Published in final edited form as:

J Am Stat Assoc. 2014 ; 109(506): 647–660. doi:10.1080/01621459.2013.866564.

Enriched Stick Breaking Processes for Functional Data

Bruno Scarpa¹ and David B. Dunson²

Bruno Scarpa: scarpa@stat.unipd.it

¹Department of Statistical Sciences, University of Padua, Padua, Italy

²Department of Statistical Sciences, Duke University, Durham, NC, USA

Abstract

In many applications involving functional data, prior information is available about the proportion of curves having different attributes. It is not straightforward to include such information in existing procedures for functional data analysis. Generalizing the functional Dirichlet process (FDP), we propose a class of stick-breaking priors for distributions of functions. These priors incorporate functional atoms drawn from constrained stochastic processes. The stick-breaking weights are specified to allow user-specified prior probabilities for curve attributes, with hyperpriors accommodating uncertainty. Compared with the FDP, the random distribution is enriched for curves having attributes known to be common. Theoretical properties are considered, methods are developed for posterior computation, and the approach is illustrated using data on temperature curves in menstrual cycles.

Keywords

Functional data analysis; Functional Dirichlet process; Nonparametric Bayes; Prior elicitation; Random curves; Random probability measure

1. Introduction

Our interest focuses on studying variability in random functions. In particular, letting f_i denote the function for subject i , $i = 1, \dots, n$, we want to estimate $f_1 \dots f_n$, cluster subjects having $f_i = f_j$, and estimate the predictive distribution of f_{n+1} for a new subject $i = n+1$. In considering methodology, a focus is on avoiding parametric assumptions about the functions and heterogeneity structure, while allowing for the incorporation of prior information. We consider prior information taking the form of frequencies of occurrence of different features of the curves.

As a motivating application, consider the study of basal body temperature (BBT) curves within menstrual cycles from different women. As illustrated in Figure 1, for healthy women in their reproductive years, temperature curves typically have a characteristic *bi-phasic* shape. In particular, during the follicular phase at the start of the menstrual cycle, there is an interval of low temperature. After ovulation, temperature then rises to a higher threshold during the luteal phase of the cycle. A bi-phasic temperature curve has been used as a clinical indicator of ovulation (Martinez et al., 1992), while the rise in temperature at the time of ovulation marks the end of the fertile interval. To better understand reproductive

functioning and causes of infertility, it is interesting to estimate a distribution of temperature curves among women.

Potentially, one could view the temperature values on the different days across the menstrual cycle as having a multivariate normal distribution, with both the mean and covariance estimated empirically (Ramsay and Silverman, 1997). However, this approach does not allow one to separate measurement errors in temperature from variability in the underlying smooth trajectories (Behseta et al., 2005). An alternative approach is to characterize the temperature measurements as error-prone realizations of functions expressed in terms of linear combinations of basis functions. By assuming normally distributed basis coefficients, one can allow variability in the functions. Related approaches have been considered by Rice and Wu (2000), James et al. (2000), Ke and Wang (2001) and James (2002). Bigelow and Dunson (2007) and Thompson and Rosen (2007) developed Bayes approaches using adaptive splines. An alternative hierarchical Gaussian process model was proposed by Behseta et al. (2005).

These approaches have the disadvantage of using a parametric distribution, typically normal, to characterize variability among the functions. A goal in modeling of temperature curves is to cluster subjects and identify outlying subjects with reproductive dysfunction, so such assumptions are not ideal. As a more flexible approach, Bigelow and Dunson (2009) proposed placing a Dirichlet process (DP) prior (Ferguson, 1973, 1974) on the distribution of the basis coefficients, following previous authors who used DP priors for random effects distributions (Bush and MacEachern, 1996; Mukhopadhyay and Gelfand, 1997; Kleinman and Ibrahim, 1998; among others).

An alternative to nonparametric modeling of the distribution of basis coefficients is to directly define a random probability measure with support on a function space. For example, letting $f_i \sim P$, with P a random probability measure, we could let $P \sim DP(\alpha G_0)$, where α is a precision parameter and G_0 is a probability measure obeying a Gaussian process (GP) law. Following the Sethuraman (1994) stick-breaking representation, we then obtain:

$$P = \sum_{h=1}^{\infty} p_h \delta_{\Theta_h}, p_h = V_h \prod_{l < h} (1 - V_l), V_h \sim \text{beta}(1, \alpha), \Theta_h \sim GP(\mu, \mathcal{K}), \quad (1)$$

where Θ_h is a functional atom generated from a Gaussian process, with mean μ and covariance function $\mathcal{K}$, and p_h is the probability mass assigned to that atom. Hence, prior (1) has an almost surely discrete form expressed as an infinite sequence of random functions that are assigned random probabilities.

Dunson et al. (2008) used (1) as a functional DP (FDP) prior for the distribution of a functional predictor modeled jointly with a response. Rodriguez et al. (2009) instead used the nested Dirichlet process (Rodriguez et al., 2008) to model the joint densities of the predictors and response for each subject, inducing DP functional clustering within a kernel model. MacLehose and Dunson (2008) proposed an alternative kernel-based approach based on a hierarchical generalization of priors for single functions proposed by Pillai et al. (2007) and Liang et al. (2006). Nguyen and Gelfand (2011) developed a Dirichlet labeling process

by assuming that the observation at a given location can be allocated to separate groups via a random allocation process.

In considering such methods, René Ecochard (personal communication) noted that they are very flexible, but seem to rely too much on data from the current study, while ignoring accumulated knowledge from past studies. Specifically, in the temperature curve application, it is known that the majority of cycles have an approximately biphasic-type temperature trajectory, with the shift in temperature typically occurring between days 10 and 17 of the cycle. Scarpa and Dunson (2009) incorporated this prior information through the use of a parametric model for the trajectory, with a nonparametric FDP contamination allowing for a subset of cycles that do not follow the anticipated pattern.

This approach only allows for the incorporation of prior knowledge through a parametric model for the functions. In many cases, one can make only a weak guess at the exact functional form *a priori*, but abundant information may be available about features of the trajectories. In the temperature curve application, such features include the occurrence of a bi-phasic shape and shift occurrence within a mid-cycle window. One can easily elicit prior probabilities for the proportions of curves having these features. To allow for the incorporation of such information in a flexible Bayesian nonparametric model, we propose a class of enriched stick-breaking priors (ESTB) for functional data, generalizing the FDP.

Section 2 introduces the structure of the model we will use, and section 3 proposes the ESTB structure, providing motivation and consideration of properties. Section 4 develops methods for posterior computation. Section 5 presents a simulated data example. Section 6 applies the method to a large data set of menstrual cycle temperature curves, and Section 7 discusses the results.

2. Functional Stick-Breaking Priors

2.1 Data and Modeling Structure

For subject i ($i = 1, \dots, n$), suppose that error-prone realizations of the function f_i are collected at times $\mathbf{t}_i = (t_{i1}, \dots, t_{i,n_i})^T$, with $t_{ij} \in \mathbf{t} = (t_1, \dots, t_J)$ for all i, j , and $\mathbf{t}$ denoting the global collection of observation times. We assume the following modeling structure:

$$\begin{aligned} y_i(t) &= f_i(t) + \varepsilon_i(t), \varepsilon_i(t) \sim t_\nu(\sigma^2) \\ f_i | P &\stackrel{iid}{\sim} P \\ P &\sim \mathcal{P} \end{aligned} \quad (2)$$

where f_i is a random function from $\mathcal{R}^+ \rightarrow \mathcal{R}$ for subject i generated according to the probability measure P , and $\varepsilon_i(t)$ is a local measurement error specific to time t . For example, f_i can be considered the true basal body temperature curve, while $\varepsilon_i(t)$ is the error in the measurement of temperature at time t of the menstrual cycle. We assume that $\varepsilon_i(t)$ is distributed as a Student- t with ν degrees of freedom and scale parameter σ , allowing for heavy tailed distributions of residuals.

Formally, f_i can be seen as a random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, where Ω is a function space, $\mathcal{B}$ is a σ -algebra of subsets of Ω , and P is a probability measure.

To allow uncertainty, we further treat P as a random variable defined on a probability space $(\Gamma, \mathcal{A}, \mathcal{P})$, with Γ the space of probability measures over $(\Omega, \mathcal{B})$, $\mathcal{A}$ a σ -algebra of subsets of Γ , and $\mathcal{P}$ a probability measure. Our focus is on the specification of $\mathcal{P}$.

2.2 Functional Stick-Breaking Process

In constructing priors for random probability measures, a useful starting point is a stick-breaking formulation. Ishwaran and James (2001) developed methods for posterior computation under a broad class of stick-breaking priors, while other authors rely on stick-breaking in constructing priors for dependent collections of random measures (Griffin and Steel, 2006; Duan et al., 2007).

Following Ishwaran and James (2001), one possible specification for $P_N \sim \mathcal{P}_N$ is

$$P_N = \sum_{h=1}^N p_h \delta_{\Theta_h}, \Theta_h \stackrel{iid}{\sim} P_0, \quad (3)$$

where $p_h = V_h \prod_{l < h} \bar{V}_l$, $V_h \stackrel{iid}{\sim} \text{beta}(a_h, b_h)$, for $h = 1, \dots, N-1$, $V_N = 1$, $\bar{V}_h = 1 - V_h$ and Θ_h is a function generated according to probability measure P_0 on $(\Omega, \mathcal{B})$. Hence, P_N is constructed by generating a sequence of functional atoms, $\{\Theta_h\}_{h=1}^N$, and assigning these atoms weights, $\{p_h\}_{h=1}^N$, which are generated according to a stick-breaking process. The N subscript on P_N denotes the number of functional atoms in the stick-breaking characterization.

The special case of (3) in which $N \rightarrow \infty$, $a_h = 1$ and $b_h = \alpha$, for $h = 1, \dots, \infty$, corresponds to a functional Dirichlet process (FDP). The relationship with the Dirichlet process (DP) is clear if we let $\mathcal{P}_\infty \equiv DP(\alpha P_0)$, with $DP(\alpha P_0)$ denoting the Dirichlet process with precision parameter α and base measure P_0 on $(\Omega, \mathcal{B})$. Then (3) is obtained (under the above conditions) using the Sethuraman (1994) representation of the DP. The FDP was applied by Dunson et al. (2008), with P_0 chosen to correspond to a Gaussian process. Petrone, Guindani and Gelfand (2009) proposed a more intricate hybrid FDP, which also draws functional atoms from a Gaussian process, while allowing realizations f_i to select different atoms in different local regions. Nguyen and Gelfand (2011) study properties of this approach, while developing improved inferential methods.

3. Enriched Functional Stick-Breaking Priors

3.1 Proposed Prior

Our primary focus is on adapting (3) to incorporate prior information on the frequencies of occurrence of different curve features. Let $\mathcal{F}: \Omega \rightarrow \mathcal{A} = \{0, 1\}^M$ denote a feature class indicator function, with $\mathcal{F}(f) = [\mathcal{F}_1(f), \dots, \mathcal{F}_M(f)]^T$ for all $f \in \Omega$, where $\mathcal{F}_m(f) = 1$ denotes that function f possesses feature m and $\mathcal{F}_m(f) = 0$ otherwise, for $m = 1, \dots, M$. There are $\#\mathcal{A} = 2^M$ elements in set $\mathcal{A}$ and choosing an arbitrary ordering for these elements, we let $\mathcal{A}_a$ denote the a th element of $\mathcal{A}$, for $a = 1, \dots, \#\mathcal{A}$. The function space Ω can be expressed as $\Omega = \bigcup_{a=1}^{\#\mathcal{A}} \Omega_a$, where $\Omega_a = \{f: \mathcal{F}(f) = \mathcal{A}_a, f \in \Omega\}$.

We focus on the case in which a finite collection of M feature classes are elicited in advance in consultation with subject matter experts. In practice, functional data are routinely reduced to a small set of features to simplify interpretation, statistical analyses and inferences. For example, clinicians categorize patients based on the average level, the change between the low and high point, the location of the peak, the average slope, and so on. Through routinely collecting such observations, a rich prior data base is obtained. Prior knowledge on the overall function shape is more elusive to quantify. A fundamental advantage of our approach is the ability to characterize the population distribution of the functional data, while maintaining parameters characterizing proportions falling into different feature classes.

For illustration consider the example in which $f_i : [0, 1] \rightarrow \mathbb{R}$, $M = 2$, $\mathcal{J}_1(f) = 1$ if f is monotone and 0 otherwise, and $\mathcal{J}_2(f) = 1$ if $\text{range}(f) \subset [0, 1]$ and 0 otherwise. Hence, there are two binary feature classes corresponding to whether or not the function is monotone and whether or not the range of the function is contained in $[0, 1]$. These binary classes partition the function space Ω into $2^M = 4$ mutually exclusive subsets, $\Omega = \bigcup_{a=1}^4 \Omega_a$, with Ω_1 containing monotone functions with $\text{range}(f) \subset [0, 1]$, Ω_2 containing monotone functions with $\text{range}(f) \not\subset [0, 1]$, Ω_3 containing non-monotone functions with $\text{range}(f) \subset [0, 1]$, and Ω_4 containing non-monotone functions with $\text{range}(f) \not\subset [0, 1]$.

As in expression (2), let P denote a probability measure characterizing the distribution of functional data $\{f_i\}$, with $\mathcal{P}$ a prior for this probability measure. Then,

$$\Pi_a = \Pr(f_i \in \Omega_a | P) = \int_{\Omega} 1\{\mathcal{J}(f) = \mathcal{A}_a\} P(df)$$

is the proportion of curves falling in feature class a under P . Modifying (3) in the $N \rightarrow \infty$ case, we let

$$P = \sum_{a=1}^{\#\mathcal{A}} \Pi_a \sum_{h \in \mathcal{C}_a} \left\{ V_h \prod_{l \in \mathcal{C}_a(\hat{\alpha}\hat{\alpha}\hat{\alpha})} \bar{V}_l \right\} \delta_{\Theta_h}, = \sum_{a=1}^{\#\mathcal{A}} \Pi_a \sum_{h \in \mathcal{C}_a} p_{a,h} \delta_{\Theta_h}, \quad (4)$$

where $V_h \stackrel{iid}{\sim} \text{beta}(a_h, b_h)$ and $\Theta_h \stackrel{iid}{\sim} P_0$, for $h = 1, \dots, \infty$, $\mathcal{C}_a = \{h : \mathcal{J}(\Theta_h) = \mathcal{A}_a, h = 1, \dots, \infty\}$, $\mathcal{C}_a(h)$ denotes the subset of elements of $\mathcal{C}_a$ having values less than h , and $p_{a,h}$ is the prior probability allocated to element h of the subset of functional atoms in the a th feature class.

To allow uncertainty in the feature probabilities, we let

$$\Pi = (\Pi_a, a = 1, \dots, \#\mathcal{A})^T \sim \text{Diri}(a_{\Pi,1}, \dots, a_{\Pi,\#\mathcal{A}}).$$

Each atom Θ_h in expression (4) belongs to a single feature class, and hence we obtain atoms representative of the different pre-specified feature classes. This helps greatly in interpretation. For example, in the BBT application, we obtain representative BBT trajectories in each of the three feature classes. Clinicians can examine these representative trajectories to obtain insight into the variation among women in BBT trajectories that are biphasic with a mid-cycle shift, biphasic with a shift outside the specified mid-cycle interval,

and lacking a biphasic pattern. This provides new insight into natural variability among women in these reproductive biomarkers across the cycle. Potentially we could attempt to model dependence in the different Θ_h s within and across feature classes, but such intricate hierarchical dependence may be difficult to infer from the data.

We refer to (4) as an enriched functional stick-breaking process (E-FSTB), as it allows enrichment of functions in feature classes less likely to be sampled from P_0 . We use the notation $P \sim \text{E-FSTB}(\Pi, \mathbf{a}, \mathbf{b}, P_0)$ to denote that P is assigned an E-FSTB prior. It is useful to define an alternative formulation of the E-FSTB, which simplifies the structure through relabeling of the atoms. Let $P_{0,a}$ denote a probability measure

$$P_{0,a}(B) = \frac{P_0(B)}{P_0(\Omega_a)}, \text{ for all } B \in \mathcal{B}_a, \quad (5)$$

for $a = 1, \dots, \#\mathcal{A}$, where $\mathcal{B}_a$ is a σ -algebra of subsets of Ω_a . Relabeling the components, we can express (4) as

$$P = \sum_{a=1}^{\#\mathcal{A}} \Pi_a \sum_{h=1}^{\infty} \left\{ V_{a,h} \prod_{l < h} \bar{V}_{a,l} \right\} \delta_{\Theta_{a,h}}, V_{a,h} \stackrel{iid}{\sim} \text{beta}(a_h, b_h), \Theta_{a,h} \stackrel{iid}{\sim} P_{0,a}, = \sum_{a=1}^{\#\mathcal{A}} \Pi_a P_a, P_a \sim \text{FSTB}(\mathbf{a}, \mathbf{b}, P_{0,a}). \quad (6)$$

Here, the notation $P_a \sim \text{FSTB}(\mathbf{a}, \mathbf{b}, P_{0,a})$ denotes that the random probability measure P_a is drawn from an FSTB with hyperparameters $\mathbf{a}, \mathbf{b}$ and base measure $P_{0,a}$.

3.2 An Illustrative Example

Returning to the illustrative example, suppose Θ_h is generated from P_0 by first generating $\beta_h = (\beta_{h1}, \dots, \beta_{hL})^T \in \mathcal{R}^L$ from G , and then letting $\Theta_h(t) = \sum_{l=1}^L \beta_{hl} b_l(t)$, with $\{b_l, l = 1, \dots, L\}$ B-spline basis functions having L equally-spaced knots in $[0, 1]$, $b_1(0) = 1$ and $b_L(1) = 1$. For each feature class a , there is a corresponding set Γ_a such that $\beta_h \in \Gamma_a$ implies that $\Theta_h \in \Omega_a$. Monotone Θ_h is obtained for $\beta_{h1} < \beta_{h2} < \dots < \beta_{hL}$, with the further restriction $0 \leq \beta_{h1}$ and $\beta_{hL} \leq 1$ ensuring $\text{range}(\Theta_h) \subset [0, 1]$. Although there is no explicit dependence included between the atoms in closely-related feature classes, the similar restrictions defined for these classes will automatically favor similarities *a priori*. For example, suppose that $f_i \in \Omega_1$ and $f_{i'} \in \Omega_2$, so that they share the monotonicity feature but have different range restrictions. As illustrated in Figure 2, the curves $\{\Theta_h, h \in \mathcal{C}_1\}$ and $\{\Theta_h, h \in \mathcal{C}_2\}$ in the closely related feature classes 1 and 2 are much more similar than the curves in the very different feature classes 1 and 4.

The prior enables global borrowing of information across all feature classes through the shared base measure P_0 (induced through G in the illustrative example). One can potentially choose hyperpriors on location and scale parameters within P_0 to allow the data to inform about an average curve as well as variability in the curves. In addition, G can be specified to accommodate basis function selection through including mass at zero values. Due to the structure, inferences tend to be much less sensitive to the choice of P_0 than they would be under a hierarchical model that lets $f_i \sim P_0$ directly.

3.3 Basic Properties

Important special cases of the E-FSTB include (1) $M = 1$, $\Pi_1 = 1$, resulting in the $\mathcal{P}_\infty$ FSTB prior of expression (3), and (2) $a_h = 1$, $b_h = \alpha$, for $h = 1, \dots, \infty$, resulting in an enriched FDP (E-FDP). In addition, many existing functional data models such as hierarchical Gaussian processes and splines with Gaussian random effects correspond to the special case in which $\alpha \rightarrow \infty$ so that $f_i \sim P_0$ directly. From (6), it is clear that the E-FSTB prior is effectively a mixture of FSTB priors, reducing to a mixture of FDPs in the special case in which $a_h = 1$ and $b_h = \alpha$.

Note that iid samples from $P_{0,a}$ can be obtained by drawing functions from P_0 and discarding samples not in Ω_a (i.e., those functions not in feature class a). Alternatively, such rejection sampling can be avoided in favor of efficient sampling methods; for example, in using splines as in the above illustrative example, efficient algorithms are available for sampling from the prior and posterior under a rich variety of constraints (Shively, Walker and Damien, 2011). Rejection sampling can be prohibitively inefficient, as the probability of drawing a function from an unconstrained prior P_0 which satisfies a stringent set of constraints can be vanishingly small. Our R code for generating Figure 2 avoided rejection sampling and produced results effectively instantaneously.

For coherence, the probabilities allocated to the atoms in (4) must sum to one a.s., which follows from Lemma 1.

Lemma 1. $\sum_{h \in \mathcal{C}_a} P_{a,h} = 1$ almost surely, for all $a : \Pi_a > 0$.

To prove Lemma 1, we require the following regularity condition on P_0 :

$$C1: \Pi_{0,a} = \int_{\Omega} 1(\mathcal{J}(\Theta) = \mathcal{A}_a) P_0(d\Theta) = \int_{\Omega_a} P_0(d\Theta) > 0, \text{ for all } a : \Pi_a > 0,$$

with $\Pi_{0,a}$ the probability allocated to functions in feature subclass a by the base measure P_0 . This condition implies that Ω_a is a measurable subset of Ω , for all subclasses a having $\Pi_a > 0$. Under condition C1, the cardinality of $\mathcal{C}_a$ is ∞ , for $a : \Pi_a > 0$, so that there are infinitely many functional atoms falling within each such subclass. Lemma 1 then follows trivially from Lemma 1 of Ishwaran and James (2001).

Lemma 2. Letting $P \sim \mathcal{P}$ denote a probability measure generated from (4),

$$\int_{\Omega} 1\{\mathcal{J}(f) = \mathcal{A}_a\} P(df) = \Pi_a.$$

Lemma 2 follows from the application of Lemma 1 to expression (4). It follows that all probability measures generated from prior (4) allocate probability Π_a to feature class a . Hence, one can elicit the feature class probabilities separately from the base measure P_0 and hyperparameters $\{a_h, b_h, h = 1, \dots, \infty\}$.

The expectation of the random probability measure P can be expressed in a set theoretic

manner. Let $B \in \mathcal{B}$ denote a Borel set, expressed as $B = \sum_{h=1}^{\#\mathcal{A}} B_h$, where $B_h \in \mathcal{B}_h$. Then, for all $B \in \mathcal{B}$, we have

$$E\{P(B)\} = \sum_{h=1}^{\#\mathcal{A}} E(\Pi_h) E\{P_h(B_h)\} = \sum_{h=1}^{\#\mathcal{A}} \Pi_{0,h} P_{0,h}(B_h), \quad (7)$$

where $\Pi_{0,h} = a_{\Pi,h}/n_{\Pi}$ for $h = 1, \dots, \#\mathcal{A}$ and $n_{\Pi} = \sum_{h=1}^{\#\mathcal{A}} a_{\Pi,h}$. Focusing on $a_h = 1$ and $b_h = \alpha$, for $h = 1, \dots, \infty$, we can also derive the second moment:

$$\begin{aligned} E\{P(B)^2\} &= \sum_{h=1}^{\#\mathcal{A}} E(\Pi_h^2) E\{P_h(B_h)^2\} \\ &\quad + \sum_{h=1}^{\#\mathcal{A}} \sum_{l \neq h} E(\Pi_h \Pi_l) P_{0,h}(B_h) P_{0,l}(B_l), \\ &= \sum_{h=1}^{\#\mathcal{A}} \Pi_{0,h} \left(\frac{\Pi_{0,h} n_{\Pi} + 1}{n_{\Pi} + 1} \right) P_{0,h}(B_h) \left(\frac{\alpha P_{0,h}(B_h) + 1}{\alpha + 1} \right) \\ &\quad + \sum_{h=1}^{\#\mathcal{A}} \sum_{l \neq h} \Pi_{0,h} \Pi_{0,l} \left(\frac{n_{\Pi}}{n_{\Pi} + 1} \right) P_{0,h}(B_h) P_{0,l}(B_l), \end{aligned} \quad (8)$$

using the property that $P_a(B_a) \sim \text{beta}(\alpha P_{0,a}(B_a), \alpha \{1 - P_{0,a}(B_a)\})$. From expressions (7) and (8), one can also obtain the variance.

Because of the stick-breaking structure, there is a positive probability that two individuals are assigned to the same function. In particular, if $f_i \sim P$, for $i = 1, \dots, n$, then

$$\Pr(f_i = f_{i'}) = \sum_{h=1}^{\#\mathcal{A}} \Pi_{0,h} \left(\frac{\Pi_{0,h} n_{\Pi} + 1}{n_{\Pi} + 1} \right) \pi(\mathbf{a}, \mathbf{b}), \quad (9)$$

where the form of $\pi(\mathbf{a}, \mathbf{b})$ depends on the choice of $\mathbf{a} = \{a_h, h = 1, \dots, \infty\}$ and $\mathbf{b} = \{b_h, h = 1, \dots, \infty\}$ in the stick-breaking process. For an enriched functional two parameter Poisson Dirichlet (Pitman-Yor) process, which lets $a_h = 1 - a$ and $b_h = b + ha$ with $0 < a < 1$ and $b > -a$, $\pi(\mathbf{a}, \mathbf{b}) = (1 - a)/(1 + b)$, with the enriched functional DP corresponding to $a = 0$ and $b = \alpha$. For an enriched functional generalized Dirichlet process, which lets $a_h = a$ and $b_h = b$ with $a > 0$ and $b > 0$, we instead obtain $\pi(\mathbf{a}, \mathbf{b}) = 2b/(a + 2b + 1)$. The proof is in Appendix A.

3.4 Base Measure Specification

An important issue in the specification of an E-FSTB prior is the choice of base measure, P_0 . Gaussian processes (GPs) have been used previously for the FDP and extensions (Petrone et al. 2009; Nguyen and Gelfand, 2011). However, GPs have some disadvantages. Posterior computation requires calculation of multivariate Gaussian densities evaluated at $\mathbf{t} = \cup_{i=1}^{n+1} \mathbf{t}_i$, the union of observation times across subjects. The dimension of $\mathbf{t}$ can be large in many applications, leading to matrix inversion bottlenecks. Although there is a rich

literature proposing approximations to accelerate computation, such approaches focus primarily on the case in which there is a single function. In functional data analysis, issues in choosing a covariance to allow unknown smoothness for each function, while also accommodating flexible differences across subjects are challenging. As an alternative to GPs, previous authors have relied on basis expansions such as splines (Bigelow and Dunson, 2009) and wavelets (Ray and Mallick, 2006) in implementing FDP priors.

Splines are particularly convenient in accommodating a variety of constraints within the different feature classes, while also freeing up the computational bottleneck (at least for 1-d curves) and allowing for rich priors to be specified without loss of computational tractability. In contrast, posterior computation becomes daunting when the prior corresponds to generating the functional clusters from a GP subject to a constraint. For the remainder of the paper, we focus on a simple spline specification, which places a hierarchical variable selection prior on the basis coefficients (Smith and Kohn, 1996; George and McCulloch, 1997).

Letting $\Theta = \{\Theta_{a,h}, h = 1, \dots, \infty, a = 1, \dots, \#\mathcal{A}\}$, and $\mathcal{F}_h = a$ denote that the h th function in Θ belongs to class a , we assume

$$\Theta_{a,h}(t) = \sum_{l=1}^L \beta_{ahl} b_l(t), \text{ for all } t \in \mathcal{T}, \quad (10)$$

where $\beta_{ah} = (\beta_{ah1}, \dots, \beta_{ahL})^T$ are basis coefficients specific to the h th functional atom of the a th feature class, and $\{b_l, l = 1, \dots, L\}$ are splines. To generate $\Theta_{a,h} \sim P_{0,a}$, we draw the basis coefficients from

$$\beta_{ahl} \sim \nu_0 \delta_0(\cdot) + (1 - \nu_0) \phi(\cdot; 0, c), l = 1, \dots, L, \quad (11)$$

restricting draws to obtain $\Theta_{a,h}$ belonging to feature class $\mathcal{F}_h = a$. Assuming the data are centered in advance by subtracting the overall mean, we center the prior for the basis coefficients on zero, while borrowing information across functional clusters within the different feature classes in estimating the basis function exclusion probability ν_0 and coefficient variance c . Equation (11) could be modified to allow a different exclusion probability ν_{0a} for each feature class a , but this would limit the advantageous borrowing of information across classes.

By replacing $\phi(\cdot; 0, c)$ with $\phi(\cdot; \beta_l, c_l)$, with hyperpriors on $\{\beta_l, c_l\}$, we can modify (11) to allow centering on a common global curve $f_0(t) = \sum_{l=1}^L \beta_l b_l(t)$. This formulation has conceptual advantages in centering our nonparametric model on a spline random effects model, which allows borrowing of information across subjects in estimating the average curve, while also characterizing the degree of heterogeneity among subjects. Such a prior is only slightly more complicated to implement. However, we focus on the simpler structure in (11) because in our experience there is little to be gained in using an elaborate hierarchical structure in the base measure of a nonparametric Bayesian model. The proposed base measure is still highly flexible in allowing different basis function selection for the different functional atoms.

3.5 Truncations

As for the FSTB in expression (3), one can truncate the infinite sum representation of the E-FSTB, either to obtain a finite approximation or as a useful finite-dimensional prior. Starting with representation (6), we consider the following representation for $P_{\mathbf{N}} \sim \mathcal{P}_{\mathbf{N}}$:

$$P_{\mathbf{N}} = \sum_{a=1}^{\#\mathcal{A}} \Pi_a \sum_{h=1}^{N_a} \left\{ V_{a,h} \prod_{l < h} V_{a,l} \right\} \delta_{\Theta_{a,h}}, = \sum_{a=1}^{\#\mathcal{A}} \Pi_a \sum_{h=1}^{N_a} p_{ah} \delta_{\Theta_{a,h}}, \quad (12)$$

where $V_{a,h} \stackrel{ind}{\sim} \text{beta}(a_h, b_h)$, for $h = 1, \dots, N_a - 1$, $V_{a,N_a} = 1$, and $\Theta_{a,h} \sim P_{0,a}$, for $h = 1, \dots, N_a$. Here, $\mathbf{N} = \{N_a\}_{a=1}^{\#\mathcal{A}}$ is the level of truncation specific to each functional class.

In the subsequent development, we focus on the case in which $a_h = 1$ and $b_h = \alpha$. To assess the accuracy of inferences based on using (12) as an approximation to (4), or equivalently (6), we follow previous authors (Ishwaran and James, 2002; Rodriguez, Dunson and Gelfand, 2012) in bounding the distance between the true and approximated marginal likelihoods. In particular, assuming model (2) with $P_{\mathbf{N}}$ in place of P , the marginal likelihood is

$$m_{\mathbf{N}}(\mathbf{y}) = \int \int_{\mathfrak{R}^+} \left(\prod_{i=1}^n \int_{\Omega} p_{n_i}(\mathbf{y}_i | f_i(\mathbf{t}_i), \sigma^2) P(df_i) \right) g(d\sigma^2) \mathcal{P}_{\mathbf{N}}(dP) = \int \left(\prod_{i=1}^n \int_{\Omega} p_{n_i}(\mathbf{y}_i | f_i(\mathbf{t}_i)) P(df_i) \right) \mathcal{P}_{\mathbf{N}}(dP), \quad (13)$$

where $f_i(\mathbf{t}_i)$ denotes the values of the function f_i at points $\mathbf{t}_i$, $p_{n_i}(\mathbf{y} | \boldsymbol{\mu}, \sigma^2)$ is the product of n_i density functions $t_v(\sigma^2)$, with mean $\boldsymbol{\mu}$, evaluated at $\mathbf{y}$, $g(\cdot)$ denotes the prior density of σ^2 , and $p_{n_i}(\mathbf{y} | \boldsymbol{\mu}) = \int_{\mathfrak{R}^+} p_{n_i}(\mathbf{y} | \boldsymbol{\mu}, \sigma^2) g(d\sigma^2)$. We use $m_{\infty}(\mathbf{y})$ to denote the limiting case of $m_{\mathbf{N}}(\mathbf{y})$ as $N_a \rightarrow \infty$ for all a .

$$\textit{Theorem 1.} \int |m_{\mathbf{N}}(\mathbf{y}) - m_{\infty}(\mathbf{y})| d\mathbf{y} \leq 4 \left[1 - \left\{ 1 - \sum_{a=1}^{\#\mathcal{A}} \Pi_a \left(\frac{\alpha}{1+\alpha} \right)^{N_a-1} \right\}^n \right].$$

The proof, which follows similar lines to Rodriguez, Dunson and Gelfand (2012), is in Appendix A. It is clear that in the limit as $N_a \rightarrow \infty$, for $a = 1, \dots, \#\mathcal{A}$, that the upper bound converges to 0. Hence, in this limiting case, $m_{\mathbf{N}}(\mathbf{y})$ converges to $m_{\infty}(\mathbf{y})$ in total variation distance.

Note that one can obtain a simplified bound in the special case in which $N_a = N$, for $a = 1, \dots, \#\mathcal{A}$, so that the same truncation level is used within each feature class:

$$\int |m_{\mathbf{N}}(\mathbf{y}) - m_{\infty}(\mathbf{y})| d\mathbf{y} \leq 4 \left[1 - \left\{ 1 - \left(\frac{\alpha}{1+\alpha} \right)^{N-1} \right\}^n \right].$$

We plot the bound as a function of N for various choices of α and n in Figure 3. Viewing

$D = \sum_{a=1}^{\#\mathcal{A}} N_a$ as a fixed positive integer, one can potentially choose $\{N_a\}_{a=1}^{\#\mathcal{A}}$ to minimize the upper bound in Theorem 1. Even supposing that D is a multiple of $\#\mathcal{A}$, the solution to

this minimization problem is not in general $N_a = D/\#\mathcal{A}$. Instead, one can potentially obtain a more accurate approximation by allocating more atoms to feature classes having higher Π_a .

4. Posterior computation

We describe an MCMC algorithm for posterior computation in hierarchical model (2), with $\mathcal{P}$ chosen as an enriched functional DP and with the base measure specified as proposed in Section 3.4. We rely on truncation approximations as in Ishwaran and James (2001) to obtain an algorithm, which is straightforward to describe and implement. This algorithm can be easily modified to avoid truncations using slice sampling by modifying the approach of Kalli, Griffin and Walker (2011). However, the simple truncation approach is easier to describe and tends to produce essentially indistinguishable performance in other settings, so we focus on truncation here.

Let $\mathbf{S} = (\mathcal{S}_1, \dots, \mathcal{S}_n)^T$ define the global configuration of subjects to functional clusters, with $\mathcal{S}_i = h$ denoting that subject i belongs to cluster h , implying $f_i = \Theta_h$. Here, the index h is defined across feature classes, so we let $\mathcal{F}_h = a$ denote that cluster h belongs to feature class a . Relying on the truncation approach of Section 3.3, the index h is defined so that $\mathcal{F}_h = 1$ for $h = 1, \dots, N_1$, $\mathcal{F}_h = 2$ for $h = N_1 + 1, \dots, N_1 + N_2$, and so on.

After defining initial values, our MCMC algorithm iterates through the following steps:

1. Update Π by sampling from the full conditional posterior:

$$(\Pi|\mathbf{S}, \mathbf{F}, \mathbf{y}) \sim \text{Diri} \left(a_{\Pi,1} + \sum_{i=1}^n 1(\mathcal{F}_{\mathcal{S}_i} = 1), \dots, a_{\Pi,\#\mathcal{A}} + \sum_{i=1}^n 1(\mathcal{F}_{\mathcal{S}_i} = \#\mathcal{A}) \right), \quad (14)$$

2. Update V_{ah} , for $a = 1, \dots, \#\mathcal{A}$ and $h = 1, \dots, N_a - 1$, by sampling from the beta full conditional posterior distribution:

$$(V_{ah}|\mathbf{S}, \mathbf{F}, \mathbf{y}) \sim \text{beta} \left(1 + \sum_{i:\mathcal{F}_{\mathcal{S}_i}=a} I_{\mathcal{S}_i}(a, h), \alpha + \sum_{i:\mathcal{F}_{\mathcal{S}_i}=a} \sum_{l=h+1}^{N_a-1} I_{\mathcal{S}_i}(a, l) \right), \quad (15)$$

where $I_h(a, l) = 1$ if Θ_h is the l th atom in class a .

3. Update parameters for $\varepsilon_i(t)$. It is known (e.g. Congdon, 2001) that the Student- t distribution can be written

$$\varepsilon_i(t_{ij}) \sim N(0, \phi_{ij}^{-1} \sigma^2), \phi_{ij} \sim \text{gamma} \left(\frac{\nu}{2}, \frac{\nu}{2} \right).$$

Assuming $\nu \sim \text{gamma}(a_\nu, b_\nu)$ and $\sigma^{-2} = \tau \sim \text{gamma}(a_\tau, b_\tau)$, we

- a. Update ϕ_{ij}

$$(\phi_{ij}|\mathbf{S}, \mathbf{F}, \mathbf{V}, \Pi, \Theta, \mathbf{y}, \mathbf{t}, \nu, \sigma^2) \sim \text{gamma} \left(\frac{\nu}{2} + \frac{1}{2}, \frac{\nu}{2} + \frac{1}{2} \sigma^{-2} \{y_{ij} - f_i(t_{ij})\}^2 \right) \quad (16)$$

- b.** Update ν using Metropolis random walk, by generating a candidate

$$\nu^* \sim N(\nu^{t-1}, 0.25) \quad (17)$$

accept with probability

$$\min \left\{ 1, \frac{\text{gamma}(\nu^*; a_\nu, b_\nu) \prod_{i=1}^n \prod_{j=1}^{n_i} \text{gamma}(\phi_{ij}; \frac{\nu^*}{2}, \frac{\nu^*}{2})}{\text{gamma}(\nu^{(t-1)}; a_\nu, b_\nu) \prod_{i=1}^n \prod_{j=1}^{n_i} \text{gamma}(\phi_{ij}; \frac{\nu^{(t-1)}}{2}, \frac{\nu^{(t-1)}}{2})} \right\} \quad (18)$$

- c.** Update $\tau = \sigma^{-2}$ by sampling from the full conditional posterior:

$$(\tau | \mathbf{S}, \mathbf{F}, \mathbf{V}, \mathbf{\Pi}, \mathbf{\Theta}, \mathbf{y}, \mathbf{t}) \sim \text{gamma} \left(a_\tau + \frac{1}{2} \sum_{i=1}^n n_i, b_\tau + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{n_i} \phi_{ij} \{y_{ij} - f_i(t_{ij})\}^2 \right). \quad (19)$$

- 4.** Update the configuration of subjects to functional atoms by sampling $\mathcal{S}_i$ from the multinomial conditional posterior distribution, with

$$\Pr(S_i = h | \mathbf{V}, \mathbf{\Pi}, \mathbf{\Theta}, \tau, \mathbf{y}, \mathbf{t}) = \frac{p_h \prod_{j=1}^{n_i} \phi(y_{ij}; \Theta_h(t_{ij}), \tau^{-1} \phi_{ij}^{-1})}{\sum_{l=1}^{N_+} p_l \prod_{j=1}^{n_i} \phi(y_{ij}; \Theta_l(t_{ij}), \tau^{-1} \phi_{ij}^{-1})}, \quad (20)$$

where $N_+ = \sum_a N_a$, $\phi(z; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-1/(2\sigma^2)(z - \mu)^2\}$, and $p_h = \Pi_a V_{al} / \prod_{m < l} V_{am}$, for $a, l : I_h(a, l) = 1$.

- 5.** Update $\Theta_h(\cdot)$, for $h = 1, \dots, N_+$, by updating the basis coefficients, β_h , by sampling from the full conditional posterior distribution:

$$(\beta_{hl} | S_i = h, \mathcal{F}_h, \mathbf{V}, \mathbf{\Pi}, \beta_{h(-l)}, \tau, \mathbf{y}, \mathbf{t}) \sim \hat{\nu}_{hl} \delta_0(\cdot) + (1 - \hat{\nu}_{hl}) \phi_{A_{hl}}(\cdot; \hat{\beta}_{hl}, \hat{\sigma}_{\beta_{hl}}^2), \quad (21)$$

where A_{hl} denotes the subset of $\mathcal{R}$ such that the resulting function, Θ_h , belongs to feature class $\mathcal{F}_h$, $\phi_A(\cdot; \mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ density truncated to the region A ,

$$\hat{\nu}_{hl} = \frac{1}{1 + \frac{(1-\nu_0)}{\nu_0} \frac{\phi_{A_{hl}}(0; 0, c)}{\phi_{A_{hl}}(0; \hat{\beta}_{hl}, \hat{\sigma}_{\beta_{hl}}^2)}}, \quad \hat{\sigma}_{\beta_{hl}}^2 = \left(c^{-1} + \sum_{i: S_i = h} \sum_{j=1}^{n_i} \phi_{ij} \tau b_l(t_{ij})^2 \right)^{-1}, \quad \hat{\beta}_{hl} = \hat{\sigma}_{\beta_{hl}}^2 \left(\sum_{i: S_i = h} \sum_{j=1}^{n_i} \phi_{ij} \tau y_{ij}^{(hl)} b_l(t_{ij}) \right),$$

where $y_{ij}^{(hl)} = y_{ij} - \sum_{m \neq l} \beta_{hm} b_m(t_{ij})$.

In general problems, it is simple to sample from (21) by drawing the basis function inclusion indicator from Bernoulli($1 - \hat{\nu}_{hl}$), and then drawing β_{hl} from a univariate truncated Gaussian if this indicator is one, with $\beta_{hl} = 0$ otherwise. The main concern is slow mixing, which can occur in general when applying Gibbs sampling in constrained parameter problems; Pakman and Paninski (2012) propose an elegant solution using Hamiltonian Monte Carlo. Alternatively, one can generate a candidate β_{hl}^* for $l = 1, \dots, L$ from the unconstrained version of (21) that sets $A_{hl} =$

$\mathfrak{R}$, with this candidate vector accepted if the resulting function belongs to feature class $\mathcal{F}_h$. This approach was found to be more efficient in our applications, having 77% acceptance probability on average in the BBT application.

6. Update the probability of including a basis function by sampling as follows:

$$(\nu_0 | -) \sim \text{beta} \left(a_\nu + \sum_h \sum_l 1(\beta_{hl} = 0), b_\nu + \sum_h \sum_l 1(\beta_{hl} \neq 0) \right). \quad (22)$$

This algorithm is repeated for a large number of iterations, and posterior summaries are calculated using draws collected after a burn-in interval.

5. Simulation Examples

We focus on cubic B-splines (de Boor, 1978) and we simulate data from a finite mixture model chosen to mimic the temperature curve data, assuming 2 different feature types and $\# \mathcal{A} = 4$ different feature classes. The first feature corresponds to the occurrence of a biphasic temperature curve. Such a curve is defined as bi-phasic if there exists, between days 5 to 25, a horizontal line that intersects with the curve at a single point; that is, for some $a \in \mathfrak{R}$, $f_i(t) < f_i(t_a) < f_i(t')$ for all $t < t_a$ and $t' > t_a$, where t_a is the value of t such that $f_i(t) = a$. In addition, we define the shift day as

$t_{0i} = \underset{t_0}{\operatorname{argmax}} \{ f_i(t_0 + 2) - 2f_i(t_0 + 1) + f_i(t_0) : f_i(t_0 + 1) > f_i(t_0) + 0.01, t_0 \in [5, 25] \}$. The second feature corresponds to the shift in the curve occurring between days 10 and 20 of the cycle.

We simulated a total of 200 trajectories by polynomial functions of fifth degree describing patterns observed in real data. Data were simulated from

$$y_i(t) = a_{0i} + a_{1i}t + a_{2i}t^2 + a_{3i}t^3 + a_{4i}t^4 + a_{5i}t^5 + \varepsilon_i(t), \varepsilon_i(t) \sim N(0, 0.0049),$$

with ten different functional clusters chosen corresponding to the different values of $(a_{0i}, \dots, a_{5i})^T$ shown in Table 1, which also shows the probabilities for each functional cluster. Figure 4 shows some of the simulated trajectories and the underlying functions.

To complete a Bayesian specification of the model, we let $a_\tau = b_\tau = 1$, we choose $a_\Pi = (0.2, 0.7, 0, 0.1)^T$ to be weakly informative, corresponding to a prior sample size of one centered on our best guess of the different feature class proportions in the BBT application. We also let $a_\nu = b_\nu = 1$ and $c = 10$. We update α by following Escobar and West (1995), assuming a gamma (0.3, 0.1) prior, which favors few clusters while being quite diffuse to allow the data to inform strongly.

We ran the algorithm for 25,000 iterations after 10,000 burn-in samples. Examination of traceplots of the parameters and standard diagnostic tests showed no evidence against convergence and mixing was good. For each simulated subject, we estimated the posterior probabilities of their falling into each of the three possible feature classes. Assuming a 0–1 loss function, the Bayes' classifier allocated subjects to the feature class having the highest

posterior probability. This lead to perfect classification of the subjects into the correct classes.

Based on MCMC samples collected after burn-in, we estimated the mean integrated squared error relative to the truth for each sample and subject. This leads to a distribution of MISEs, which is summarized in Table 2 along with results for a functional Dirichlet process (FDP) and a piecewise linear parametric model contaminated with an FDP proposed by Scarpa and Dunson (2009). Clearly, the posterior for the subject-specific curves under the E-FDP is much more concentrated on the truth; the gains over the FDP seem likely due to enrichment. Interestingly, even though we chose a vague prior for the feature class probabilities, we observed substantial improvements relative to the FDP and the posterior distribution for the feature class probabilities was quite concentrated near the true values.

The posterior distribution of the measurement error variance is summarized in Table 3; note that the variance of the error term in the simulated data was 0.0049 so we slightly overestimate it. It is common in functional data analysis to overestimate the measurement error variance, as it is often difficult to distinguish the signal and noise processes. The two competitors overestimate the variance by a much larger amount.

Table 4 compares the proportion of functions falling in each feature class in the three models, showing that while E-FDP classifies all curves in the correct feature class, both the alternatives have some misclassification. In addition, as we can see in Figure 5, E-FDP more accurately estimates the curves in the different feature classes compared with FDP which inappropriately clusters certain curves. The contaminated FDP has worse fitting performance, as expected given the parametric model used. We additionally noticed a substantial MCMC efficiency difference in the proportion of curves allocated to each feature class, with E-FDP having close to 25,000 effective sample size for each category while FDP had approximately 2,300 effective sample sizes for classes 1–2 and contaminated FDP ranges from 3,700 – 6,200. The traceplots also exhibit a disturbing tendency for instability, jumping between widely different values.

6. Application: Temperature Curves in Menstrual Cycles

We applied the E-FDP to temperature curves in menstrual cycles from the European Fertility Study (Colombo and Masarotto, 2000). We analyzed data on daily observations of BBT for 1,118 non-conception cycles from 157 women in the Verona center of the study, excluding cycles that had < 7 days or > 50 days of cycle length. The women were followed prospectively and they recorded daily BBT and the days during which intercourse and menstrual bleeding occurred. The women had an average age of 28.6 years (standard deviation 3.54) and each woman contributed an average of 7.14 (6.62) cycles.

In healthy cycles, BBT trajectories tend to follow a characteristic bi-phasic pattern with the last day of hypothermia prior to the post-ovulatory rise in BBT commonly used as a marker of ovulation in scientific studies and for women timing their intercourse to attempt or avoid conception. Monitoring of BBT is an inexpensive non-invasive way of assessing a woman's reproductive functioning, and it is of substantial interest to develop statistical methods for automatically characterizing heterogeneity in BBT trajectories and classifying women into

different groups based on their trajectories. Currently this is done by expert examination of the BBT trajectories, with cycles then classified subjectively or according to simple rules applied to the daily observations. Abundant prior information is known about BBT for the majority of women. For example, it is well established in the cycle biometry literature (see for example Marshall, 1979) that the percentage of bi-phasic cycles is greater than 90% and among these more than 60% will likely present the shift in the central window between day 10 and day 20 of the cycle. We use these two features in our analysis, by formally defining the shift as in Section 5, with $a_{\Pi} = (0.65, 0.3, 0, 0.05)^T$ in our Dirichlet prior for the feature class probabilities. Class 1 is bi-phasic with a shift in $[10, 20]$, class 2 is bi-phasic with a shift outside $[10, 20]$, class 3 is empty since non bi-phasic curves do not have a defined shift, and class 4 is non bi-phasic.

There are several goal of our analysis. We would like to obtain a model for accurately characterizing the distribution of the BBT trajectories within and across the pre-specified feature classes, while also obtaining updated knowledge of the feature class probabilities. It is of substantial interest to examine how the BBT curves vary adjusting for measurement errors and missing data. Through subjective examination of the curves, it is likely that some aspects of the curves have been overlooked and the story is not as simple as presented in the literature. Our approach automatically allows learning of new trajectory clusters within each of the pre-specified feature classes. Using pre-specified feature classes in no way limits the overall exibility of the curves accommodated, but is highly useful for incorporating prior information, for classification and for producing interpretable results. Learning of the variability in curves within a feature class is of substantial interest, and allows reassessing of feature class definitions.

We choose weakly informative priors for the remaining parameters by letting $a_{\tau} = b_{\tau} = a_v = b_v = 1$, $c = 10$ and $\alpha \sim \text{gamma}(0.3, 0.1)$. We ran the Gibbs sampling algorithm for 20,000 iterations after a 8,000 burn-in period. In assessing convergence and mixing, we focus on quantities of inferential interest including functional trajectories for representative cycles and time points, the proportions of cycles falling into each feature class, and the measurement error variance. Trace plots demonstrate that mixing ranges from adequate to good. Tests of convergence including Heidelberger and Welch and Geweke showed no evidence for lack of convergence. Effective sample sizes were adequately large so that Monte Carlo errors in our posterior summaries are small.

The posterior mean of Π_1 , the proportion of cycles included in the first feature class, was 0.69 with 95% credible interval (0.68, 0.71) while the same quantities for the 2nd and 4th feature classes are $\hat{\Pi}_2 = 0.29$ (ci: 0.27–0.31) and $\hat{\Pi}_4 = 0.016$ (ci: 0.013–0.021), respectively. We estimated the posterior probability of falling in each of the three feature classes for each cycle under study, and cycles were allocated to the feature class having the highest posterior probability. Figure 6 shows the concentrations of the classes among samples: in the first panel we plot the distribution of the feature classes among all the samples for subjects classified by majority vote in feature class 1. The other panels show similar distributions for subject included by majority vote in feature class 2 and 4, respectively.

Figure 7 shows the posterior mean of the trajectory f_i for each cycle within each of the three feature classes. The variety of trajectories in the different feature classes is of substantial interest. The first feature class is interesting in having a broad variety of temperature levels and additionally exhibiting substantial variability in whether curves were at or not early and late in the cycle. A similar pattern was observed in feature class two but the rate of mid-cycle increase was much slower with broader differences, including curves that are biphasic but do not satisfy the usual expected pattern. The non bi-phasic class was similarly interesting, exhibiting a dominant cluster which is quite at and a singleton cluster with an erratic shape. This variety of curves is of substantial interest to reproductive biologists and clinicians. In addition, temperature trajectories for a future woman can be automatically classified using the proposed methods into the feature classes and sub-clusters, leading to insight into whether her pattern is consistent with typical variability in the bi-phasic shape or is possibly suggestive of dysfunction.

Figure 8 shows the raw BBT observations for some cycles in the sample, as well as the posterior mean and pointwise 95% credible intervals for the curves. Cycles were selected to be representative of the 3 feature classes.

Figures 7 and 8 illustrate the characteristics of the proposed process. We introduced weak prior information through the two features of presence and location of the shift, and the prediction of each single temperature curve was more precise with respect to a simple FDP. However, the choice to use only priors on the proportions of curves having these features, compared with the parametrically contaminated model (Scarpa and Dunson, 2009), allows for a more flexible fit of the curves belonging to each feature class, as can be clearly seen on the trajectories in Figure 7 and even better on the comparison between data and temperature curves in Figure 8.

7. Discussion

This article proposes an approach for including prior information on the frequencies of occurrence of different curve features; given that many fields routinely collapse curves into features prior to analysis, there is often rich prior knowledge of relevant feature classes. Such prior knowledge is not straightforward to include in existing functional data analysis methods. Even if one does not have perfect prior knowledge of features, choosing initial features to organize the functional data into categories is useful, can lead to improvements in efficiency, and additional features can be learned through the sub-clustering that occurs within feature classes. Curve features can be elicited collaboratively by statisticians and subject matter experts, with statisticians choosing mathematical definitions of feature classes based on expert input, showing prior realizations to experts, and refining definitions depending on feedback. When prior data bases are available for curves that have been labeled by experts to fall within different feature classes, one can automatically select a small number of features to maximize classification accuracy.

Although our focus has been on curve data, the enriched stick-breaking process can be applied directly in other contexts in which one has prior information on the proportions of atoms in a discrete distribution falling into particular feature classes. One extension is to

surface data; for example, in modeling reproductive hormones for women, it is natural to define a surface with the x-axis time relative to start of the cycle and the y-axis time relative to ovulation (Bigelow and Dunson, 2007). Our methods can be applied directly to surface applications, with the main hurdle being implementation of Step 5 in our Gibbs sampler for updating the atoms subject to feature class constraints. The projective Gaussian process of Lin and Dunson (2013) may be useful in this setting. Outside of functional data analysis, in density estimation, one may have prior information on the probability of a random variable being positive, while in regression prior information may be available on the probability the slope is positive or even that the mean is increasing across levels of a predictor.

A straightforward extension of substantial applied interest is to include predictors of class membership; the E-FSTB allows a simple approach in which the feature class allocation probabilities are dependent on predictors through a usual categorical response regression model embedded in the E-FSTB framework.

Acknowledgments

The authors dedicate this paper in memory of Professor Bernardo Colombo. This research was partially supported by grant number R01 ES017240 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH) and by the University of Padua CPDA097208/09 grant.

Appendix A

Proof of Clustering Property

Consider the model specification (6). Within each feature class, clustering is driven by the Chinese restaurant process induced from the assumed stick-breaking process, and in particular the choice of $\mathbf{a}$ and $\mathbf{b}$. For example, when $a_h = 1$ and $b_h = \alpha$ either $f_i = f_{i'}$, with prior probability $1/(1+\alpha)$, or f_i and $f_{i'}$ correspond to two independent draws from P_0 (assuming nonatomic P_0). We refer to the within-class cluster probability as $\pi(\mathbf{a}, \mathbf{b})$. For the two-parameter Poisson-Dirichlet process, the clustering probability is well known; refer to Ishwaran and James (2001). For the generalized Dirichlet process, the presented form for $\pi(\mathbf{a}, \mathbf{b})$ is a straightforward consequence of Theorem 1 in Rodriguez and Dunson (2010).

Given Π_a , the probability allocated to curves in the a th feature class by probability measure P , the probability that two independent individuals belong to the same feature class is Π_a^2 . Therefore within each feature class the probability that two individuals belong to the same feature class is $P(f_i = f_{i'} | f_i \text{ and } f_{i'} \text{ belong to feature class } a) = \Pi_a^2 \pi(\mathbf{a}, \mathbf{b})$. Since different feature classes are disjoint, the probability that two individuals are assigned to the same cluster is the sum of these probabilities:

$$\Pr(f_i = f_{i'} | \Pi) = \sum_{a=1}^{\#\mathcal{A}} \Pi_a^2 \pi(\mathbf{a}, \mathbf{b}), \quad (23)$$

and marginalizing out Π over the Dirichlet prior we obtain the result.

Proof of Theorem 1

Denote the joint probability measures for $\mathbf{f} = \{f_i\}_{i=1}^n$ under $\mathcal{P}_{\mathbf{N}}$ and $\mathcal{P}_{\infty}$ as $\pi_{\mathbf{N}}$ and π_{∞} , respectively. Then,

$$\begin{aligned} & \int \left| m_{\mathbf{N}}(\mathbf{y}) - m_{\infty}(\mathbf{y}) \right| d\mathbf{y} \\ &= \int \left| \int \prod_{i=1}^n \phi_{n_i}(\mathbf{y}_i | f_i(\mathbf{t}_i)) (\pi_{\mathbf{N}}(df) \right. \\ & \quad \left. - \pi_{\infty}(df)) \right| d\mathbf{y} \leq \int \int \prod_{i=1}^n \phi_{n_i}(\mathbf{y}_i | f_i(\mathbf{t}_i)) d\mathbf{y} \left| \pi_{\mathbf{N}}(df) \right. \\ & \quad \left. - \pi_{\infty}(df) \right| \\ &= 2D(\pi_{\mathbf{N}}, \pi_{\infty}), \end{aligned}$$

where $D(P_1, P_2) = \sup_A |P_1(A) - P_2(A)|$ is the total variation distance between the probability measures P_1 and P_2 . Let $\mathbf{n} = (n_1, \dots, n_{\#\mathcal{A}})^T \in \mathcal{N}_n(\mathcal{A})$ denote the number of

subjects allocated to feature classes 1, ..., $\#\mathcal{A}$, respectively, with $\sum_{a=1}^{\#\mathcal{A}} n_a = n$ and $\mathcal{N}_n(\mathcal{A})$ the space of $\#\mathcal{A} \times 1$ vectors of non-negative integers adding to n . In addition, let h_{ai} denote the atom number for subject i in feature class a . When $h_{ai} < N_a$, the probabilities allocated to the functional atoms are the same under $\pi_{\mathbf{N}}$ and π_{∞} . Hence, we have

$$\begin{aligned} D(\pi_{\mathbf{N}}, \pi_{\infty}) &\leq 2 \left(1 - \pi_{\mathbf{N}} \{ h_{ai} < N_a, \text{ for } a = 1, \dots, \#\mathcal{A}, i = 1, \dots, n_a \} \right) \\ &= 2 \left[1 - \sum_{\mathbf{n} \in \mathcal{N}_n(\mathcal{A})} \mathbb{P}(n_1, \dots, n_{\#\mathcal{A}}) \prod_{a=1}^{\#\mathcal{A}} \mathbb{E} \left\{ \left(\sum_{h=1}^{N_a-1} p_{ah} \right)^{n_a} \right\} \right] \leq 2 \left[1 - \sum_{\mathbf{n} \in \mathcal{N}_n(\mathcal{A})} \mathbb{P}(n_1, \dots, n_{\#\mathcal{A}}) \prod_{a=1}^{\#\mathcal{A}} \left(\sum_{h=1}^{N_a-1} \mathbb{E}(p_{ah}) \right)^{n_a} \right] \\ &= 2 \left[1 - \sum_{\mathbf{n} \in \mathcal{N}_n(\mathcal{A})} \mathbb{P}(n_1, \dots, n_{\#\mathcal{A}}) \prod_{a=1}^{\#\mathcal{A}} \left[1 - \left(\frac{\alpha}{1+\alpha} \right)^{N_a-1} \right]^{n_a} \right] \\ &= 2 \left(1 - \sum_{\mathbf{n} \in \mathcal{N}_n(\mathcal{A})} \binom{n}{n_1 \dots n_{\#\mathcal{A}}} \prod_{a=1}^{\#\mathcal{A}} \left[1 - \left(\frac{\alpha}{1+\alpha} \right)^{N_a-1} \right]^{n_a} \right) \\ &= 2 \left(1 - \left[\sum_{a=1}^{\#\mathcal{A}} \Pi_a \left\{ 1 - \left(\frac{\alpha}{1+\alpha} \right)^{N_a-1} \right\} \right]^n \right), \\ &= 2 \left[1 - \left\{ 1 - \sum_{a=1}^{\#\mathcal{A}} \Pi_a \left(\frac{\alpha}{1+\alpha} \right)^{N_a-1} \right\}^n \right], \end{aligned}$$

where $P(n_1, \dots, n_{\# \mathcal{A}})$ denotes the multinomial($n, \Pi_1, \dots, \Pi_{\# \mathcal{A}}$) probability of $\mathbf{n}$. This proof relies on a related approach to that used by Ishwaran and James (2002) and Rodriguez, Dunson and Gelfand (2012).

References

- Behseta S, Kass RE, Wallstrom GL. Hierarchical Models for Assessing Variability Among Functions. *Biometrika*. 2005; 92:419–434.
- Bigelow JL, Dunson DB. Bayesian Adaptive Regression Splines for Hierarchical Data. *Biometrics*. 2007; 63:724–732. [PubMed: 17403106]
- Bigelow JL, Dunson DB. Bayesian Semiparametric Joint Models for Functional Predictors. *Journal of the American Statistical Association*. 2009; 104:26–36.
- Bush CA, MacEachern SN. A Semiparametric Bayesian Model for Randomised Block Designs. *Biometrika*. 1996; 83:275–285.
- Colombo B, Masarotto G. Daily Fecundability: First Results from a New Data Base. *Demographic Research*. 2000; 3:5.
- Congdon, P. *Bayesian Statistical Modeling*. Wiley; 2001.
- de Boor, C. *A Practical Guide to Splines*. Springer; 1978.
- De Iorio M, Müller P, Rosner GL, MacEachern SN. An Anova Model for Dependent Random Measures. *Journal of the American Statistical Association*. 2004; 99:205–215.
- Duan J, Guindani M, Gelfand AE. Generalized Spatial Dirichlet Process Models. *Biometrika*. 2007; 94:809–825.
- Dunson DB, Herring AH, Siega-Riz Anna Maria. Bayesian Inference on Changes in Response Densities Over Predictor Clusters. *Journal of the American Statistical Association*. 2008; 103:1508–1517.
- Escobar MD, West M. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*. 1995; 90:577–588.
- Ferguson TS. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*. 1973; 1:209–230.
- Ferguson TS. Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*. 1974; 2:615–629.
- Gelfand AE, Kottas A, MacEachern SN. Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing. *Journal of the American Statistical Association*. 2005; 100:1021–1035.
- George EI, McCulloch RE. Approaches for Bayesian variable selection. *Statistica Sinica*. 1997; 7:339–373.
- Griffin J, Steel M. Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association*. 2006; 101:179–194.
- Ishwaran H, James LF. Gibbs Sampling for Stick-Breaking Priors. *Journal of the American Statistical Association*. 2001; 96:161–173.
- Ishwaran H, James LF. Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. *Journal of Computational and Graphical Statistics*. 2002; 11:1–26.
- James GM. Generalized Linear Models with Functional Predictors. *Journal of the Royal Statistical Society B*. 2002; 64:411–432.
- James GM, Hastie TJ, Sugar CA. Principal Component Models for Sparse Functional Data. *Biometrika*. 2000; 87:587–602.
- Ke C, Wang Y. Semiparametric Nonlinear Mixed-Effects Models and their Application (with Discussion). *Journal of the American Statistical Association*. 2001; 96:1272–1298.
- Kleinman K, Ibrahim J. A Semi-Parametric Bayesian Approach to Generalized Linear Mixed Models. *Statistics in Medicine*. 1998; 17:2579–2596. [PubMed: 9839349]
- Liang, F.; Liao, M.; Mukherjee, S.; West, M. ISDS Discussion Paper, Duke University. 2006. *Nonparametric Bayesian Kernel Models*.

- Lin, L.; Dunson, DB. Bayesian monotone regression using Gaussian process projection. Duke University; 2013. Available at <http://arxiv.org/abs/1306.4041>
- MacEachern SN. Dependent Nonparametric Processes. ASA Proceedings of the Section on Bayesian Statistical Science. 1999;50–55.
- MacEachern, SN. Technical Report. Department of Statistics, Ohio State University; 2000. Dependent Dirichlet Processes.
- MacLehose RF, Dunson DB. Nonparametric Bayes Kernel-based Priors for Functional Data Analysis. *Statistica Sinica*. 2008; 19:611–629.
- Marshall, J. Planning for a Family. An Atlas of Mucothermic Charts. 2nd edition.. London: Faber and Faber; 1979.
- Martinez AR, Vanhooff MHA, Schoute E, Vandermeer M, Broekmans FJM, Hompes PGA. The Reliability, Acceptability and Applications of Basal Body-Temperature (BBT) Records in the Diagnosis and Treatment of Infertility. *European Journal of Obstetrics & Gynecology and Reproductive Biology*. 1992; 47:121–127. [PubMed: 1459325]
- Medvedovic M, Sivaganesan S. Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles. *Bioinformatics*. 2002; 18:1194–1206. [PubMed: 12217911]
- Morris JS, Carroll RJ. Wavelet-Based Functional Mixed Models. *Journal of the Royal Statistical Society B*. 2006; 68:179–199.
- Mukhopadhyay S, Gelfand A. Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association*. 1997; 92:633–639.
- Neal RM. Slice sampling. *Annals of Statistics*. 2003; 31(3):705–767.
- Nguyen X, Gelfand AE. The Dirichlet labeling process for clustering functional data. *Statistica Sinica*. 2011; 21(3):1249–1289.
- Pakman, A.; Paninski, L. Exact hamiltonian monte carlo for truncated multivariate gaussians. Columbia University; 2012. Available at <http://arxiv.org/abs/1208.4118>
- Petrone S, Guindani M, Gelfand AE. Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society B*. 2009; 71:755–782.
- Pillai NS, Wu Q, Liang F, Mukherjee S, Wolpert RL. Characterizing the Function Space for Bayesian Kernel Models. *Journal of Machine Learning Research*. 2007; 8:1769–1797.
- Ramsay, JO.; Silverman, BW. *Functional Data Analysis*. New York: Springer; 1997.
- Rice J, Wu C. Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves. *Biometrics*. 2001; 57:253–259. [PubMed: 11252607]
- Ray S, Mallick B. Functional Clustering by Bayesian Wavelet Methods. *Journal of the Royal Statistical Society B*. 2006; 68:305–332.
- Rodriguez, A.; Dunson, DB. Discussion Paper, 2010-03. Department of Statistical Science, Duke University; 2010. Functional Clustering in Nested Designs.
- Rodriguez A, Dunson DB, Gelfand AE. The Nested Dirichlet Process. *Journal of the American Statistical Association*. 2008; 103:1131–1154.
- Rodriguez A, Dunson DB, Gelfand AE. Bayesian Nonparametric Functional Data Analysis through Density Estimation. *Biometrika*. 2009; 96:149–162. [PubMed: 19262739]
- Rodriguez A, Dunson DB, Gelfand AE. Latent Stick-Breaking Process. *Journal of the American Statistical Association*. 2012; 105:647–659. [PubMed: 23559690]
- Scarpa B, Dunson DB. Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors. *Biometrics*. 2009; 65:772–780. [PubMed: 19173703]
- Sethuraman J. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*. 1994; 4:639–650.
- Shively TS, Walker SG, Damien P. Nonparametric Function Estimation Subject to Monotonicity, Convexity and Other Shape Constraints. *Journal of Econometrics*. 2011; 161:166–181.
- Smith M, Kohn R. Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics*. 1996; 75(2):317–343.
- Thompson WK, Rosen O. A Bayesian Model for Sparse Functional Data. *Biometrics*. 2007; 64:54–63. [PubMed: 17573864]

Figure 1.
Fertility study: The BBT data for two cycles from two women. These data exhibit the characteristic biphasic shape expected for a healthy, ovulating woman.

Figure 2. 10 samples from the prior for Θ_h for functions falling in each feature class, including (1) monotone and range in $[0,1]$; (2) monotone and range not in $[0,1]$; (3) non-monotone and range in $[0,1]$; and (4) non-monotone and range not in $[0,1]$.

Figure 3. Upper bound of the distance between the true and approximated marginal likelihoods as function of N for various choices of α and n

E-FSTB

Figure 4.
Simulated data: noisy trajectories and underlying curves

Figure 5.
Simulated data: variety of curves represented in each feature class for some models

Figure 6. Fertility study: distribution of feature classes for cycles classified (by majority vote) in feature class 1, 2 and 4

Figure 7.
Fertility study: posterior mean trajectories for cycles belonging to each feature class

Figure 8.

Fertility study: BBT raw data and estimated curves, including posterior means and point wise 95% intervals, for representative women.

Table 1
Simulated data: coefficients used for the fifth degree polynomial function in each cluster within each feature class.

Cluster	Feature Class	Pr	a_0	a_1	a_2	a_3	a_4	a_5
A	1	0.115	36.22	0.1420	-0.03460	$3.25 \cdot 10^3$	$-1.23 \cdot 10^4$	$1.61 \cdot 10^6$
B	1	0.105	36.65	-0.0730	0.01390	$-7.26 \cdot 10^4$	$1.18 \cdot 10^5$	0.00
C	2	0.115	36.17	0.3201	-0.03774	$1.89 \cdot 10^3$	$-3.90 \cdot 10^5$	$2.10 \cdot 10^7$
D	2	0.125	36.30	0.1700	-0.02345	$1.88 \cdot 10^3$	$-7.71 \cdot 10^5$	$1.20 \cdot 10^6$
E	2	0.105	36.50	-0.0190	0.00800	$-2.50 \cdot 10^4$	0.00	0.00
F	1	0.120	36.60	-0.0700	0.00800	$-2.00 \cdot 10^4$	0.00	0.00
G	2	0.115	36.30	0.1300	-0.00630	$9.00 \cdot 10^5$	0.00	0.00
H	2	0.125	36.30	0.1270	-0.00850	$1.72 \cdot 10^4$	0.00	0.00
I	4	0.040	37.17	-0.0960	0.00300	0.00	0.00	0.00
L	4	0.035	36.40	0.0000	0.00000	0.00	0.00	0.00

Table 2

Simulated data: summary of the mean integrated squared error for some models

	Models		
	E-FDP	FDP	FDP contaminated
First quartile	4.07	11.10	14.83
Median	4.16	11.24	15.19
Third quartile	4.25	11.38	15.55
Mean	4.17	11.25	15.19
SD	0.144	0.211	0.536

Table 3

Simulated data: summary of the estimate of the measurement error variance for some models

	Models		
	E-FDP	FDP	FDP contaminated
First quartile	0.00513	0.0103	0.00603
Median	0.00521	0.0104	0.00612
Third quartile	0.00530	0.0106	0.00621
Mean	0.00522	0.0104	0.00612
SD	0.00013	0.00022	0.00024

Table 4

Simulated data: percentages of misclassified units for each model.

Actual feature class	Predicted feature class	Models		
		E-FDP	FDP	FDP contaminated
1	2	0	16.31%	17.02%
	4	0	0	0
2	1	0	54.54%	36.36%
	4	0	0	0
4	1	0	0	46.67%
	2	0	0	0