all cycles i', j' excluding cycle i, j. The steps are outlined below.

Step 0: Initialization. The algorithm starts with all subjects in the parametric component, so that $k = 0$. The parameters k_{ij} and r_{ij} will be initialized to a fixed number for all women and cycles, while the other parameters of the parametric component as well as the common parameters (σ^2 and the parameters of the covariance function) are initialized by sampling from the prior distributions.

Step 1: Update cluster indicators. For each cycle of every woman, sample from the multinomial posterior distribution of cluster indicators $S_{ij}^{(ij)}$ with probabilities

$$\Pr\left(S_{ij}^{(ij)} = h \middle| \ldots\right) = q_{ij,h}, \quad h = -1, 0, 1, \ldots, k^{(ij)},$$

where $S_{ij}^{(ij)} = 0$ denotes that the cycle belongs to a new cluster in the nonparametric component, and the $q_{ij,h}$ are obtained as follows:

(i) Considering that the parametric model may be written as a hierarchical linear model (see Section 2), we obtain

$$q_{ij,-1} \propto c\pi_{ij} \int N_{n_{ij}}\left(\mathbf{y}_{ij}; \mathbf{X}_{ij}\boldsymbol{\theta}_{ij}, \sigma^2\mathbf{I}_{n_{ij}}\right) \times N_{n_{ij}}(\boldsymbol{\theta}_{ij}; \boldsymbol{\alpha}_i, \boldsymbol{\Omega})\, d\boldsymbol{\theta}_{ij},$$

where $N_m(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.

(ii) For $h = 0$, we have

$$q_{ij,0} \propto c(1-\pi_{ij})\alpha \int N_{n_{ij}}\left(\mathbf{y}_{ij}; \boldsymbol{\mu}, \sigma^2\mathbf{I}_{n_{ij}}\right) \times N_{n_{ij}}(\boldsymbol{\mu}; \boldsymbol{\mu}_{0,ij}, \mathbf{C}_{n_{ij}})\, d\boldsymbol{\mu},$$

where $N_{n_{ij}}(\boldsymbol{\mu}; \boldsymbol{\mu}_{0,ij}, \mathbf{C}_{n_{ij}})$ is the multivariate normal distribution induced by evaluating $\mathrm{GP}(\mu, \mathcal{C})$ at the n_{ij} observation times for cycle i, j.

(iii) When $h > 0$ the procedure allocates the subject to the cluster indicated by h

$$q_{ij,h} \propto c(1-\pi_{ij}) n_h^{(ij)} N_{n_{ij}}\left(\mathbf{y}_{ij}; \psi_h^{(ij)}, \sigma^2\mathbf{I}_{n_{ij}}\right),$$

where $n_h^{(ij)}$ is the number of cycles already allocated to the group h before considering the jth cycle from subject i.

For any i and j such that $S_{ij}^{(ij)} = 0$, k is updated by $k = k + 1$ and a new cluster parameter ψ_k is drawn. The posterior distribution of the new group mean is

$$[\psi_k | \ldots] \propto N(\psi_k; \boldsymbol{\mu}_{0,ij}, \mathbf{C}_{n_{ij}}) N_{n_{ij}}\left(\mathbf{y}_{ij}; \psi_k, \sigma^2\mathbf{I}_{n_{ij}}\right).$$

Step 2: Update parametric model. Update the parameters of the parametric model using Gibbs sampler for hierarchical linear model. Only subjects drawn from the parametric model will be used. The algorithm used is reported in Web Supplementary Materials.

Step 3: Update nonparametric components. Update the cluster specific parameters using only subjects in these clusters. Define $\mathbf{t}_{ij}$ as the vector of observation times at which the data are collected for cycle j of subject i and let $\psi_h(\mathbf{t})$ denote the values of ψ_h at times $\mathbf{t}$. The likelihood for the subjects in cluster h of the nonparametric components is

$$\prod_{ij:S_{ij}=h} N_{n_{ij}}\left(\mathbf{y}_{ij}; \psi_h(\mathbf{t}_{ij}), \sigma^2 I_{n_{ij}}\right).$$

The posterior for the value of ψ_h evaluated at any finite collection of points $\mathbf{t}$ is proportional to this likelihood multiplied by the normal prior for $\psi_h(\mathbf{t})$ implied by the prior: $\mathrm{GP}(\psi_h; \mu, \mathcal{C})$. We update α following Escobar and West (1995), assuming a $\mathcal{G}(a_\alpha, b_\alpha)$ prior.

Step 4: Update the covariate effect parameters. To update π_{ij}, the probability, for a cycle, of falling in the parametric component, a data-augmentation step is proposed. Let $z_{ij} \sim N(\mathbf{w}_{ij}^T\ \boldsymbol{\beta},\ 1)$ denote an underlying normal variable with $S_{ij} = -1$ if $z_{ij} < 0$ and $S_{ij} > 0$ if $z_{ij} > 0$.

Sample the z_{ij}'s from truncated normal conditional posteriors given S_{ij} and then once the z_{ij}'s are imputed the conditional posterior of $\boldsymbol{\beta}$ are multivariate normal given a multivariate normal prior.

Step 5: Update overall parameters. The variance σ^2 of the noise component $\varepsilon_{ij}(t)$ is common to the parametric and nonparametric components and will be updated from its conditionally conjugate inverse-gamma posterior. We include Metropolis–Hastings steps for updating the covariance parameters.

5. Simulation Example

To illustrate the method, we simulated data from 20 women each with 10 cycles. For each cycle, we generated 25 values corresponding to daily temperatures on different days after the end of menses. We selected the numerical specification of the simulation by choosing typical values described in the medical literature (e.g., Vincent, 1964; Marshall, 1979; see also Section 6.2). Of the 200 cycles, 15 were selected as abnormal cycles, with 12 following one of two systematic patterns (seven type 1, and five type 2) with autocorrelated residuals and the remaining three consisting of uncorrelated Gaussian observations with a single spike at a random time between days 13 and 20.

The remainder of the cycles had data generated under the parametric model of Section 2.2. In particular, we let $\boldsymbol{\alpha} = (36.5, 0.4)\prime$, $\omega_{11} = \omega_{12} = 0.01$ and $\omega_1 = \omega_2 = 0.01$ in generating the cycle-specific follicular phase baseline bbt and rate of increase at ovulation. In addition, we generated $k_{ij} \sim \mathcal{U}(1, 14)$ and $r_{ij} \sim \mathcal{U}(1, 5)$ to allow variability in the last day of hypothermia and number of days between the high and low plateaus, and we generated the measurement error for a Normal distribution with mean 0 and $\sigma^2 = 0.1^2$.

For simplicity, in the simulation, we consider the probability of falling in the parametric component, π, as fixed for all women and cycles, but allow uncertainty in this fixed value by letting $\pi \sim \mathcal{B}(a, b)$, where a and b are prespecified hyperparameters.

To specify the hyperparameters, we purposely move the prior distributions from the true parameters to allow for some errors in elicitation and we consider somewhat vague priors,

$$\boldsymbol{\alpha} \sim N_2\left(\begin{bmatrix}36\\1\end{bmatrix}, \begin{bmatrix}0.5^2 & 0\\0 & 0.1^2\end{bmatrix}\right) \quad \sigma^{-2} \sim \mathcal{G}(1,1)$$

$$\omega_h^{-1} \sim \mathcal{G}(1,1) \qquad \omega_{1h}^{-1} \sim \mathcal{G}(1,1) \qquad \pi \sim \mathcal{B}(1,5) \qquad \alpha \sim \mathcal{G}(0.1,1).$$

As motivated in Section 3.3, we choose $\mathcal{C}$ to correspond to the exponential covariance, with $\mathcal{G}(1,1)$ hyperpriors for the two unknown parameters.

We ran the algorithm for 40,000 iterations after an 11,000 iteration burn-in. Examination of traceplots of all parameters showed no evidence against convergence. Figure 2 provides the true and estimated trajectories in each of the abnormal groups. The estimated posterior probability of allocating a normal curve to the abnormal component was 0.00 for all of the simulated curves, while the average posterior probability of allocation to the nonparametric component for an abnormal

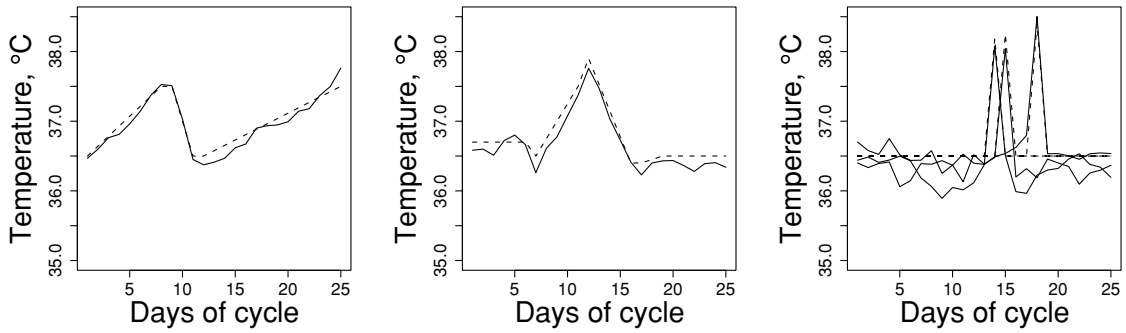

Figure 2. Simulated data. True and estimated trajectories for each of the types of abnormal cycles.

curve was 0.82. In addition, the approach did a good job at estimating and clustering of the abnormal curves.

To assess the performance in estimation of the parameters in the parametric component model, in Table 1 we compare true and estimated summaries of the distribution of various cycle- and woman-specific parameters in the hierarchical model, including k_{ij}, r_{ij}, θ_{1ij}, θ_{2ij}, α_{i1}, and α_{i2}. Values for cycle-specific parameters are summaries across all the cycles from all the women, while values for woman-specific parameters are summaries across all women. As you can see, the distribution of each of these coefficients was well estimated.

6. European Fertility Study

6.1 *Data and Model Specification*

We analyzed data on daily observations of bbt for 1121 nonconception cycles from 157 women in the Verona center of the Colombo and Masarotto (2000) study, excluding cycles that resulted in a conception or had <7 days of bbt data. The participants were followed prospectively during one or more menstrual cycles, as they recorded daily bbt and the days during which intercourse and menstrual bleeding occurred. The average age in years of the women in the Verona center was 28.6 (standard deviation, 3.54), and women contributed an average of 7.14 (6.62) cycles to the data set.

Our focus was on characterizing interesting features of the bbt curves for normal cycles, including the average temperature level of the follicular phase, the day of ovulation, suggested by the last day of the low plateau, the number of days between the low and high temperature plateaus, and the magnitude of the shift in temperature between the two phases of the cycle. We are also interested in identifying and studying heterogeneity in abnormal trajectories, which do not follow the characteristic biphasic pattern. We anticipate that less than 5% of the cycles will be allocated to this group, because the women in the study are of reproductive age and have no history of infertility or reproductive problems. Woman's age and number of prior pregnancies are used as covariates to help predict whether a woman had one or more irregular cycles.

Informative prior distributions are chosen for each of the parameters, except for the covariate parameters. For each of them we choose as prior a weakly informative Normal distribution. Marshall (1979) and Vincent (1964) observed that the temperature during the first phase of the cycle is around 36.5°C, while the total increment in temperature between the low and the high plateaus has been reported around 0.3°C with some variability around these values. Therefore, we set $\boldsymbol{\alpha} \sim N([36.5,\ 0.3]^T,\ \text{diag}\ (0.2,\ 0.1))$ as the hyperprior for the overall mean of the temperature during the first phase and for the increment, with the variance selected to allow a plausible amount of uncertainty. Based on descriptive statistics for data from a different center, we chose the hyperpriors $\omega_h^{-1} \sim \mathcal{G}(0.1, 0.005)$ and $\omega_{1h}^{-1} \sim \mathcal{G}(0.1, 0.03)$. Based on expert elicitation, we chose $\sigma^{-2} \sim \mathcal{G}(0.1, 0.02)$, set the GP mean to $\mu(t) = 36.5$ and chose an exponential correlation function, with $\mathcal{G}(1, 1)$ on the two parameters.

6.2 *Analysis and Results*

We ran the algorithm for 5,500 iterations after a 500 iteration burn-in period. Convergence was assessed by examining the stationarity of the distribution of the parameters for the parametric component and for the mixture weights π_{ij}, the overall variance σ^2, and the means of the groups for the nonparametric component. Traceplots of these parameters provided no evidence against convergence.

Table 1
Simulated data. Posterior summaries of the cycle- and woman-specific parameters.

	True distribution					Estimates				
Parameter	Mean	Standard deviation	First quartile	Median	Third quartile	Mean	Standard deviation	First quartile	Median	Third quartile
k_{ij}	7.58	3.70	5.00	8.00	11.00	8.84	4.00	5.43	8.94	11.93
r_{ij}	8.00	4.34	4.00	8.00	11.00	7.98	2.63	6.18	8.38	10.12
θ_{ij1}	36.53	0.15	36.42	36.54	36.65	36.56	0.11	36.43	36.55	36.67
θ_{ij2}	0.40	0.13	0.33	0.40	0.49	0.40	0.18	0.30	0.41	0.51
α_{i1}	36.52	0.13	36.44	36.55	36.63	36.56	0.11	36.49	36.56	36.63
α_{i2}	0.41	0.07	0.36	0.40	0.45	0.39	0.11	0.31	0.40	0.49

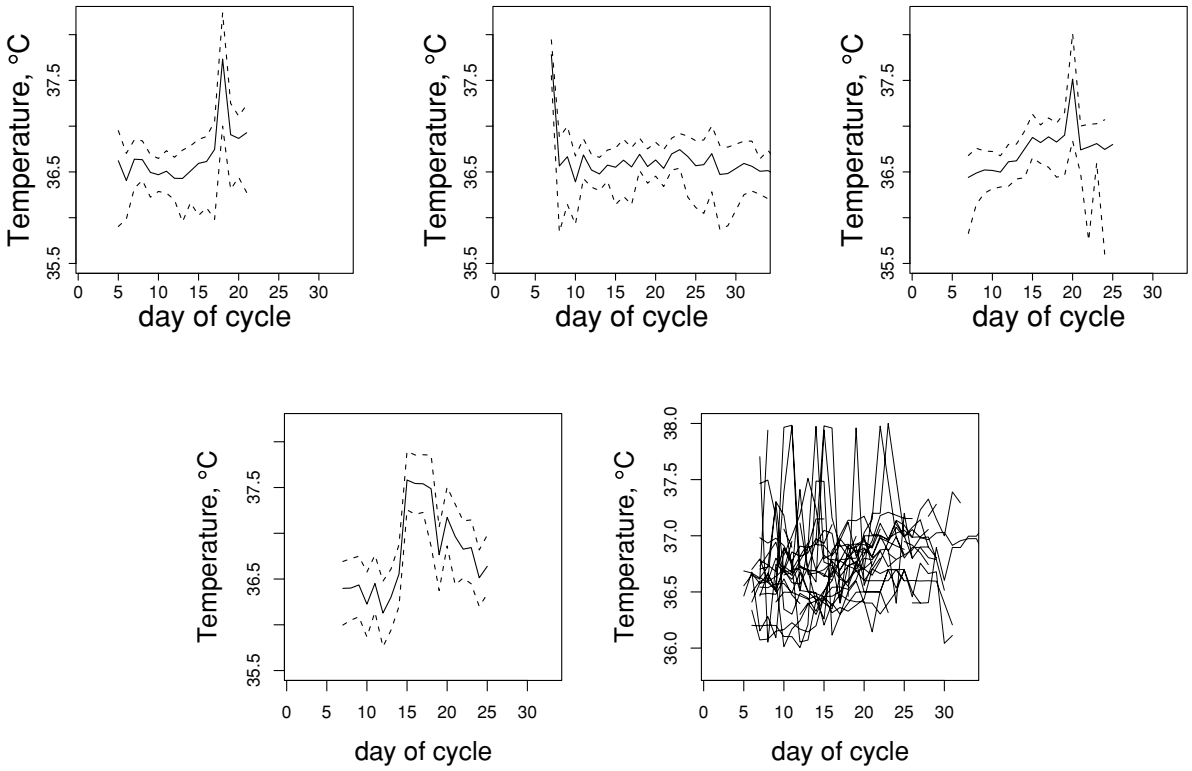

Figure 3. Fertility study. Cycles falling in the nonparametric component with at least 70% posterior probability. Solid lines are posterior means and dashed lines are 95% pointwise credible intervals. The first four panels are clusters of two cycles having clustering probabilities of 0.86, 0.53, 0.50, and 0.48, respectively, while the fifth panel shows singleton clusters.

The posterior probability of allocation to the parametric component was greater than 50% for 94.11% of the menstrual cycles under study, and there were 36 cycles from 26 women that had a posterior probability higher than 70% of being allocated to the nonparametric component. These cycles do not fit the biphasic pattern and are potentially abnormal. Web Table 1 in Supplementary Materials summarizes some information about these 36 cycles. We identified four pairs of cycles having a nonnegligible probability of being clustered together (0.86, 0.53, 0.50, and 0.48). Figure 3 shows the cluster-specific estimated trajectories for all cycles with at least 70% posterior probability of being allocated to the nonparametric component.

Figure 4 provides histograms showing the variability among cycles in different features of the bbt trajectories, including last day of hypothermia (i.e., the last day of the low temperature plateau), the number of days between the plateaus, the degree Celsius in the rise, and the level of the low plateau. It is clear from this figure that there is substantial variability among cycles in the last day of hypothermia, which is consistent with our prior knowledge of high variability in follicular phase lengths. Interestingly, there was also substantial variability in the duration of the rise between the low and high plateaus, with many women having a rapid rise in 3–4 days, while others take 10–11 days. The estimated mean temperature in the first region shows more moderate variability, with a mean value of 36.53 and 91% of the values falling with 36.2°C and 36.8°C. In addition, the mean temperature shift was 0.4°C, with 99.6% of the values between 0.1 and 0.7.

To assess whether the variability observed in these features is primarily due to variability across cycles within a woman or across women, we examined the estimated variance components. The posterior mean of the variance within a woman for the temperature level of the low plateau, ω_1, is 0.0078 with 95% credibility interval (0.0070, 0.0087), while the corresponding posterior mean for the variance across women, ω_{11}, is 0.037 with 95% credibility interval (0.030, 0.047). As expected, variability of the temperature level of the low plateau between women is much higher than that within women, with a variance of about five times higher. We observe, also, a big difference in the variability between women (ω_{12}) and within women (ω_2) for the increase in temperature between low and high plateau: the variance between women (posterior mean 0.020, 95% CI: 0.016–0.026) is twice as large as the one within women (posterior mean 0.0094, 95% CI: 0.0083–0.0107), with about the same standard deviation.

Our results show no evidence that the probability of having an abnormal cycle depends on age (β_1, mean = −0.04, sd = 0.31) or parity (β_2, mean = −0.11, sd=0.34). Web Figure 1 in Supplementary Material provides histograms of posterior samples of β_0, β_1, and β_2, the parameters characterizing covariate effects on the probability of allocation to the nonparametric component.

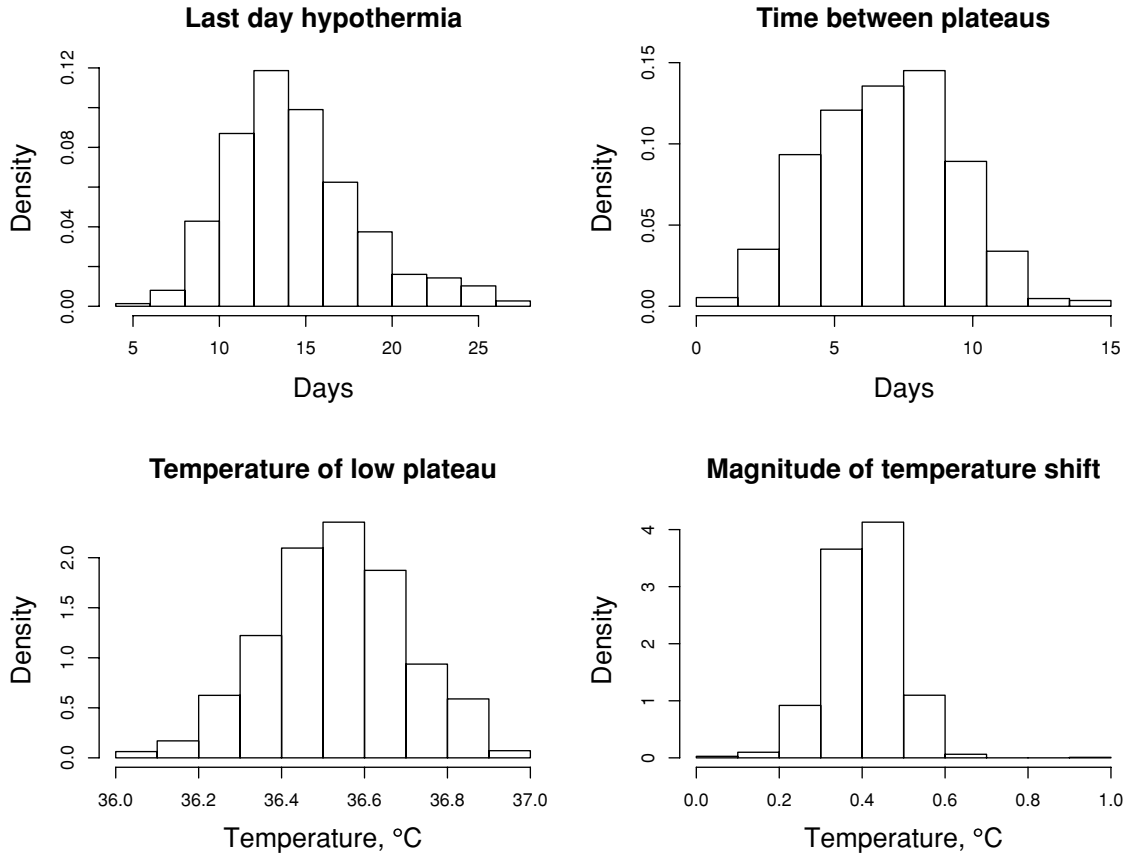

Figure 4. Fertility study. Distributions among all the cycles of the cycle-specific posterior means for the last day hypothermia k, the magnitude of temperature shift θ_2, the time between plateaus r, and the temperature of the low plateau θ_1.

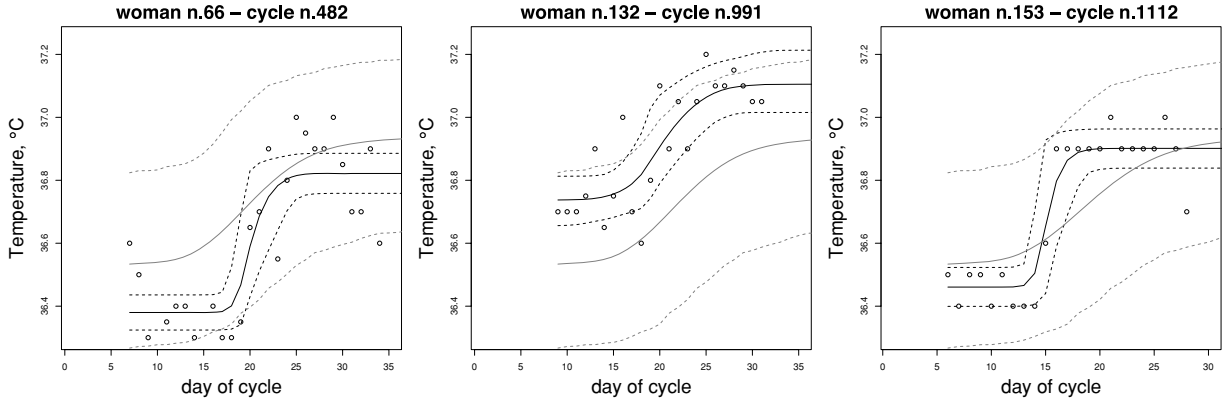

Figure 5. Fertility study. Mean (solid black lines) and 90% credible intervals (black dashed lines) of the estimated trajectories of three typical cycles along with overall mean (gray solid lines) and 90% credible intervals (gray dashed lines) of the trajectories.

Figure 5 shows the estimated trajectories and raw data for three cycles randomly chosen from among the curves falling in the parametric component. Mean, median, and 90% credible interval are also plotted in the same figure. This figure is illustrative of our general observation that the biphasic parametric model provided a good fit to the overwhelming majority of curves allocated to the parametric component model.

We assessed sensitivity to the prior specification by repeating the analysis using a default approach in which we

standardized the temperature values prior to analysis and then chose hyperpriors to have variance one, with mean zero for location parameters and mean one for scale parameters. The results were robust, though MCMC mixing was slower and a somewhat greater proportion of cycles were allocated to the nonparametric component (4.1% instead of 1.2%). We also repeated the analysis including woman's age and number of previous pregnancies as covariates impacting the bbt level in the parametric component model. However, there was no evidence that bbt trajectories vary with age or parity based on our results. The other results presented above did not change with the covariate adjustment.

7. Discussion

This article proposes an approach for including prior information in functional data analyses, with this prior information taking the form of a parametric hierarchical model that is believed to approximately characterize regularly behaved functions obtained for healthy subjects. To allow such a parametric model to provide a poor characterization of a minority of subjects, possibly having a latent adverse health condition, we propose to contaminate the parametric model with a nonparametric component, with this component characterized as a FDP. This structure allows one to obtain more robust inferences on the parametric component through effectively discarding outliers. In addition, one can perform a sort of unsupervised learning in which unusual subjects are automatically identified as those in the nonparametric component. Finally, one can estimate functional clusters within the nonparametric component, allowing one to do inferences on the commonalities in the abnormal curves.

In the application to bbt trajectories in menstrual cycles, we found that only a few percent of the cycles were allocated to the nonparametric component, with the remaining cycles adequately characterized using a very simple parametric hierarchical model. The proposed method may be very useful in clinical and diagnostic practice. In particular, the model fitted using the Verona data may be used to obtain a fast approach to classify a new cycle as abnormal. For diagnostic purposes, it would be possible to produce a simple program or device that outputs the classification when the temperatures are input. As a measure of uncertainty, the posterior probabilities could also be output in the automated diagnosis. Such an approach could also be used for other biomarkers, such as progesterone levels.

8. Supplementary Materials

The Web Table and Figure referenced in Section 6.2 and the algorithm used for estimate of the parametric component are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgement

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The authors thank Bernardo Colombo for his helpful comments and for generously providing the data.

References

Baladandayuthapani, V., Mallick, B. K., Hong, M. W., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* **64,** 64–73.

Bigelow, J. and Dunson, D. B. (2005). Posterior simulation across nonparametric models for functional clustering. *Discussion Paper*, 2005-18, Department of Statistical Science, Duke University, Durham, North Carolina.

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

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 semi-parametric Bayesian model for randomized block designs. *Biometrika* **83,** 275–286.

Carter, R. L. and Blight, B. J. N. (1981). A Bayesian change-point problem with an application to the prediction and detection of ovulation in women. *Biometrics* **37,** 743–751.

Colombo, B. and Masarotto, G. (2000). Daily fecundability: First results from a new data base. *Demographic Research* **3**, 5.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An Anova model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

Dunlop, A. L., Schultz, R., and Frank, E. (2005). Interpretation of the BBT chart: Using the "gap" technique compared to the coverline technique. *Contraception* **71,** 188–192.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Ferguson, T. S. (1973). Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1,** 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2,** 615–629.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100,** 1021–1035.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **101,** 18–29.

MacEachern, S. N. (1999). Dependent nonparametric processes, in *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, Virginia: American Statistical Association.

MacLehose, R. F. and Dunson, D. B. (2008). Nonparametric Bayes kernel-base priors for functional data analysis. *Statistica Sinica*, in press.

Marshall, J. (1979). *Planning for a Family. An Atlas of Mucothermic Charts*, 2nd edition. London: Faber and Faber.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68,** 179–199.

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association* **98,** 573–583.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society, Series B* **68,** 305–332.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, in press.

Royston, J. P. and Abrams, R. M. (1980). An objective method for detecting the shift in basal body temperature in women. *Biometrics* **36,** 217–224.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4,** 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101,** 1566–1581.

Thompson, W. K. and Rosen, O. (2007). A Bayesian model for sparse functional data. *Biometrics* **64,** 54–63.

Vincent B. (1964). *Atlas de Courbes Thermiques*, 4th edition. Nantes: Centre de Documentation et d'Information Conjugale.

Received August 2007. *Revised June* 2008.
Accepted June 2008.