Statistica Sinica **20** (2010), 1395-1422

MULTIVARIATE KERNEL PARTITION PROCESS MIXTURES

David B. Dunson

Duke University

Abstract: Mixtures provide a useful approach for relaxing parametric assumptions. Discrete mixture models induce clusters, typically with the same cluster allocation for each parameter in multivariate cases. As a more flexible approach that facilitates sparse nonparametric modeling of multivariate random effects distributions, this article proposes a kernel partition process (KPP) in which the cluster allocation varies for different parameters. The KPP is shown to be the driving measure for a multivariate ordered Chinese restaurant process that induces a highly-flexible dependence structure in local clustering. This structure allows the relative locations of the random effects to inform the clustering process, with spatially-proximal random effects likely to be assigned the same cluster index. An exact block Gibbs sampler is developed for posterior computation, avoiding truncation of the infinite measure. The methods are applied to hormone curve data, and a dependent KPP is proposed for classification from functional predictors.

Key words and phrases: Chinese restaurant process, Dirichlet process, discriminant analysis, local clustering, longitudinal data, nonparametric Bayes, random effects.

1. Introduction

Mixture models are used for addressing a broad variety of problems including model-based clustering (Banfield and Raftery (1993)), density estimation (Fraley and Raftery (2002)), and supervised classification (Hastie and Tibshirani (1996)). In multivariate cases, the vast majority of the literature focuses on *global* mixture models in which subjects are allocated to the same mixture component index for all of their parameters. To provide motivation for local alternatives to global mixture models, consider the functional data analysis model with

$$f_i(x) = \sum_{j=1}^{p} \theta_{ij} b_j(x), \quad \boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ip})' \sim P, \tag{1.1}$$

where $\mathbf{b} = \{b_j\}_{j=1}^p$ is a collection of basis functions, such as splines or kernels, $\boldsymbol{\theta}_i$ is a random effects vector, and P is a random effects distribution.

To allow flexible modeling of functional data, such as longitudinal trajectories, P is commonly assumed to have a discrete form with

$$P = \sum_{h=1}^{k} \pi_h \delta_{\boldsymbol{\Theta}_h}, \tag{1.2}$$

where k is the number of mixture components, π_h is the probability of allocation to component h, δ_Θ is a degenerate distribution with unit probability mass at Θ and $\boldsymbol{\Theta}_h = (\Theta_{h1}, \ldots, \Theta_{hp})'$ is a vector of random effects coefficients specific to component h. Subjects allocated to component h have function $f_i(x) = f_h^*(x) = \mathbf{b}(x)'\boldsymbol{\Theta}_h$. For applications of mixture modeling to functional clustering applications, refer to Heard, Holmes and Stephens (2006) and Ray and Mallick (2006). In longitudinal data applications, latent class trajectory models (Muthén and Shedden (1999)) are widely used. These models allow $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ to depend on predictors. Implementation proceeds via the EM algorithm, with the BIC used to select the number of classes k.

The global mixture model that results from using (1.2) for the random effects distribution P has some distinct disadvantages when P is moderate to high dimensional, as is typically the case for functional data. As motivation, consider an application to modeling of PdG (a metabolite of progesterone) in early pregnancy; these data have previously been analyzed by Bigelow and Dunson (2009) and Dunson (2009). Figure 1 shows the data for five of the 165 women in the study. After ovulation, which is estimated using a highly-accurate hormonal marker, progesterone tends to rise to a plateau in healthy pregnancies. However, the shapes vary substantially, and pregnancies that result in an early loss exhibit a decline, as is apparent for two of the women in Figure 1. Assuming for simplicity that $\mathbf{b}$ corresponds to a pre-specified collection of 20 Gaussian kernels placed at equally-spaced times and applying the model specified in (1.1)−(1.2), it is not clear whether the model with $k = 2$ or $k = 5$ components is preferred. Although there are two groups of curves having similar shapes, there are also substantial local differences between the individual curves in these two groups. In addition, all the curves are similar up to day 10-12, which corresponds approximately to the timing of implantation.

As the dimension p increases, it is increasingly unlikely that two subjects are similar with respect to all the elements of their random effects vectors. Hence, global mixture models will either allocate subjects to many clusters, leading to an inefficient characterization of the data, or will inappropriately cluster subjects that are similar at most locations, obscuring local differences. In practice, both of these problems occur and one can obtain a simultaneous improvement in goodness of fit and reduction in model complexity using a carefully-specified local mixture model.

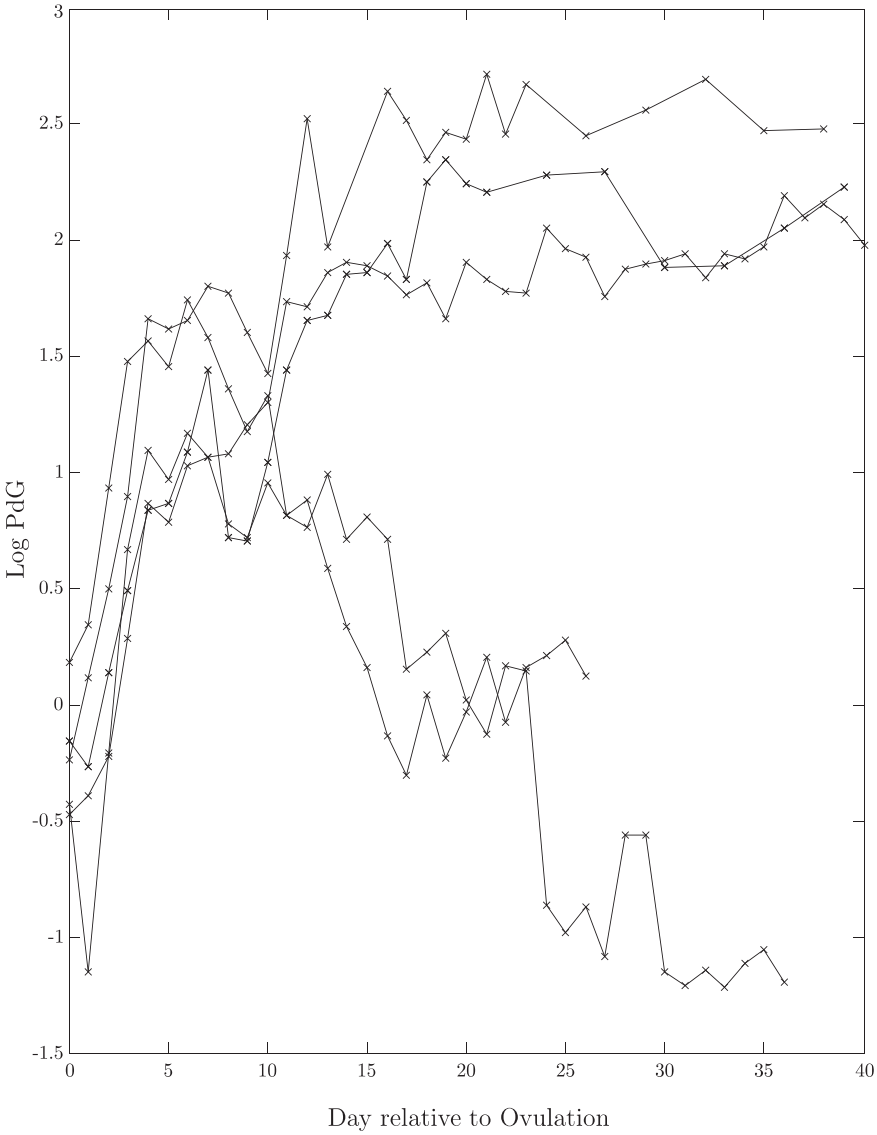

Figure 1. Log PdG data in early pregnancy for five randomly selected women.

Expression (1.2) is appealing in facilitating a reduction of dimensionality from n random effects vectors, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, to k coefficient vectors, $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k$. To maintain this characteristic, while relaxing the assumption of global clustering, one can replace (1.2) with a local mixture model of the form

$$P = \sum_{h_1=1}^{k} \cdots \sum_{h_p=1}^{k} \pi_{h_1 \cdots h_p} \delta_{\boldsymbol{\Theta}_{\mathbf{h}}}, \quad \boldsymbol{\Theta}_{\mathbf{h}} = \big(\Theta_{h_1 1}, \ldots, \Theta_{h_p p}\big)', \tag{1.3}$$

where $\pi_{h_1 \cdots h_p}$ is the probability of allocation to component h_1 for the first element of the random effect vector, h_2 for the second element, and so on up to h_p for the pth element. Hence, as in (1.2), there are k clusters. However, instead of forcing the cluster index to be the same for every parameter, I allow a different cluster allocation for each parameter. This additional flexibility comes at the price of incorporating an array of $k^p - 1$ allocation probabilities $\pi = \{\pi_{h_1 \cdots h_p}, h_j = 1, \ldots, k, j = 1, \ldots, p\}$ instead of a vector of $k - 1$ allocation probabilities. However, this increase in dimensionality of the model for the allocation probabilities is typically more than made up for by a reduction in the value of k needed to characterize the data.

As for the global latent class trajectory model that uses (1.2) for the random effects distribution in (1.1), the EM algorithm can potentially be used for maximum likelihood estimation under the local mixture model (1.3). However, to improve estimation of π and the coefficient vectors $\{\boldsymbol{\Theta}_h\}_{h=1}^k$, it is appealing to incorporate a shrinkage prior. In addition, the assumption that all subjects can be allocated to a finite number of groups k regardless of sample size is unappealing. For example, suppose that we estimate that $k = 3$ based on an initial sample of 100 subjects, and data later become available for an additional 1,000 subjects. Ideally, the model would be flexible enough to allow new sub-populations, and hence mixture components, to be represented in the new sample. Following this line of reasoning, the number of mixture components *occupied* by the subjects in a sample should grow without bound as the sample size increases. This can be accomplished automatically within a coherent nonparametric Bayes framework by setting $k = \infty$.

A Dirichlet process (DP) prior (Ferguson (1973, 1974)) on the random effects distribution induces the global mixture model in (1.2) with

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\Theta}_h}, \ \pi_h = \pi_h^* \prod_{l<h} (1 - \pi_l^*), \ \pi_h^* \sim \text{beta}(1, \alpha), \ \boldsymbol{\Theta}_h \sim P_0, \qquad (1.4)$$

$h = 1, \ldots, \infty$. This is the so-called stick-breaking representation of Sethuraman (1994). Although $k = \infty$, the weights decrease stochastically as the index h increases, so that the n subjects will tend to be allocated to a small number of components relative to the sample size, particularly if α is small. Indeed, the expected number of occupied components is proportional to $\alpha \log n$, suggesting a slow rate of discovery of new clusters as additional subjects are added. DP priors and DP mixtures (Lo (1984); Escobar and West (1995)) have been widely used for modeling of unknown random effects distributions (Bush and MacEachern (1996); Müller and Rosner (1997); Kleinman and Ibrahim (1998)).

The goal of this article is to generalize the DP prior to develop useful classes of priors that have the local mixture form in (1.3), with $k = \infty$, instead of the

global form in (1.2). Dunson, Xue, and Carin (2008) and Dunson (2009) proposed approaches for addressing this problem based on matrix stick-breaking process (MSBP) and local partition process (LPP) priors, respectively. However, these approaches simplify modeling of the infinite probability tensor π controlling the local allocation to mixture components by assuming an exchangeable dependence structure. In particular, it is assumed that $\Pr(\theta_{ij} = \theta_{i'j} \,|\, \theta_{il} = \theta_{i'l}) = \Pr(\theta_{ij} = \theta_{i'j} \,|\, \theta_{im} = \theta_{i'm})$ for all $l, m \in \{1, \ldots, p\}$ with $l \neq m$. This assumption is an over-simplification in applications in which the random effects have locations or a natural ordering. For example, suppose b_j in model (1.1) is a kernel located at $\mathbf{z}_j \in \mathcal{Z}$, for $j = 1, \ldots, p$, and let

$$\Delta_{jj'} = \Pr(\theta_{ij} = \theta_{i'j} \,|\, \theta_{ij'} = \theta_{i'j'}) - \Pr(\theta_{ij} = \theta_{i'j})$$

represent the increase in the probability of allocating two subjects to the same cluster for the jth random effect given knowledge that they are allocated to the same cluster for the j'th random effect. Then, intuitively, the value of $\Delta_{jj'}$ should be greater *a priori* when $\mathbf{z}_{j'}$ is close to $\mathbf{z}_j$, since then the subjects are known to be similar at a location close to the location $\mathbf{z}_j$ of the jth random effect. This article proposes a kernel partition process (KPP) prior for P that allows incorporation of the feature matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p)$.

Also motivated by functional data analysis, Petrone, Guindani, and Gelfand (2009) proposed a hybrid Dirichlet process prior that has a fundamentally different structure than the KPP. Their approach formulates the individual functions as hybrids of global functions drawn from a Gaussian process (GP). A latent GP copula is used to allow local surface selection, inducing local clustering of functions. A related alternative that also includes a GP copula was proposed by Rodriguez, Dunson, and Gelfand (2010). These approaches are quite flexible, but can be computationally intensive due to the need to update high-dimensional latent variables within MCMC sampling. The KPP has computational advantages and avoids the need to truncate the infinite-dimensional representation.

Mixed membership models allow subjects to be proportionally allocated to different clusters (Woodbury, Clive, and Garson (1978); Rosenberg et al. (2002); Erosheva, Fienberg, and Lafferty (2006)). Mixture model (1.3) instead allows different allocation for different parameters, and hence is related to methods for characterizing dependence in latent class analysis (Chung, Lanza, and Loken (2008)). However, previous methods for dependent latent allocation rely on Markov or other restrictive assumptions, while the KPP allows the dependence structure to be unknown.

Section 2 proposes the KPP formulation, discusses properties and considers relationships with previous methods. Section 3 develops an MCMC algorithm

for posterior computation. Section 4 considers a simulation example, Section 5 applies the methods to hormone curve data, and Section 6 discusses the results.

2. Kernel Partition Processes

2.1. Proposed prior

Letting $\boldsymbol{\theta}_i \sim P$, I propose a prior $P \sim \text{KPP}(\alpha, \beta, \psi, P_0)$ for the unknown random effects distribution, with $\alpha > 0, \beta > 0, \psi > 0$ scalar hyperparameters and $P_0 = \otimes_{j=1}^{p} P_{0j}$ a base measure providing an initial guess for P. The KPP prior for P, which is explicitly specified at (2.3), leads to a hierarchical model for the random effects:

$$\begin{aligned} &\boldsymbol{\theta}_i = \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}, \quad \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i} = (\Theta_{\gamma_{i1}1}, \ldots, \Theta_{\gamma_{ip}p})', \ \boldsymbol{\Theta}_h \sim P_0, \ h = 1, \ldots, \infty, \\ &\gamma_{ij} \sim \sum_{g=1}^{p} \omega_{jg} \delta_{\phi_{ig}}, \quad \phi_{ig} \sim \sum_{t=1}^{\infty} \nu_t \delta_t, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{ip})' \in \{1, \ldots, \infty\}^p$ is a multivariate cluster index for subject i, $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{ip})'$ is a vector of group-specific latent cluster indices, $\omega_{jg} = \Pr(\gamma_{ij} = \phi_{ig})$ is the probability of allocating the jth random effect to group g, and $\nu_t = \Pr(\gamma_{ij} = t)$ is the marginal probability of allocation to cluster t. For parsimony and computational efficiency, the marginal distribution of γ_{ij} is assumed constant for $j = 1, \ldots, p$, though this assumption can be relaxed by replacing ν_t with ν_{gt}.

Letting $G_{ij} = g$ denote that the jth random effect from subject i is allocated to group g, all random effects in the same group are assigned the same cluster index, with $\gamma_{ij} = \phi_{ig}$ for all $j : G_{ij} = g$. The cluster indices for different groups are generated independently. Hence, dependence in the elements of $\boldsymbol{\gamma}_i$ is entirely induced through the process of grouping the random effects, which is controlled by the choice of $\{\omega_{jg}\}$. Letting $\mathbf{z}_j$ and $\mathbf{z}_g$ denote features for the jth random effect and gth group, respectively, it is appealing to allow these features to inform the grouping process. For example, considering the functional data analysis model in (1.1), $\mathbf{z}_j$ could be chosen to correspond to the location of the jth kernel basis, for $j = 1, \ldots, p$. Then, by allowing the probability of allocating the jth random effect to group g to decrease with distance between $\mathbf{z}_j$ and $\mathbf{z}_g$, as measured by a kernel $K_\psi : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$, one automatically makes it more likely to allocate random effects for nearby basis functions to the same group.

To complete the specification, I choose explicit processes for the probability

weights in (2.1) by letting

$$\omega_{jg} = \frac{\lambda_g K_\psi(\mathbf{z}_j, \mathbf{z}_g)}{\sum_{g'=1}^{p} \lambda_{g'} K_\psi(\mathbf{z}_j, \mathbf{z}_{g'})}, \quad \lambda_g \sim \text{gamma}(\frac{\beta}{p}, 1),\ j, g = 1, \ldots, p,$$
$$\nu_t = \nu_t^* \prod_{s<t}(1-\nu_s^*), \quad \nu_t^* \sim \text{beta}(1, \alpha),\ t = 1, \ldots, \infty, \tag{2.2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ are random weights for groups $g = 1, \ldots, p$, respectively, ψ is a kernel precision, and $\Pr(\phi_{ig} = t) = \nu_t$ follows a stick-breaking process, with precision α controlling how rapidly the probabilities decrease as t increases. The hyperparameter β controls the variability in the group-specific weights, with values of β close to zero leading to a small number of dominant groups having much higher weights. Smaller β tends to induce greater dependence in the elements of $\boldsymbol{\gamma}_i$ in leading to few large groups. Tighter kernels K_ψ induce a local dependence structure in which the elements of $\boldsymbol{\gamma}_i$ tend to be identical within local blocks, particularly for small β. In practice, hyperpriors are chosen for α, β, ψ to allow the data to inform about their values.

Expression (2.1) is consistent with an explicit KPP prior for P,

$$P = \sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{\infty} \pi_{h_1 \cdots h_p} \delta_{\boldsymbol{\Theta}_\mathbf{h}},\ \boldsymbol{\Theta}_\mathbf{h} = (\Theta_{h_1 1}, \ldots, \Theta_{h_p p})',\ \boldsymbol{\Theta}_h \sim P_0,$$
$$\pi_{h_1 \cdots h_p} = \sum_{J \in \mathcal{J}_\mathbf{h}} \prod_{g: J_g \neq \emptyset} \nu_{h(J_g)} \prod_{j \in J_g} \omega_{jg}, \tag{2.3}$$

where $\Pr(\gamma_1 = h_1, \ldots, \gamma_p = h_p) = \pi_{h_1 \cdots h_p}$, $J = (J_1, \ldots, J_p)$ is a partition of the index set $\{1, \ldots, p\}$ into p groups, with $J_g \subset \{1, \ldots, p\}$ the indices in group g, $h(J_g) = \{h_j, j \in J_g\}$, and $\mathcal{J}_\mathbf{h}$ is the set of all partitions J satisfying $\#h(J_g) = 1$ for all g such that $J_g \neq \emptyset$, with $\#A$ denoting the cardinality of set A.

Let P correspond to a probability measure over $\{\Omega, \mathcal{B}(\Omega)\}$, where Ω is a measurable Polish space, and $\mathcal{B}(\Omega)$ is the Borel σ-algebra of subsets of Ω. The prior expectation of P can be expressed as

$$\begin{aligned} \mathrm{E}\big\{P(B)\big\} &= \mathrm{E}\Big[\sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{\infty} \pi_{h_1 \cdots h_p} 1\big\{(\Theta_{h_1 1}, \ldots, \Theta_{h_p p})' \in B\big\}\Big] \\ &= \mathrm{E}\Big\{ \sum_{h_1=1}^{\infty} \cdots \sum_{h_p=1}^{\infty} \pi_{h_1 \cdots h_p} \Big\} \int 1(\boldsymbol{\Theta} \in B) dP_0(\boldsymbol{\Theta}) = P_0(B) \end{aligned} \tag{2.4}$$

for all $B \in \mathcal{B}(\Omega)$. Hence, the prior for P is centered on P_0.

The hybrid Dirichlet process (hDP) of Petrone, Guindani, and Gelfand (2009) also induces a prior having the form in (1.3), but with a carefully-specified

joint Dirichlet distribution on the weights $\pi = \{\pi_{h_1 \cdots h_p}, h_j = 1, \ldots, k, j = 1, \ldots, p\}$. They focus on the finite k case in developing the prior and in conducting posterior computation and inferences, though they show weak limit properties as $k \to \infty$. To accommodate spatial dependence in the allocation to components, they incorporate a hidden label process, which leads to challenging computation and identifiability issues. The LPP prior of Dunson (2009) has the form shown in the first line of (2.3), but with a fundamentally different prior on π, that is induced through mixing global and independent Dirichlet process priors to obtain an exchangeable dependence structure for $\boldsymbol{\gamma}$. The KPP has advantages over the LPP in terms of allowing inclusion of spatial dependence that can lead to substantial improvements in interpolations and predictions.

2.2. Predictive probability function

To better understand the KPP prior and the associated clustering process, it is useful to marginalize out the infinitely-many stick-breaking random variables $\boldsymbol{\nu}^* = \{\nu_t^*\}_{t=1}^{\infty}$ and random atoms $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_h\}_{h=1}^{\infty}$ to obtain a conditional prediction rule. Before deriving a prediction rule for the KPP, I review the Dirichlet process prediction rule, and propose a modification, which is then generalized to the KPP case in Theorem 1.

In the case in which $\boldsymbol{\theta}_i \sim P$, with $P \sim DP(\alpha P_0)$ assigned a Dirichlet process prior, Blackwell and MacQueen (1973) showed that the conditional distribution of $\boldsymbol{\theta}_{n+1}$ given $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, but marginalizing out P, is

$$(\boldsymbol{\theta}_{n+1} \,|\, \tilde{\boldsymbol{\gamma}}^{(n)}, \tilde{\boldsymbol{\theta}}^{(n)}) = \left(\frac{\alpha}{\alpha + n}\right) P_0 + \sum_{h=1}^{\tilde{k}^{(n)}} \left(\frac{\tilde{n}_h^{(n)}}{\alpha + n}\right) \delta_{\tilde{\boldsymbol{\theta}}_h}, \quad n = 0, 1, \ldots, \infty, \tag{2.5}$$

where $\tilde{\boldsymbol{\gamma}}^{(n)} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_n)'$, $\tilde{\gamma}_i = h$ denotes that the ith subject is allocated to cluster h, with clusters indexed in the order of appearance as subjects $i = 1, \ldots, n$ are added, $\tilde{\boldsymbol{\theta}}_h$ is the value of $\boldsymbol{\theta}_i$ for the $\tilde{n}_h^{(n)} = \sum_{i=1}^{n} 1(\tilde{\gamma}_i = h)$ subjects in cluster h, and $\tilde{k}^{(n)}$ is the number of unique values of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)'$.

To obtain intuition for the Polya urn scheme in (2.5), it is common to rely on a Chinese restaurant process (CRP) metaphor (Aldous (1985); Ishwaran and James (2003)). In particular, consider a Chinese restaurant with infinitely many tables, with each table containing a dish generated from P_0. The first customer entering the restaurant is seated at the first table, and the second customer is either seated at the first table with probability $1/(\alpha + 1)$ or is seated at the unoccupied second table with probability $\alpha/(\alpha + 1)$. As the CRP proceeds, the nth customer entering the restaurant is seated at an occupied table with probability proportional to the number of customers seated at that table and is seated at a new table with probability proportional to α.

In order for the CRP to be more closely linked to the DP stick-breaking process in (1.4), it is convenient to define a modification, which I refer to as the ordered CRP (oCRP). The distinguishing characteristic of the oCRP is that the tables are indexed not by the order in which they are occupied but by the order of the components in the stick-breaking representation. Because the stick-breaking weights are stochastically-decreasing, one can modify the metaphor so that the restaurant has infinitely-many tables ordered in terms of desirability, with table $t = 1$ the most desirable. Lemma 1 provides the predictive probability of allocation to table t for individual $i = n + 1$ under the oCRP.

Lemma 1. *Assuming that* $\boldsymbol{\theta}_i \sim P$ *for* $i = 1, 2, \ldots, n+1$, *with* $P \sim DP(\alpha P_0)$ *as in* (1.4), $\gamma_i = t$ *if* $\boldsymbol{\theta}_i = \boldsymbol{\Theta}_t$, *and* $\boldsymbol{\gamma}^{(n)} = (\gamma_1, \ldots, \gamma_n)'$, *we have*

$$Pr(\gamma_{n+1} = t \,|\, \boldsymbol{\gamma}^{(n)}) = \left(\frac{1 + n_t^{(n)}}{\alpha + 1 + \sum_{s=t}^{k^{(n)}} n_s^{(n)}}\right) \prod_{h=1}^{t-1} \left(\frac{\alpha + \sum_{s=h+1}^{k^{(n)}} n_s^{(n)}}{\alpha + 1 + \sum_{s=h}^{k^{(n)}} n_s^{(n)}}\right),$$

$t = 1, \ldots, \infty$, *where* $n_t^{(n)} = \sum_{i=1}^{n} 1(\gamma_i = t)$ *is the number of customers sitting at table* t *and* $k^{(n)} = \max\{\boldsymbol{\gamma}^{(n)}\}$ *is the maximum index of an occupied table.*

The KPP prior proposed in Section 2.1 induces a multivariate extension of the oCRP. Suppose households $i = 1, \ldots, n+1$ each contain p types of individuals who are allocated to groups. Individuals of type j have traits $\mathbf{z}_j$ and select group g with probability proportional to $\lambda_g K(\mathbf{z}_j, \mathbf{z}_g)$, for $g = 1, \ldots, p$. Individuals within a group are more likely to have similar traits than individuals in different groups, though very popular groups having high λ_g values may contain widely different types of individuals unless the kernel is tight. Individuals from a household that are in the same group are seated together when they arrive at the restaurant, with the seating following an oCRP that depends on the number of previous groups seated at each table instead of the number of individuals. Theorem 1 shows the resulting multivariate predictive probability function.

Theorem 1. *Let* $G_{ij} = g$ *if individual* j *in household* i *is allocated to group* g, $T_{ig} = t$ *if group* g *from household* i *is seated at table* t, $\mathbf{G}_i = (G_{i1}, \ldots, G_{ip})'$, $\mathbf{G}^{(n)} = \{\mathbf{G}_i\}_{i=1}^n$, $n_{ig} = \sum_{j=1}^p 1(G_{ij} = g)$ *denotes the number of individuals from household* i *in group* g, $\mathbf{T}_i = \{T_{ig}, g : n_{ig} > 0\}$ *denotes the table assignment for the groups from household* i, *and* $\mathbf{T}^{(n)} = \{\mathbf{T}_i\}_{i=1}^n$. *Then,*

$$\begin{aligned}&\big(\boldsymbol{\theta}_{n+1}, \mathbf{G}_{n+1} = \mathbf{g}, \mathbf{T}_{n+1} = \mathbf{t} \,|\, \boldsymbol{\theta}^{(n)}, \mathbf{G}^{(n)}, \mathbf{T}^{(n)}\big) \\ &= \prod_{j=1}^{p} \omega_{jg_j} \pi_{jt_{g_j}}^{(n)} \big\{1\big(w_j^{(n)} = 0\big) P_{0j}(\theta_{n+1,j}) + 1(w_j^{(n)} > 0) \delta_{\Theta_{t_{g_j} j}}(\theta_{n+1,j})\big\},\end{aligned}$$

where $\mathbf{g} = (g_1, \ldots, g_p)'$ *are the groups for individuals* $1, \ldots, p$ *in household* $n + 1$, $\mathbf{t} = \{t_g, g \in \mathbf{g}\}$ *are the tables for each group in household* $n + 1$, $w_j^{(n)} =$

$\sum_{i=1}^{n} 1(T_{iG_{ij}} = t_{g_j})$ *is the number of individuals of type* j *seated at table* t_{g_j}, ω_{jg} *is defined in* (2.2), $\pi_{jt}^{(n)} = 1(t = t_{g_l}, l : g_l = g_j, l < j)$ *if* $g_j \in \{g_1, \ldots, g_{j-1}\}$ *and otherwise*

$$\pi_{jt}^{(n)} = \left(\frac{1 + n_{jt}^{(n)}}{\alpha + 1 + \sum_{s=t}^{k_j^{(n)}} n_{js}^{(n)}} \right) \prod_{h=1}^{t-1} \left(\frac{\alpha + \sum_{s=h+1}^{k_j^{(n)}} n_{js}^{(n)}}{\alpha + 1 + \sum_{s=h}^{k_j^{(n)}} n_{js}^{(n)}} \right), \quad t = 1, \ldots, \infty,$$

with $n_{jt}^{(n)} = \sum_{i=1}^{n} \sum_{g: n_{ig} > 0} 1(T_{ig} = t) + \sum_{g \in \{g_1, \ldots, g_{j-1}\}} 1(t_g = t)$ *the number of groups seated at table* t, *including individuals* $1, \ldots, j-1$ *in household* $n+1$, *and* $k_j^{(n)} = \max\{\mathbf{T}^{(n)}, t_{g_1}, \ldots, t_{g_{j-1}}\}$.

Proofs of Lemma 1 and Theorem 1 are provided in an Appendix. If an individual of type j is seated at a table that does not already contain an individual of that type, they are given a new dish sampled from P_{0j}, with this dish served to all future individuals of type j seated at that table. In sequentially applying Theorem 1 for $n = 1, 2, \ldots$, the first member of a group to be seated is assigned to the first few tables with high probability if α is small, with the subsequent members in the same household and group automatically assigned to the same table. Small α leads to a sparse formulation with few large tables, and hence a small number of high probability clusters. In addition, when β is small, the popularity of the groups varies greatly, leading to a small number of dominant groups within each household unless the kernel is very tight. When all members of a household are in the same group, global clustering is induced.

Proposition 1. *Under* (2.1),

$$\text{(i)} \quad Pr(\gamma_{ij} = \gamma_{ij'}) = \frac{1 + \alpha \boldsymbol{\omega}_j' \boldsymbol{\omega}_{j'}}{1 + \alpha} = \rho_{jj'}, \quad \textit{for } j \neq j',$$

$$\text{(ii)} \quad Pr(\gamma_{ij} = \gamma_{i'j'}) = \frac{1}{1 + \alpha} = \kappa_{jj'}, \qquad \textit{for any } j, j',$$

where $\boldsymbol{\omega}_j = (\omega_{j1}, \ldots, \omega_{jp})'$.

Proposition 1 (i) gives the probability of allocating two individuals in a household of type j and j' to the same table and hence the same cluster index, while Proposition 1 (ii) gives the probability of assigning individuals from two different households to the same table. For concreteness, consider a longitudinal data application in which $f_i(x) = \mathbf{b}(x)'\boldsymbol{\theta}_i$, with $b_j(x) = \exp\{-\psi_b(x - z_j)^2\}$ a kernel basis function located at time z_j for $j = 1, \ldots, p$. The within-subject pairwise dependence structure in the vector of cluster indices $\boldsymbol{\gamma}_i$ is characterized by Proposition 1 (i), while the cross-subject dependence is characterized by Proposition 1 (ii). Within a subject it is more likely for random effects that are located close

together to be allocated to the same cluster index. The $\boldsymbol{\omega}_j'\boldsymbol{\omega}_{j'}$ term is a cross-product of weighted and normalized kernel functions, and so induces a highly flexible dependence structure.

The hyperparameter α provides a global control on clustering, with small values implying a high probability of being allocated to the same cluster index even for random effects located far apart. Indeed, in the limiting case $\lim_{\alpha\to 0}\rho_{jj'} = \kappa_{jj'} = 1$, and $\boldsymbol{\theta}_i \sim \delta_{\boldsymbol{\Theta}_1}$, so that all subjects are allocated to a single global cluster. A small but non-zero α instead implies a sparse formulation with a combination of short and long range dependence in clustering.

Proposition 2. *Let $A : G_{ij} = G_{ij'}$ denote allocation of individuals j and j' from household i to the same group, $B : G_{i'j} = G_{i'j'}$ denote allocation of individuals j and j' from household i' to the same group, and $\overline{A}$ and $\overline{B}$ denote the complements of A and B, respectively. Then, under the KPP prior,*

$$Pr(\theta_{ij} = \theta_{i'j} \,|\, \theta_{ij'} = \theta_{i'j'}) = \frac{1}{1+\alpha}\left\{\Delta_1 + \frac{\Delta_2}{2+\alpha} + \frac{(6+\alpha)\Delta_3}{(2+\alpha)(3+\alpha)}\right\},$$

where $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \Delta_3)'$ is a probability vector, with $\Delta_1 = Pr(A\cap B) = (\boldsymbol{\omega}_j'\boldsymbol{\omega}_{j'})^2$, $\Delta_2 = Pr\{(A\cap\overline{B})\cup(\overline{A}\cap B)\} = 2\boldsymbol{\omega}_j'\boldsymbol{\omega}_{j'}\boldsymbol{\omega}_j'(1-\boldsymbol{\omega}_{j'})$, and $\Delta_3 = Pr(\overline{A}\cap\overline{B}) = \{\boldsymbol{\omega}_j'(1-\boldsymbol{\omega}_{j'})\}^2$.

As $\Pr(\theta_{ij} = \theta_{i'j}) = 1/(1+\alpha)$, it is clear from Proposition 2 that the probability of clustering subjects i and i' for their jth random effect increases given information that these subjects are clustered for their j'th random effect. This is an appealing property in allowing borrowing of information in local clustering. The magnitude of the multiplicative increase in $\{\cdot\}$ tends to decrease as the distance between the jth and j'th random effect increases, though the function can take an extremely wide variety of shapes and is not restricted to be monotone due to the incorporation of the adaptive weights $\boldsymbol{\lambda}$.

2.3. Some special cases

We consider special cases of the KPP. First, suppose $K_\psi(\mathbf{z}_j, \mathbf{z}_g) = 1(j = g)$ for all j, g. In this case, each individual is assigned to their own group, and hence there is no within-household dependence in the seating process. Interestingly, in this case we obtain a fixed-π dependent Dirichlet process (DDP) prior for $\{P_j\}_{j=1}^p$ (De Iorio et al. (2004)), with

$$\theta_{ij} \sim P_j, \quad P_j = \sum_{h=1}^{\infty}\nu_h\delta_{\Theta_{hj}}, \quad \boldsymbol{\Theta}_h = (\Theta_{h1}, \ldots, \Theta_{hp})' \sim P_0.$$

Although we assumed $P_0 = \otimes_{j=1}^p P_{0j}$ for simplicity, it is straightforward to modify the specification to include dependence in P_0.

As another extreme case, let $K_\psi(\mathbf{z}_j, \mathbf{z}_g) = 1$ for all j, g, which leads to $\boldsymbol{\omega}_j = \boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)'$ with $\omega_g = \lambda_g / \sum_{l=1}^p \lambda_l$. Because $\lambda_g \sim \text{gamma}(\beta/p, 1)$, we obtain $\boldsymbol{\omega} \sim \text{Dirichlet}(\beta/p, \ldots, \beta/p)$. As β decreases, $\boldsymbol{\omega}$ converges in distribution to $\delta_{\mathbf{v}_\xi}$, with $\mathbf{v}_\xi$ a $p \times 1$ vector of zeros with a single one in position $\xi \sim \sum_{l=1}^p (1/p)\delta_l$. Hence, from (2.1) it is clear we obtain $\gamma_{ij} = \gamma_i = \phi_{i\xi}$ for all i, j, leading to $P \sim \text{DP}(\alpha P_0)$. By replacing the DP-type stick-breaking prior on the distribution of ϕ_{ig} in (2.1), we can induce an arbitrary species sampling prior on P.

In addition to fixed-π DDP and global DP priors, we can induce an exchangeable within-subject dependence structure by letting $K_\psi(\mathbf{z}_j, \mathbf{z}_g) = 1$ for all j, g with $\beta > 0$. In this case, it follows from Proposition 1 that

$$\Pr(\gamma_{ij} = \gamma_{ij'}) = \frac{1 + \alpha \boldsymbol{\omega}'\boldsymbol{\omega}}{1 + \alpha}, \quad \text{for all } i, j, \quad \boldsymbol{\omega} \sim \text{Dirichlet}(\beta/p, \ldots, \beta/p).$$

As $\beta \to 0$, $\boldsymbol{\omega} \xrightarrow{\mathcal{D}} \delta_{\mathbf{v}_\xi}$, so that $\boldsymbol{\omega}'\boldsymbol{\omega} \xrightarrow{\mathcal{D}} \delta_1$ and $\Pr(\gamma_{ij} = \gamma_{ij'}) \xrightarrow{\mathcal{D}} \delta_1$.

3. Posterior Computation

For posterior computation, we propose a Markov chain Monte Carlo (MCMC) algorithm, which is a hybrid of data augmentation, the exact block Gibbs sampler of Papaspiliopoulos (2008) and Yau et al. (2008), and Metropolis sampling. Papaspiliopoulos (2008) proposed the exact block Gibbs sampler as an efficient approach to posterior computation in Dirichlet process mixture models, modifying the block Gibbs sampler of Ishwaran and James (2001) to avoid truncation approximations. The exact block Gibbs sampler combines characteristics of the retrospective sampler (Papaspilioupoulos and Roberts (2008)) and the slice sampler (Walker (2007)).

For concreteness, we focus on a functional data analysis model with $y_{it} \sim t_\upsilon\big(f_i(x_{it}), \tau\big)$, where $t_\upsilon(\mu, \tau)$ denotes the t-density centered on μ, with υ degrees of freedom and scale parameter τ. In addition, $f_i(x)$ follows (1.1) with $\boldsymbol{\theta}_i \sim P$, $P \sim \text{KPP}(\alpha, \beta, \psi, P_0)$, and $P_0 = \otimes_{j=1}^p P_0^*$, where P_0^* denotes a Cauchy prior centered on zero. The Cauchy prior is an appealing choice for robust shrinkage of the basis coefficients, as small coefficients tend to be shrunk to very close to zero, while larger coefficients fall in the tails (Bhuiyan, Ahmad, and Swamy (2007)). To complete a Bayes specification, let $\alpha \sim \text{gamma}(a_\alpha, b_\alpha)$, $\beta \sim \text{gamma}(a_\beta, b_\beta)$, $\psi \sim \text{gamma}(a_\psi, b_\psi)$, $\upsilon \sim \text{gamma}(a_\upsilon, b_\upsilon)$, and $\tau \sim \text{gamma}(a_\tau, b_\tau)$. From West (1987), the t-density can be expressed as a scale mixture of Gaussians, which results in $y_{it} \sim N(\mathbf{b}_{it}'\boldsymbol{\theta}_i, \tau^{-1}\varphi_{it}^{-1})$, $\mathbf{b}_{it} = \mathbf{b}(x_{it})$, $\varphi_{it} \sim \text{gamma}(\upsilon/2, \upsilon/2)$, $P_0^* = N(0, \kappa^{-1})$, and $\kappa \sim \text{gamma}(1/2, 1/2)$.

The algorithm proceeds as follows.

1. Update G_{ij} for all i, j from the multinomial conditional posterior, with

$$\Pr(G_{ij} = g \,|\, -) = \frac{\omega_{jg} \prod_{t=1}^{n_i} N\big(y_{it}; \mathbf{b}_{it}' \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i(G_{ij}=g)}, \tau^{-1}\varphi_{it}^{-1}\big)}{\sum_{l=1}^{p} \omega_{jl} \prod_{t=1}^{n_i} N\big(y_{it}; \mathbf{b}_{it}' \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i(G_{ij}=l)}, \tau^{-1}\varphi_{it}^{-1}\big)}, \quad (3.1)$$

$h = 1, \ldots, p$, where $\boldsymbol{\Theta}_{\boldsymbol{\gamma}_i(G_{ij}=g)}$ is $(\Theta_{\gamma_{i1}1}, \ldots, \Theta_{\gamma_{ip}p})'$ but with the current value of γ_{ij} set to ϕ_{ig}.

2. To update λ_g, for $g = 1, \ldots, p$, use a data augmentation approach related to Holmes and Held (2006) and Dunson, Pillai, and Park (2007). Letting $K_{jg} = K_\psi(\mathbf{z}_j, \mathbf{z}_g)$ and $K_{jg}^* = K_{jg}/\big(\sum_{l \neq g} \lambda_l K_{jl}\big)$, the conditional likelihood for λ_g is

$$L(\lambda_g) = \prod_{j=1}^{p} \left(\frac{\lambda_g K_{jg}^*}{1 + \lambda_g K_{jg}^*}\right)^{\sum_i 1(G_{ij}=g)} \left(\frac{1}{1 + \lambda_g K_{jg}^*}\right)^{\sum_i 1(G_{ij}\neq g)},$$

obtained via $1(G_{ij} = g) = 1(Z_{ijg}^* > 0)$, with $Z_{ijg}^* \sim \text{Poisson}(\lambda_g \xi_{ijg} K_{jg}^*)$ and $\xi_{ijg} \sim \exp(01)$. Update $\{Z_{ijg}^*, \xi_{ijg}\}$ and $\{\lambda_g\}$ in Gibbs steps as follows.

   (a) Let $Z_{ijg}^* = 0$ if $G_{ij} \neq g$ and otherwise $Z_{ijg}^* \sim \text{Poisson}(\lambda_g \xi_{ijg} K_{jg}^*) 1(Z_{ijg*} > 0)$.
   (b) $\xi_{ijg} \sim \text{gamma}(1 + Z_{ijg}^*, 1 + \lambda_g K_{jg}^*)$.
   (c) $\lambda_g \sim \text{gamma}\big(\beta/p + \sum_{i,j} Z_{ijg}^*, 1 + \sum_{i,j} \xi_{ijg} K_{jg}^*\big)$.

3. Update ψ, β, υ using Metropolis steps.

4. Update φ_{it}, $i = 1, \ldots, n, t = 1, \ldots, n_i$, and τ by sampling from gamma full conditionals

$$(\varphi_{it} \,|\, -) \sim \text{gamma}\left(\frac{\upsilon + 1}{2}, \frac{\upsilon}{2} + \frac{\tau}{2}(y_{it} - \mathbf{b}_{it}'\boldsymbol{\theta}_i)^2\right),$$

$$(\tau \,|\, -) \sim \text{gamma}\left(a_\tau + \frac{1}{2}\sum_{i=1}^{n} n_i, b_\tau + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{n_i} \varphi_{it}(y_{it} - \mathbf{b}_{it}'\boldsymbol{\theta}_i)^2\right).$$

5. Implement exact block Gibbs sampler steps as follows.

   (a) Sample $u_{ig} \sim \text{uniform}(0, \nu_{\phi_{ig}})$, for $i = 1, \ldots, n$, with $\nu_h = \nu_h^* \prod_{l<h}(1 - \nu_l^*)$.
   (b) Sample the stick-breaking random variables,

$$\nu_h^* \sim \text{beta}\left(1 + \sum_{g=1}^{p}\sum_{i=1}^{n} 1(\phi_{ig} = h), \alpha + \sum_{g=1}^{p}\sum_{i=1}^{n} 1(\phi_{ig} > h)\right),$$

   for $h = 1, \ldots, \phi^*$, with ϕ^* the minimum value satisfying $\nu_1 + \cdots + \nu_{\phi^*} > 1 - \min\{u_{ig}\}$.

(c) Update $\boldsymbol{\Theta}_h$, for $h = 1, \ldots, \phi^*$, from $N_p(\widehat{\boldsymbol{\Theta}}_h, \Sigma_{\boldsymbol{\Theta}_h})$ with

$$\widehat{\boldsymbol{\Theta}}_h = \Sigma_{\boldsymbol{\Theta}_h} \sum_{i=1}^{n} \sum_{t=1}^{n_i} \tau \varphi_{it} \Gamma_{ih} \mathbf{b}_{it} (y_{it} - \mathbf{b}_{it}' \Gamma_{i(-h)} \boldsymbol{\theta}_i),$$

$$\Sigma_{\boldsymbol{\Theta}_h} = \left(\kappa \mathbf{I}_p + \sum_{i=1}^{n} \sum_{t=1}^{n_i} \tau \varphi_{it} \Gamma_{ih} \mathbf{b}_{it} \mathbf{b}_{it}' \Gamma_{ih} \right),$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\Gamma_{ih} = \operatorname{diag}\big(1(\gamma_{i1} = h), \ldots, 1(\gamma_{ip} = h)\big)$, and $\Gamma_{i(-h)} = \operatorname{diag}\big(1(\gamma_{i1} \neq h), \ldots, 1(\gamma_{ip} \neq h)\big)$.

(d) Update ϕ_{ig}, $i = 1, \ldots, n$, $g = 1, \ldots, p$ from the multinomial conditional with

$$\Pr(\phi_{ig} = h \mid -) \propto 1(u_{ig} < \nu_h) \prod_{t=1}^{n_i} N\left(y_{it}^{(g)}; \sum_{j=1}^{p} 1(G_{ij} = g) b_{itj} \Theta_{hj}, \tau^{-1} \varphi_{it}^{-1} \right),$$

$h = 1, \ldots, \phi^*$, where $y_{it}^{(g)} = y_{it} - \sum_{j=1}^{p} 1(G_{ij} \neq g) b_{itj} \theta_{ij}$.

6. Update α by sampling from the conditional posterior

$$(\alpha \mid -) \sim \text{gamma}\left(a_\alpha + \phi^*, b_\alpha - \sum_{h=1}^{\phi^*} \log(1 - \nu_h^*) \right).$$

7. Update κ by sampling from the conditional posterior

$$(\kappa \mid -) \sim \text{gamma}\left(\frac{1}{2} + \frac{p\phi^*}{2}, \frac{1}{2} + \frac{1}{2} \sum_{h=1}^{\phi^*} \sum_{j=1}^{p} \Theta_{hj}^2 \right).$$

This algorithm is straightforward to program and has exhibited good rates of convergence and mixing in simulations and in data applications. I also considered using the slice sampler of Walker (2007) in place of the exact block Gibbs sampling steps, but this resulted in considerably slower rates of convergence and mixing. Adaptations for more complex random effects models, which include fixed and random effects and other complications, are trivial by imbedding steps similar to those outlined above in an MCMC algorithm that includes additional steps to update fixed effects and any additional unknowns involved in the more complex model.

4. Simulation Example

To assess the performance of the approach, I assumed $n = 100$ and simulated data under the functional data analysis model described in Section 3, with $n =$

100, $\upsilon = 10$, $\tau = 20$, $\mathcal{X} = (0,1)$, $b_1(x) = 1$, $b_{j+1}(x) = \exp\{-25(x - z_j)^2\}$, and $z_j = (j-1)/19$, for $j = 1, \ldots, 19$. In addition, I let $n_i = 10$ plus a discrete uniform random variable on $[1, \ldots, 10]$, for $i = 1, \ldots, n$, and simulated $t_{ij} \sim \text{Uniform}(0, C_i)$, with $C_i = 1$ for the first 50 subjects and $C_i = 2/3$ for the second 50 subjects. I let $\boldsymbol{\Theta}_h = (\boldsymbol{\Theta}_{h1}, \ldots, \boldsymbol{\Theta}_{l20})'$, for $h = 1, 2, 3$, with the elements $\Theta_{hl} \sim 0.8\delta_0 + 0.2N(0, 2)$ independently. Then, letting $\theta_{ij} = \Theta_{\gamma_{ij} j}$, for $j = 1, \ldots, p$, I simulated the elements of $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{ip})'$ from a Markov chain, with γ_{i1} set to a random element of $\{1, 2, 3\}$, $\gamma_{ij} = \gamma_{i,j-1}$ with probability 0.9, and γ_{ij} otherwise set to a random element of $\{1, 2, 3\}$.

The approach of Section 3 was applied after choosing hyperpriors by letting $a_\alpha = b_\alpha = 1$, $a_\beta = b_\beta = 1$, $a_\psi = 25$, $b_\psi = 1$, $a_\upsilon = 4$, $b_\upsilon = 1$, and $a_\tau = b_\tau = 0.1$. These were considered reasonable default values when $\mathcal{X} = (0,1)$ and the overall variance of y is within a factor of 10 of 1. Values of $\alpha \approx 1$ favor a sparse specification, with most of the probability mass placed on the first few cluster indices, while a gamma(1,1) prior for β allows substantial variability in the group-specific weights, tending to favor a few dominant groups. However, the hyperpriors are vague enough to allow substantial uncertainty. The prior for υ favors heavy-tailed measurement errors and hence robust estimation, while the prior for the scale parameter τ is vague if the data are normalized in advance. In addition, I let $K_\psi(z, z') = \exp\{-\psi(z - z')^2\}$ and $z_j = z_j^*$, for $j = 1, \ldots, 19$. Note that these values of z_j, z_j^*, and p provide a reasonable default for smooth curves, as there are sufficient numbers of equally-spaced kernels to capture a very wide variety of smooth curve shapes. The results are robust to increases in the number of kernels due to the adaptive shrinkage resulting from allowing the data to inform about κ and β. The MCMC algorithm was run for 22,500 iterations including a 7500 iteration burn-in, with the chain thinned by collecting every 15 iterations due to storage constraints.

Based on examination of trace plots of the parameters and function estimates at a variety of points for different subjects, convergence was rapid and mixing was good. Note that it is important to avoid diagnosing convergence based on examination of the unique coefficient vectors, $\boldsymbol{\Theta}_h$, due to the well-known label switching issue that is omnipresent in mixture models (Jasra, Holmes, and Stephens (2005)). This label switching does not create problems as long as the focus of inference is not on mixture component-specific quantities and one obtains good rates of convergence and mixing in quantities of interest, such as individual-specific function estimates.

Figure 2 shows the data, estimated posterior mean curves (solid lines), 95% pointwise credible intervals (dashed lines), and true curves (dotted lines) for subjects $1, \ldots, 8$ (top 8 panels) and subjects $51, \ldots, 58$ (bottom 8 panels). Recall that subjects $1, \ldots, 50$ have observations throughout the $[0, 1]$ interval, while

subjects $51, \ldots, 100$ have no observations in the $[2/3, 1]$ interval. For subjects $1, \ldots, 50$ the estimates are very close to the true curves, while there is some deviation after the 2/3 point for some of the $51, \ldots, 100$ subjects. However, in general, predictions across this interval are quite good, with the true curves enclosed in the credible bounds. The average mean square error across a dense grid of times between 0 and 1 is 0.156 and the mean width of 95% credible intervals is 0.916.

Repeating the analysis for the LPP prior of Dunson (2009), the results were very good at locations close to data points. However, interpolations and predictions were not as good as for the KPP. For example, for the subject in the (2,3) panel there was a substantial gap in the observations. Across this gap, the 95% credible intervals were much wider in the LPP analysis. In addition, across the [2/3,1] region for many of the subjects $51, \ldots, 100$, the LPP estimates were substantially farther from the truth than the KPP-based estimates. The mean square error for the LPP was 0.183.

5. Application to Hormone Curve Data

5.1. Function estimation

I consider progesterone curve data in early pregnancy previously analyzed by Dunson (2009) using a functional data analysis model with kernel basis functions and t-distributed measurement errors. That article demonstrated improved performance for a local partition process (LPP) prior on the random effects distribution relative to a DP. Data consisted of daily urinary measurements of PdG, a metabolite of progesterone, starting on the identified day of ovulation and continuing for up to 40 additional days. There were 165 women, with an average of 23 measurements per women (range $= [4, 41]$). To analyze these data using the KPP prior for the random effects distribution, I implemented the approach used in Section 4, with the same prior specification.

In examining plots of the individual curve estimates for each of the 165 women, it is apparent that the estimates fit the data very well. Figure 3 shows the estimated curves and 95% pointwise credible intervals for 16 randomly selected women. There is a small but notable improvement in fit in using the KPP prior for the random effects distribution instead of the LPP prior. The improvement in fit is not attributable to over-fitting, as it is clear that a very sparse representation of the data is obtained in examining the estimated parameters in Table 1. In particular, the small α value suggests that a small number of unique coefficient vectors are sufficient to characterize the data. Indeed, the estimate of ϕ^* was 4.72, implying that the basis coefficient vectors for all 165 women are constituted of elements selected from ~ 5 unique coefficient vectors. In addition, the small

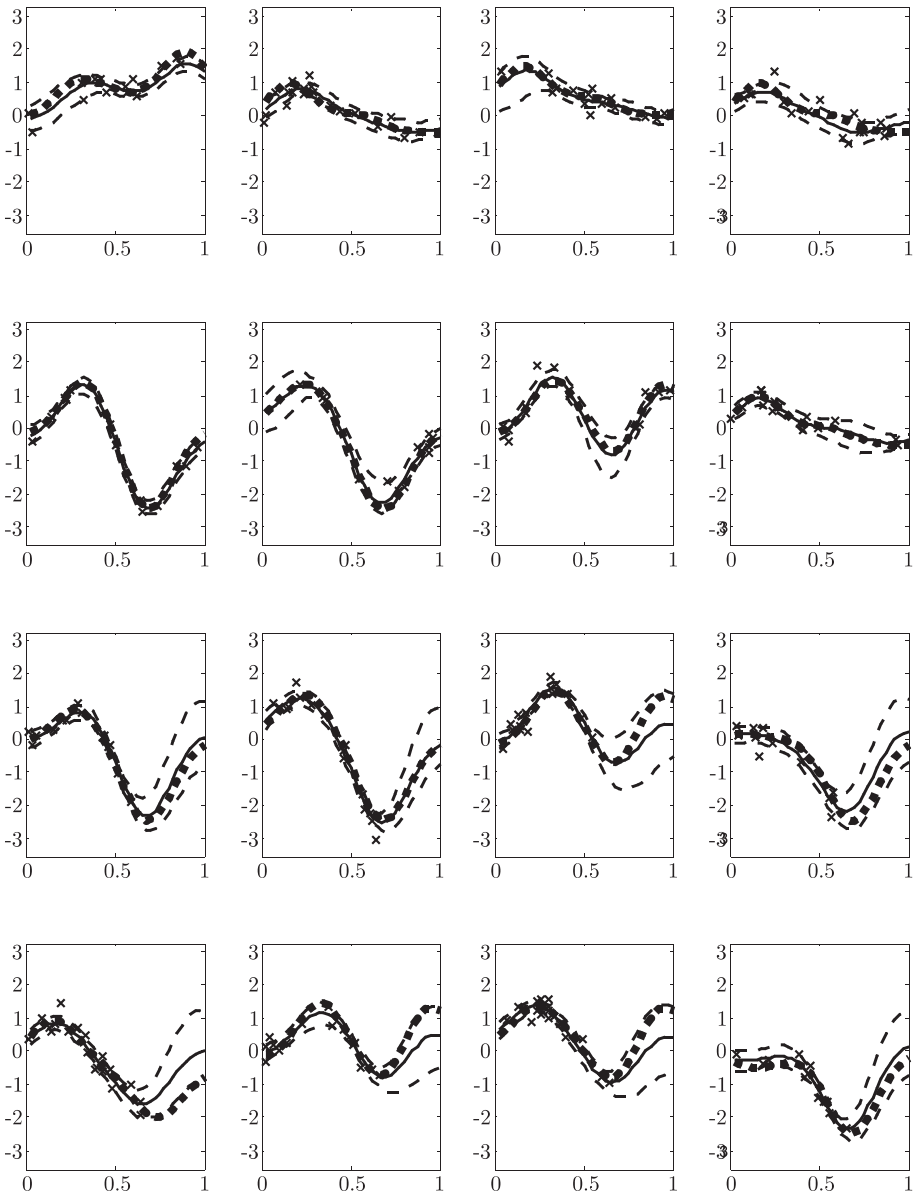

Figure 2. Results for simulation example. The upper 8 plots are for subjects having observations distributed randomly in [0,1], while the lower 8 plots are for subjects having observations in [2/3,1]. The solid lines are posterior mean curves, dashed lines are 95% pointwise credible intervals, and dotted lines are true curves.

value of β suggests the occurrence of a few dominate locations with much higher λ_h values, again leading to a sparse characterization.

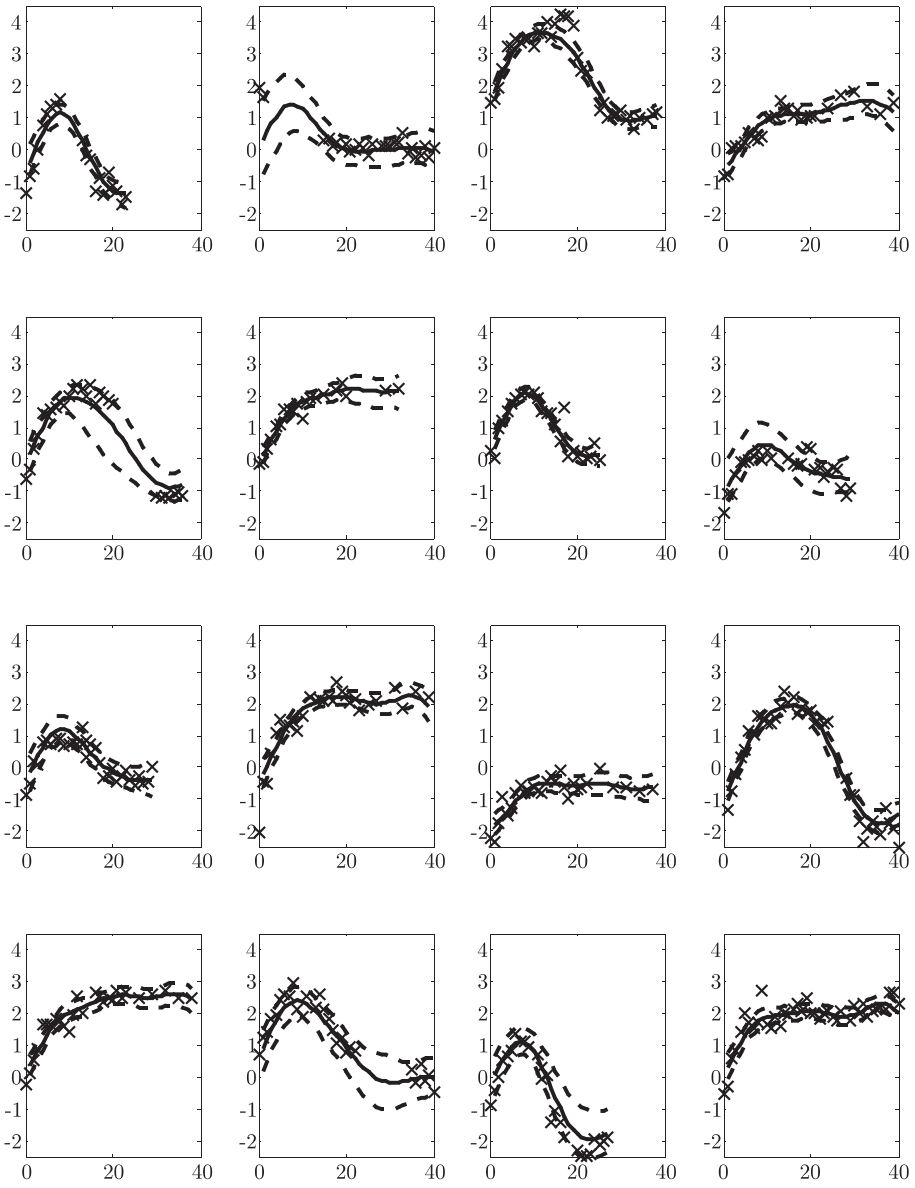

Figure 3. Log(PdG) data and KPP-based function estimates for 16 randomly selected women. The data points are marked with $\times$, the posterior means are solid lines, and 95% credible intervals are dashed lines.

A primary motivation for the KPP over the LPP is that the incorporation of information on the relative locations of the basis functions should allow improvements in prediction. Although the LPP is flexible enough to provide a good fit to complex functional data, one may expect diminished predictive performance

Table 1. Posterior summaries of parameters in KPP analysis of hormone curve data.

	Posterior Summary			
Parameter	Mean	Median	SD	95% CI
α	0.43	0.38	0.26	[0.08, 1.11]
β	0.30	0.30	0.07	[0.18, 0.45]
τ	12.94	12.93	0.84	[11.27, 14.71]
ψ	4.41	4.34	0.94	[2.81, 6.56]
υ	2.06	2.06	0.02	[2.02, 2.11]
κ	30.42	25.61	17.04	[12.26, 78.63]
ϕ^*	4.72	4.00	1.96	[3.00, 9.50]

when there is no observed data available for a subject at times close to the time of interest. To assess predictive performance, we repeated the analysis holding out the last 5 observations for 50 women randomly selected from among the women having at least 10 observations. Figure 4 shows the true values of y_{it} for these held-out observations versus the predicted values, with 95% predictive intervals shown with light dotted lines. The correlation between the true and predicted values was 0.866 and the mean square predictive error was 2.05. Given that hormone trajectories are difficult to predict more than a few days out, this is very good performance. For the first held out observation the correlation was very high at 0.939, while for the second to fifth observations the correlations were 0.898, 0.765, 0.741 and 0.685, respectively.

For comparison, we repeated the analysis using the LPP prior with the same held-out observations. Figure 5 shows the true values of y_{it} versus the predictive values. The results are clearly not as good as those shown in Figure 4 for the KPP, with a small but non-neglible subset of the points moving much further away from the line. The correlation between y_{it} and $\hat{y}_{it}$ diminished to 0.795, the mean square predictive error was 2.340 and the correlations for observations 1-5 were 0.939, 0.853, 0.580, 0.500, and 0.389, respectively. As expected, the LPP has good predictive performance when the subject has data available close to the time of interest, but the performance decays rapidly with an increasing time gap. Repeating the analysis also for a DP prior on the random effects, the mean square predictive error was 4.920 and the correlation between y_{it} and $\hat{y}_{it}$ was 0.681, suggesting substantially worse predictive performance than for either the KPP or LPP.

The computations were implemented on a MacBook Pro laptop with a 2.6 GHz Intel Core 2 Duo processor. For the KPP analysis, MCMC iterations proceeded at a rate of approximately 1/sec using non-optimized Matlab code, so that 10,000 iterations required several hours to run. Convergence was assessed by examining trace plots of the subject-specific functions at a wide variety of

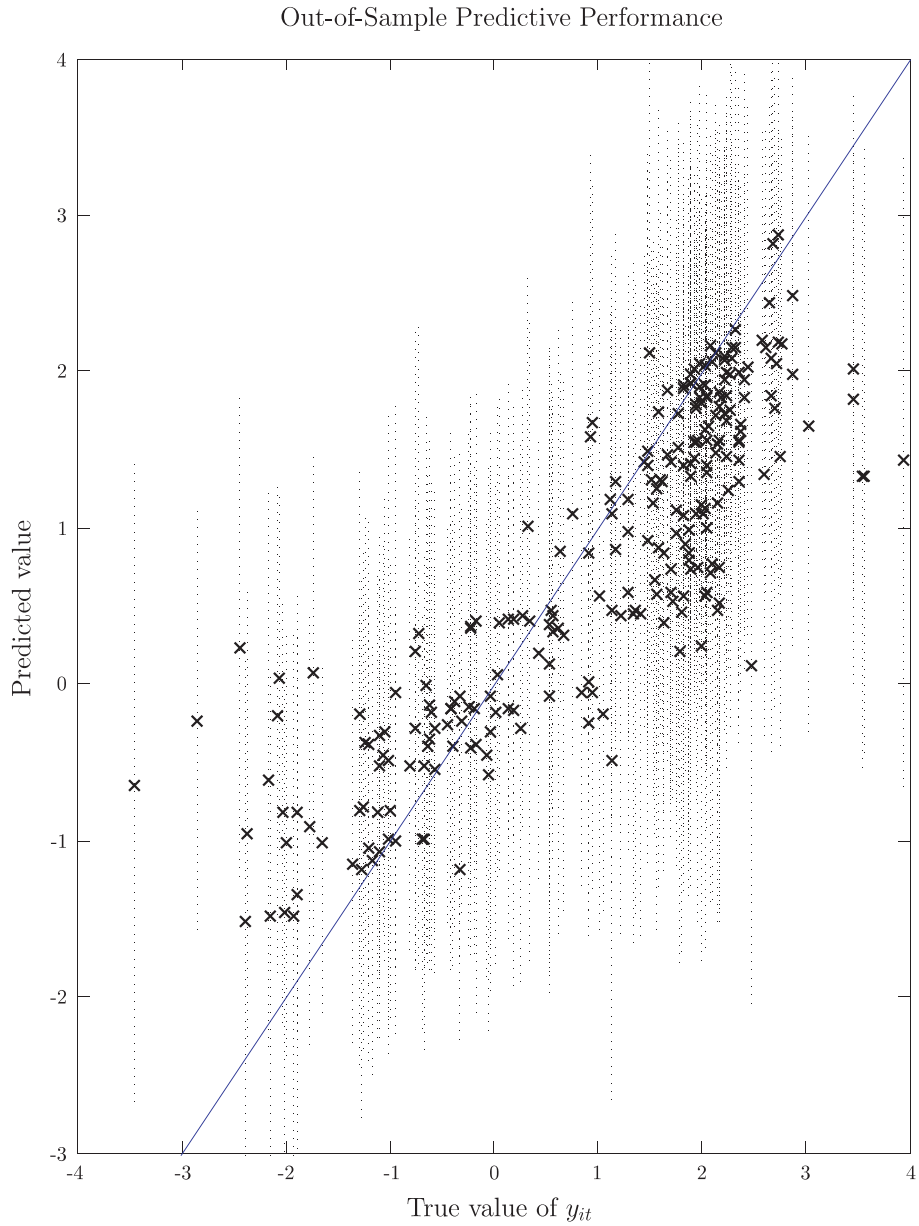

Figure 4. Out-of-sample predictive performance for the KPP. The last 5 log(PdG) observations for 50 women randomly selected from those with 10 or more observations were held out.

times for a random selection of subjects, while also examining trace plots of the hyperparameters in the KPP prior and parameters characterizing the residual distribution. In addition, correlograms were estimated to assess mixing. These

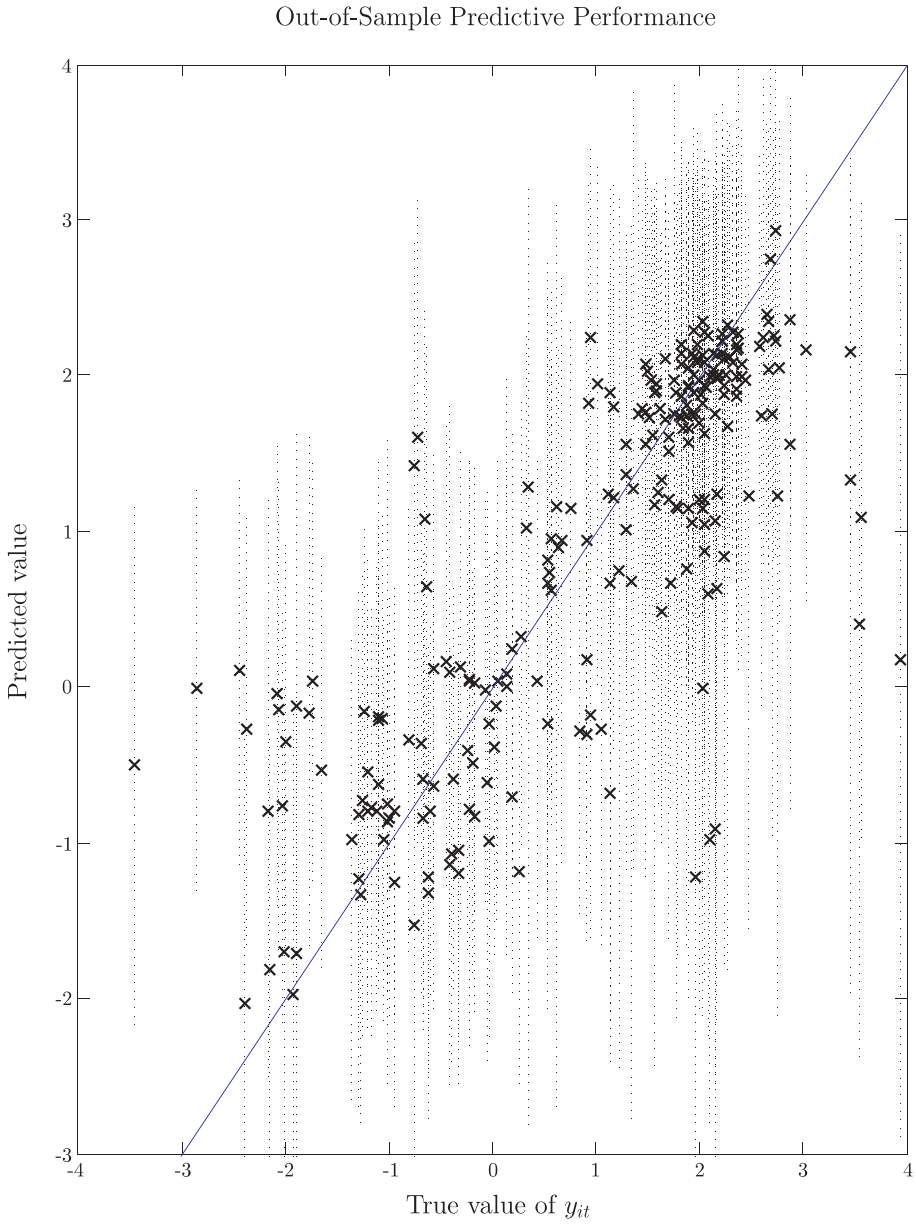

Figure 5. Out-of-sample predictive performance for the LPP (Dunson (2009)). The last 5 log(PdG) observations for 50 women randomly selected from those with 10 or more observations were held out.

exercises showed that apparent convergence occurred rapidly, and autocorrelation functions declined to zero quickly for most of the unknowns, suggesting good mixing.

5.2. Dependent KPP and classification

As motivated in Bigelow and Dunson (2009), it is of interest to use the PdG measurements following ovulation to predict impending early pregnancy loss. For $i = 1, \ldots, 165$, let $l_i = 1$ denote that an early pregnancy loss occurred, with $l_i = 0$ otherwise. Let $i = 1, \ldots, n_0$ index the women in a training sample, and let $i = n_0 + 1, \ldots, n = 165$ index women in a test sample. The overall proportion of early pregnancy losses was $1/n \sum_{i=1}^{n} l_i = 0.29$. To assess the impact of the size of the training sample on predictive performance, I varied $n_0 \in \{85, 65, 45\}$, running each case for two different randomly chosen splits of the data and averaging the results.

To adapt the model implemented in Section 5.1 for prediction, I propose a nonparametric Bayes discriminant analysis approach based on a dependent KPP (DKPP) mixture model. De la Cruz-Mesia, Quintana and Müller (2007) proposed a related discriminant analysis approach for classification based on longitudinal markers, but they used a dependent DP (DDP) instead of a DKPP. To predict l_i given $\mathbf{y}_i$, write

$$\Pr(l_i = 1 \,|\, \mathbf{y}_i) = \frac{\Pr(l_i = 1)\, L(\mathbf{y}_i \,|\, l_i = 1)}{\Pr(l_i = 1)\, L(\mathbf{y}_i \,|\, l_i = 1) + \Pr(l_i = 0)\, L(\mathbf{y}_i \,|\, l_i = 0)},$$

where $\Pr(l_i = 1) = \zeta$ is the marginal probability of an early pregnancy loss, with $\zeta \sim \text{beta}(1, 1)$ to allow uncertainty in this probability under a uniform prior, and

$$L(\mathbf{y} \,|\, l_i) = \prod_{t=1}^{n_i} t_\kappa(y_{it}; \mathbf{b}_{it}'\boldsymbol{\theta}_i, \tau), \quad (\boldsymbol{\theta}_i \,|\, l_i = l) = P^{(l)}, \tag{5.1}$$

is the conditional likelihood of $\mathbf{y}_i$ given $l_i = l$, for $l = 0, 1$. I consider DDP and DKPP priors for the random effects distributions $(P^{(0)}, P^{(1)})$, with the same probability weights but varying coefficient vectors $\boldsymbol{\Theta}_h^{(0)}$ and $\boldsymbol{\Theta}_h^{(1)}$ for $P^{(0)}$ and $P^{(1)}$, respectively. In particular, $\boldsymbol{\Theta}_h^{(1)}$ is expressed as $\boldsymbol{\Theta}_h^{(0)}$ plus a component-specific shift.

The MCMC algorithm of Section 3 is straightforward to modify to accommodate the DDP and DKPP models. However, in implementing the exact block Gibbs sampler for the DDP model, results tended to be quite sensitive to the initial cluster allocation, with a slow rate of mixing between different configurations. To address this problem, I modified the MCMC implementation to incorporate the label-switching moves recommended by Papaspiliopoulos and Roberts (2008). This modification was included for both the DDP and DKPP models to make the results comparable, and led to greatly improved mixing. For each training-test split of the data, the MCMC algorithm was run for 10,000 iterations with a

1,500 iteration burn-in for both the DDP and DKPP analyses. In the DDP analyses, the proportion of test subjects correctly classified was 0.88, 0.84 and 0.86 for $n_0 = 85, 65, 45$, respectively, while the corresponding results for the DKPP were 0.94, 0.96 and 0.92. Hence, the DKPP had consistently better predictive performance even for small training samples.

6. Discussion

This article has proposed a new nonparametric Bayes prior for unknown random effects distributions, allowing for flexible local borrowing of information across subjects through dependent local partitioning. The KPP has clear advantages over previous nonparametric Bayes methods based on global partitioning, such as the Dirichlet process. In particular, the KPP favors a sparser representation of the data in allowing subjects to have clustered random effects, with cluster allocation varying for different random effects. Clustering is viewed as an approach for dimensionality reduction and flexible modeling, though one could perform inferences on the local clustering structure based on the proposed methods.

The proposed approach was illustrated through applications to functional data analysis (FDA) models having an unknown distribution for subject-specific random effects. Behseta, Kass, and Wallstrom (2005) and Kaufman and Sain (2009) instead avoid an explicit basis representation through hierarchical Gaussian processes. Such models do not allow local borrowing of information or clustering, and results may be sensitive to the choice of covariance function. Morris and Carroll (2006) proposed a wavelet-based functional mixed model that incorporates independent Gaussian random effects. Flexible semiparametric Bayes approaches for FDA were recently proposed by MacLehose and Dunson (2009) and Rodriguez, Dunson, and Gelfand (2009). The MacLehose and Dunson (2009) model relies on a kernel convolution of a random signed measure, leading to appealing theoretical properties but challenging computation. Rodriguez, Dunson, and Gelfand (2009) induced an FDA model through a nonparametric Bayes prior for a collection of related multivariate densities, leading to global functional clustering.

In addition to sparse characterization of unknown random effects distributions, there are clear applications of the KPP to multiple changepoint detection and image segmentation. In such settings, it is useful to consider a minor modification of the formulation to replace the vector $\boldsymbol{\Theta}_h = (\Theta_{h1}, \ldots, \Theta_{hp})'$ with a scalar Θ_h, letting $\theta_{ij} = \Theta_{\gamma_{ij}}$, for $j = 1, \ldots, p$. Then, using piecewise constant or linear basis functions, so that $\mathbf{z} = (z_1, \ldots, z_p)'$ is a vector of potential knot locations, a changepoint in f_i occurs at z_j if $\gamma_{ij} \neq \gamma_{ij-1}$. The KPP will automatically borrow information on changepoint locations within- and across-subjects using a more

flexible dependence structure than commonly-used Markov models. In addition, computation is straightforward using the proposed MCMC algorithm.

Acknowledgement

The author thanks two anonymous referees for comments that lead to substantial improvements in the manuscript. This research was partially support by grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).

Appendix

Proof of Lemma 1. Letting $k^{(n)} = \max\{\gamma_i\}_{i=1}^n$ and $n_t = \sum_{i=1}^n 1(\gamma_i = t)$, we have

$$\begin{aligned}\Pr(\gamma_1,\ldots,\gamma_n) &= \mathrm{E}\left[\prod_{t=1}^{k^{(n)}}\left\{\pi_t^*\prod_{s<t}(1-\pi_s^*)\right\}^{n_t}\right]\\ &= \prod_{t=1}^{k^{(n)}}\mathrm{E}\Big\{\big(\pi_t^*\big)^{n_t}\big(1-\pi_t^*\big)^{n-\sum_{s=1}^t n_s}\Big\}\\ &= \prod_{t=1}^{k^{(n)}}\frac{\mathrm{Be}\Big(n_t+1,\sum_{s=t+1}^{k^{(n)}}n_s+\alpha\Big)}{\mathrm{Be}(1,\alpha)},\end{aligned}$$

where $\mathrm{Be}(\cdot)$ is the beta function. Letting $k^{(n+1)} = \max\{k^{(n)}, m\}$, $a_t = n_t + 1$, and $b_t = \sum_{s=t+1}^{k^{(n)}} n_s + \alpha$, we have

$$\Pr(\gamma_1,\ldots,\gamma_n,\gamma_{n+1}=m) = \prod_{t=1}^{k^{(n+1)}}\frac{\mathrm{Be}\Big(a_t+1(m=t),b_t+1(m>t)\Big)}{\mathrm{Be}(1,\alpha)}.$$

Noting that $\mathrm{Be}(a+1,b)/\mathrm{Be}(a,b) = a/(a+b)$ and $\mathrm{Be}(a,b+1)/\mathrm{Be}(a,b) = b/(a+b)$ from properties of the beta function, Lemma 1 is obtained as the ratio of the previous two expressions,

$$\Pr(\gamma_{n+1}=m\,|\,\boldsymbol{\gamma}^{(n)}) = \left\{\prod_{t=1}^{m-1}\left(\frac{b_t}{a_t+b_t}\right)\right\}\frac{a_m}{a_m+b_m},\quad m=1,\ldots,\infty.$$

Proof of Theorem 1. Let $\mathbf{D}^{(n)} = \{\boldsymbol{\theta}^{(n)}, \mathbf{G}^{(n)}, \mathbf{T}^{(n)}\}$, $n_g^{(j)} = \sum_{l=1}^j 1(g_l = g)$ denote the number of type $l \le j$ individuals from household $n+1$ that belong to group g, $\mathbf{T}_{n+1}^{(j)} = \{T_{n+1,g}, g : n_g^{(j)} > 0\}$ denote the tables for the groups formed by individuals of type $l \le j$, $\mathbf{t}^{(j)} = \{t_g, g : n_g^{(j)} > 0\}$, $\mathbf{G}_{n+1}^{(j)} = (G_{n+1,1},\ldots,G_{n+1,j})'$,

and $\mathbf{g}^{(j)} = (g_1, \ldots, g_{j-1})'$. Then, the predictive distribution in Theorem 1 can be factorized as

$$\begin{aligned}
&\big(\boldsymbol{\theta}_{n+1}, \mathbf{G}_{n+1} = \mathbf{g}, \mathbf{T}_{n+1} = \mathbf{t} \,|\, \mathbf{D}^{(n)}\big) \\
&= \prod_{j=1}^{p} \Big\{ \big(\theta_{n+1,j} \,|\, \boldsymbol{\theta}_{n+1}^{(j-1)}, \mathbf{T}_{n+1}^{(j)} = \mathbf{t}^{(j)}, G_{n+1}^{(j)} = \mathbf{g}^{(j)}, \mathbf{D}^{(n)}\big) \\
&\qquad \Pr(T_{n+1,g_j} = t_{g_j} \,|\, \mathbf{T}_{n+1}^{(j-1)}, \mathbf{G}_{n+1}^{(j)} = \mathbf{g}^{(j)}, \mathbf{D}^{(n)}\big) \\
&\qquad \Pr(G_{n+1,j} = g_j \,|\, \mathbf{G}_{n+1}^{(j-1)} = \mathbf{g}^{(j-1)}, \mathbf{D}^{(n)}\big) \Big\},
\end{aligned}$$

Simplifying each of these conditional distributions in turn, we obtain

- $\underline{\big(\theta_{n+1,j} \,|\, \boldsymbol{\theta}_{n+1}^{(j-1)}, \mathbf{T}_{n+1}^{(j)} = \mathbf{t}^{(j)}, G_{n+1}^{(j)} = \mathbf{g}^{(j)}, \mathbf{D}^{(n)}\big)}$ is either P_{0j} if $\sum_{i=1}^{n} 1(T_{iG_{ij}} = t_{g_j}) = 0$, so that table t_{g_j} does not yet have an individual of type j, or is $\delta_{\theta_{ij}}$ for $i : T_{iG_{ij}} = t_{g_j}$.
- $\underline{\Pr\big(T_{n+1,g_j} = t_{g_j} \,|\, \mathbf{T}_{n+1}^{(j-1)}, \mathbf{G}_{n+1}^{(j)} = \mathbf{g}^{(j)}, \mathbf{D}^{(n)}\big)}$. There are two possibilities: (i) one of the previous $1, \ldots, j-1$ individuals in household $n+1$ has been allocated to the same group as individual i, so that $g_j \in \mathbf{G}_{n+1}^{(j-1)}$; or (ii) the jth individual in household $n+1$ is the first member of group g_j to be seated. In case (i), T_{n+1,g_j} is the table assignment for previous members of the group from household $n+1$ with probability one. In case (ii), the table assignment follows the predictive rule of the oCRP shown in Lemma 1, but with the number of past subjects seated at each table replaced with the number of past groups seated at each table including groups $\mathbf{G}^{(n)}$ and $\mathbf{G}_{n+1}^{(j-1)}$. This follows from the form for ϕ_{ig} in expression (2.1).
- $\underline{\Pr\big(G_{n+1,j} = g_j \,|\, \mathbf{G}_{n+1}^{(j-1)} = \mathbf{g}^{(j-1)}, \mathbf{D}^{(n)}\big) = \omega_{jg_j}}$.

Proof of Proposition 1. Proposition 1 can be calculated as the probability that individuals i, j and i, j' choose the same group, which is trivially calculated as $\boldsymbol{\omega}_j' \boldsymbol{\omega}_{j'}$, plus the probability that these individuals choose different groups $(1 - \boldsymbol{\omega}_j' \boldsymbol{\omega}_{j'})$ multiplied by the probability of these groups being assigned to the same table,

$$\mathrm{E}\Big[\sum_{t=1}^{\infty}\Big\{\nu_t^* \prod_{s<t}(1-\nu_s^*)\Big\}^2\Big] = \Big(\frac{1}{1+\alpha}\Big)\Big(\frac{2}{2+\alpha}\Big)\sum_{t=0}^{\infty}\Big(\frac{\alpha}{2+\alpha}\Big)^t = \Big(\frac{1}{1+\alpha}\Big). \quad \text{(A.1)}$$

Here the last identity is a consequence of the infinite geometric series identity. In addition, $\Pr(\gamma_{ij} = \gamma_{i'j})$ is the probability that the jth individual in households i and i' are allocated to groups that choose the same table, which is also $1/(1+\alpha)$.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilitiés de Saint-Flour XII* (Edited by P. L. Hennequin). Springer Lecture Notes in Mathematics **1117**.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803-821.

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

Bigelow, J. L. and Dunson, D. B. (2009). Bayesian semiparametric joint models for functional predictors. *J. Amer. Statist. Assoc.* **104**, 26-36.

Bhuiyan, M. I. H., Ahmad, M. O. and Swamy, M. N. S. (2007). Spatially adaptive wavelet-based method using the Cauchy prior for denoising SAR images. *IEEE Trans. Circuits Systems for Video Technology* **17**, 500-507.

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

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

Chung, H., Lanza, S. T. and Loken, E. (2008). Latent transition analysis: Inference and estimation. *Statist. Medicine* **27**, 1834-1854.

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

De la Cruz-Mesia, R., Quintana, F. A. and Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *Appl. Statist.* **56**, 119-137.

Dunson, D. B. (2009). Nonparametric Bayes local partition models for random effects. *Biometrika* **96**, 249-262.

Dunson, D. B., Pillai, N. and Park, J.-H. (2007). Bayesian density regression. *J. Roy. Statist. Soc. Ser. B* **69**, 163-183.

Dunson, D. B., Xue, Y. and Carin, L. (2008). The matrix stick-breaking process: Flexible Bayes meta analysis. *J. Amer. Statist. Assoc.* **103**, 317-27.

Erosheva, E., Fienberg, S. and Lafferty, J. (2006). Mixed-membership models of scientific publications. *Proc. Nat. Acad. Sci.* **101**, 5220-5227.

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

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

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

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97**, 611-631.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *J. Roy. Statist. Soc. Ser. B* **58**, 155-176.

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: An application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.* **101**, 18-29.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145-168.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161-173.

Ishwaran, H. and James, L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211-1235.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20**, 50-67.

Kaufman, C. and Sain, S. (2009). Bayesian functional ANOVA modeling using Gaussian process prior distributions. Technical Report, 773, Department of Statistics, University of California, Berkeley.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54**, 921-938.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. 1. density estimates. *Ann. Statist.* **12**, 351-357.

MacLehose, R. F. and Dunson, D. B. (2009). Nonparametric Bayes kernel-based priors for functional data analysis. *Statist. Sinica* **19**, 611-629.

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

Müller, P. and Rosner, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Amer. Statist. Assoc.* **92**, 1279-1292.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463-469.

Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet process mixture models. Working Paper **08-20**, Centre for Research in Statistical Methodology, University Warwick, Coventry, UK.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika* **95**, 169-186.

Petrone, S., Guindani, M. and Gelfand, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. Roy. Statist. Soc. Ser. B* **71**, 755-782.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *J. Roy. Statist. Soc. Ser. B* **68**, 305-322.

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96**, 149-162.

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2010). Latent stick-breaking processes. *J. Amer. Statist. Assoc.* **105**, 647-659.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002). Genetic structure of human populations. *Science* **298**, 2381-2385.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639-50.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36**, 45-54.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646-648.

Woodbury, M. A., Clive, J. and Garson, A. (1978). Mathematical typology - Grade of membership technique for obtaining disease definition. *Computers and Biomedical Research* **11**, 277-298.

Yau, C., Papaspiliopoulos, O., Roberts, G. O. and Holmes, C. (2008). Bayesian nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. Working Paper, Oxford-Man Institute, University of Oxford, Oxford, UK.

Department of Statistical Science, Box 90251, 218 Old Chemistry Building, Duke University, Durham, NC 27708-0251, U.S.A.

E-mail: dunson@stat.duke.edu

(Received February 2009; accepted June 2009)