

# Kernel local partition processes for functional data

David B. Dunson

*Department of Statistical Science*

*Box 90251, 218 Old Chemistry Building*

*Duke University*

*Durham, NC 27708-0251*

dunson@stat.duke.edu

**SUMMARY** Functional data analysis commonly relies on the incorporation of basis functions having subject-specific coefficients, with the choice of basis and random effects distribution important. To allow the random effects distribution to be unknown, while inducing subject-specific basis selection and local borrowing of information across subjects, this article proposes a kernel local partition process (KLPP) prior. The KLPP selects the elements in a subject's random effects vector locally from a collection of unique coefficient vectors, leading to a flexible local generalization of the Dirichlet process and to a sparse representation of complex functional data. Basic theoretical properties are considered, an MCMC algorithm is developed for posterior computation and the methods are applied to hormone data.

*Key words:* Basis functions; Dirichlet process; Longitudinal data; Nonparametric Bayes; Random effects; Sparsity.

## 1. Introduction

Functional data analysis considers a random function as the basic unit of analysis for a subject, with the data consisting of error-prone measurements at a (often sparse) number of locations. Some examples of functional data include longitudinal trajectories in a biomarker and brain images. Our particular interest is in sparse functional data measured at unequally-spaced and varying locations for the different subjects. In such settings, a variety of modeling strategies have been proposed, including hierarchical Gaussian processes (Behseta, Kass and Wallstrom, 2005; Kaufman and Sain, 2007) and random effects models relying on basis function representations (Thompson and Rosen, 2008; Bigelow and Dunson, 2007). For Gaussian process (GP) models, the choice of mean and covariance function plays a critical role, while for random effects models the choice of basis and parametric form for the random effects distribution is important.

Let  $f_i : \mathcal{X} \rightarrow \mathfrak{R}$  denote the function for subject  $i$ , for  $i = 1, \dots, n$ . As a more flexible alternative to the GP, one can use a functional Dirichlet process (FDP) in which  $f_i \sim P$  and  $P \sim DP(\alpha P_0)$ , with  $\alpha$  a scalar precision parameter and  $P_0$  a base probability measure obeying a GP law. Under this approach, subjects are allocated to functional clusters, with  $f_i = \Theta_h$  for all subjects in cluster  $h$ , where  $\Theta_h$  is a realization from a GP. By clustering subjects, one obtains replicates so that through posterior updating the function estimates become less sensitive to the choice of GP covariance function. However, as noted by Petrone, Guindani and Gelfand (2008), the FDP has the disadvantage of inducing global functional clusters, which does not allow local clustering of functions that differ only in local regions. They proposed a hybrid FDP (hFDP), which instead allows individual functions to be formed as a patchwork of segments locally selected from a collection of global GP realizations.

An alternative to relying on generalizations of GP priors is to use a basis representation,

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

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 the distribution of the random effects. The two primary questions that arise in using (1) are how to choose  $\mathbf{b}$  and  $P$ . Addressing these questions is challenging, since the appropriate basis to use is typically not known in advance and one may need to allow varying basis functions for different subjects for sufficient flexibility. In addition,  $p$  is often large, so that  $P$  has many parameters, with simple parametric choices (e.g., Gaussian with diagonal covariance) providing a poor representation of the data.

A common strategy for accommodating uncertainty in basis function selection is to pre-specify a large number of potential basis functions, and then choose a shrinkage prior distribution for the basis coefficients. For example, the relevance vector machine (Tipping, 2001) specifies a  $t$ -prior with zero degrees of freedom, and then obtains maximum a posteriori (MAP) estimates. Unfortunately, this choice of prior does not result in a proper posterior. An alternative is to choose a Cauchy prior, which following West (1987) can be expressed as  $\theta_{ij} \sim N(0, \kappa_{ij}^{-1})$ , with  $\kappa_{ij} \sim \text{gamma}(1/2, 1/2)$ . This prior is concentrated at zero with very heavy tails, so that coefficients for unnecessary basis functions are shrunk to zero, while coefficients for important basis functions fall in the tails. To borrow information across coefficients about the degree of shrinkage, let  $\kappa_{ij} = \kappa$ . This approach can be generalized to the functional data analysis setting, while allowing  $P$  to be unknown in (1), by letting

$$P \sim DP(\alpha P_0), \quad P_0 = \bigotimes_{j=1}^p P_0^*, \quad P_0^* \equiv \text{Cauchy}(\kappa), \quad (2)$$

where  $\bigotimes_{j=1}^p P_0^j$  denotes the product measure. A related specification to (1) - (2), but without the Cauchy shrinkage, was proposed by Ray and Mallick (2006) for wavelet-based functional clustering.

Note that specification (1) - (2) leads to global functional clustering, with  $f_i(x) = \mathbf{b}(x)' \boldsymbol{\Theta}_h$  for subjects in cluster  $h$ , where one allows for differential basis function selection automatically through letting the elements of the basis coefficient vector  $\boldsymbol{\Theta}_h = (\Theta_{h1}, \dots, \Theta_{hp})'$

that are set close to zero vary across the clusters. However, the performance of the approach in accurately and sparsely characterizing complex functional data is critically dependent on global clustering. Dunson, Xue and Carin (2008) and Dunson (2008) proposed local generalizations of the Dirichlet process, which allow for more flexible borrowing of information through dependent local clustering. Both of these approaches allow the probability of  $\theta_{ij} = \theta_{i'j}$  to increase conditionally on  $\theta_{ij'} = \theta_{i'j'}$ , but without including information on the relative locations of basis functions  $b_j$  and  $b_{j'}$ .

Accurate interpolations or predictions of  $f_i(x)$  across regions of  $\mathcal{X}$  for which data are not directly available for subject  $i$  can potentially be achieved by borrowing of information from other subjects that are locally similar to  $i$ . In particular, in estimating  $f_i(x)$  it is important to borrow most strongly from subjects  $i'$  that are similar to  $i$  in regions of  $\mathcal{X}$  close to  $x$ . As such a shrinkage structure is not induced by current nonparametric Bayes methods, new methods are needed. This article proposes a class of kernel local partition processes (KLPPs), which can be used to induce priors for unknown random effects distributions in Bayesian hierarchical models. Letting  $\boldsymbol{\theta}_i \sim P$ , the proposed specification induces a prior on  $P$  through letting

$$\begin{aligned}\boldsymbol{\theta}_i &= \boldsymbol{\Theta}\boldsymbol{\gamma}_i, & \boldsymbol{\Theta}_h &\sim P_0, \\ \boldsymbol{\gamma}_i &\sim Q, & Q &\sim \mathcal{Q},\end{aligned}\tag{3}$$

where  $\boldsymbol{\Theta}\boldsymbol{\gamma} = (\Theta_{\gamma_{11}}, \dots, \Theta_{\gamma_{pp}})'$ ,  $\boldsymbol{\Theta}_h = (\Theta_{h1}, \dots, \Theta_{hp})'$  is a unique coefficient vector,  $P_0$  is as defined in (2),  $\boldsymbol{\gamma}_i = (\gamma_{i1}, \dots, \gamma_{ip})'$  is a local allocation vector,  $Q$  is a distribution over  $\{1, \dots, \infty\}^p$ , and  $\mathcal{Q}$  is a distribution on the space of distributions over  $\{1, \dots, \infty\}^p$ . The innovative aspect of the KLPP is the choice of  $\mathcal{Q}$ , which is induced through a kernel specification that allows greater dependence in local allocation for similar basis functions.

Section 2 proposes the KLPP formulation and derives basic properties. 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 Local Partition Process

Letting  $\boldsymbol{\theta}_i \sim P$ , a kernel local partition process (KLPP) prior for  $P$  is specified according to expression (3), with  $Q \sim \mathcal{Q}$  induced through letting

$$\begin{aligned} (\gamma_{ij} | \boldsymbol{\lambda}, \phi_i) &\sim \sum_{h=1}^r \left( \frac{\lambda_h K_\psi(\mathbf{z}_j, \mathbf{z}_h^*)}{\sum_{l=1}^r \lambda_l K_\psi(\mathbf{z}_j, \mathbf{z}_l^*)} \right) \delta_{\phi_{ih}}, \quad j = 1, \dots, p, \\ \lambda_h &\sim \text{gamma}(\beta/r, 1), \quad h = 1, \dots, r, \\ \phi_{ih} &\sim H, \quad h = 1, \dots, r, \quad H \sim \mathcal{H}, \end{aligned} \quad (4)$$

where  $\mathbf{z}_j = (z_{j1}, \dots, z_{jq})' \in \mathcal{Z}$  is a feature vector for the  $j$ th basis function,  $b_j$ , for  $j = 1, \dots, p$ ,  $K_\psi : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]$  is a bounded kernel that depends on precision parameter  $\psi$ ,  $\beta > 0$  is a hyperparameter controlling the distribution of the  $\lambda_h$  weights,  $\phi_{ih}$  is a latent cluster index specific to location  $\mathbf{z}_h^*$ ,  $H$  is a distribution on  $\{1, \dots, \infty\}$  and  $\mathcal{H}$  is a prior on the space of distributions on  $\{1, \dots, \infty\}$ .

Under this specification, a random weight,  $\lambda_h$ , and subject-specific latent cluster index,  $\phi_{ih}$ , are assigned to location  $\mathbf{z}_h^*$ , for  $h = 1, \dots, r$ , with  $\mathbf{z}_h^*$  potentially set equal to  $\mathbf{z}_h$ , for  $h = 1, \dots, r = p$ , though other choices are possible. The local cluster index  $\gamma_{ij}$  specific to the  $j$ th basis function is then set equal to  $\phi_{ih}$  with probability proportional to the location-specific weight,  $\lambda_h$ , multiplied by a kernel weight,  $K_\psi(\mathbf{z}_j, \mathbf{z}_h^*)$ , which decreases with distance between  $\mathbf{z}_j$  and  $\mathbf{z}_h^*$ . This structure was carefully chosen to induce the desired dependence structure in local partitioning.

**Proposition 1.** It follows from (3) and (4) that

$$\begin{aligned} \Pr(\gamma_{ij} = \gamma_{ij'} | \boldsymbol{\lambda}) &= \mathcal{H}(\phi_1 = \phi_2) + \{1 - \mathcal{H}(\phi_1 = \phi_2)\} \left\{ \frac{K(\mathbf{z}_j)' \boldsymbol{\Lambda}^2 K(\mathbf{z}_{j'})}{\boldsymbol{\lambda}' K(\mathbf{z}_j) K(\mathbf{z}_{j'})' \boldsymbol{\lambda}} \right\} \\ &= \rho_{\boldsymbol{\lambda}}(\mathbf{z}_j, \mathbf{z}_{j'}), \quad j' \in \{1, \dots, p\} \setminus j \end{aligned}$$

where  $\mathcal{H}(\phi_1 = \phi_2) = \int_{\mathcal{H}} H(\phi_1 = \phi_2) dH$ ,  $H(\phi_1 = \phi_2)$  denotes the probability of a tie in two independent draws from  $H$ ,  $K(\mathbf{z}) = [K_\psi(\mathbf{z}, \mathbf{z}_1^*), \dots, K_\psi(\mathbf{z}, \mathbf{z}_r^*)]'$  and  $\Lambda^2 = \text{diag}(\lambda_1^2, \dots, \lambda_r^2)$ .

Proposition 1 describes the pairwise dependence structure in the elements of the local partition index vector  $\gamma_i$ . For concreteness, consider a longitudinal data application in which  $b_j$  is a kernel basis function located at time  $z_j$ , for  $j = 1, \dots, p$ . If subjects  $i$  and  $i'$  have identical coefficients for the  $j$ th basis ( $\theta_{ij} = \theta_{i'j}$ ), then these subject's functions are locally similar in a neighborhood around time  $z_j$ . Hence, they should have an increased probability of having identical coefficients for other bases located close to  $z_j$ . Expression (4) induces such a dependence structure through setting  $\gamma_{ij} = \gamma_{ij'}$  and  $\gamma_{i'j} = \gamma_{i'j'}$  with higher probability if  $z_{j'}$  is close to  $z_j$  (as formalized by Proposition 1). Both the kernel,  $K$ , and the hyperparameter,  $\beta$ , are important in controlling the degree of dependence. For small  $\beta$ , the weight vector  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)'$  will have a small number of dominate elements with the remaining values close to zero. This choice will encourage setting  $\gamma_{ij} = \phi_{ih}$  for all  $j$  such that  $\mathbf{z}_j$  is in a neighborhood around  $\mathbf{z}_h^*$ , with  $\mathbf{z}_h^*$  a dominate location having a relative large  $\lambda_h$ . The size of the neighborhood is controlled by  $\beta$  and  $K$ , with smaller  $\beta$  and higher variance kernels encouraging larger neighborhoods.

**Proposition 2.** It follows from (3) and (4) that

$$\Pr(\gamma_{ij} = \gamma_{i'j'} \mid \boldsymbol{\lambda}) = \mathcal{H}(\phi_1 = \phi_2), \quad i' \neq i, j' = 1, \dots, p,$$

To demonstrate the flexibility of the KLPP and obtain further insight into the structure, it is useful to consider some special cases. When  $\mathcal{H}(\phi_1 = \phi_2) = 1$ , it is straightforward to show that  $P = \delta_{\Theta}$ , with  $\Theta \sim P_0$ , so that  $\theta_i = \Theta$  for all  $i$  and all subjects have identical functions. In the case in which  $\mathcal{H}(\phi_1 = \phi_2) < 1$  and  $K_\psi(\mathbf{z}_j, \mathbf{z}_h^*) = 1(j = h)$  with  $r = p$ ,  $\Pr(\gamma_{ij} = \gamma_{ij'}) = \mathcal{H}(\phi_1 = \phi_2) = \Pr(\gamma_{ij} = \gamma_{i'j'})$ , so that within- and across-subject dependence

is identical. At the other extreme in which  $\mathcal{H}(\phi_1 = \phi_2) < 1$  and  $K_\psi(\mathbf{z}_j, \mathbf{z}_h^*) = 1$  for all  $j, h$ ,

$$\Pr(\gamma_{ij} = \gamma_{ij'} | \boldsymbol{\lambda}) = \mathcal{H}(\phi_1 = \phi_2) + \{1 - H(\phi_1 = \phi_2)\} \frac{\mathbf{1}'_r \boldsymbol{\Lambda}^2 \mathbf{1}_r}{(\boldsymbol{\lambda}' \mathbf{1}_r)^2},$$

so that within-subject dependence is greater than across-subject dependence by a factor that does not depend on  $j$ . In this case,  $\Pr(\gamma_{ij} = \phi_{ih}) = \nu_h$ , with  $\nu_h = \lambda_h / \sum_l \lambda_l$  and  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_r)' \sim \text{Diri}(\beta/r, \dots, \beta/r)$ . If we also let  $\beta \rightarrow 0$ , we obtain  $\gamma_{ij} = \gamma_i \sim H$ , for  $j = 1, \dots, p$ , with  $H \sim \mathcal{H}$ . This special case includes a very broad class of priors, including the Dirichlet process, the two-parameter Poisson-Dirichlet process, and the class of stick-breaking priors considered by Ishwaran and James (2001). For example, letting

$$H = \sum_{l=1}^{\infty} V_l \prod_{m<l} (1 - V_m) \delta_l, \quad V_l \sim \text{beta}(1, \alpha), \quad (5)$$

results in  $P \sim DP(\alpha P_0)$  from the Sethuraman (1994) stick-breaking representation.

When  $\mathcal{H}$  is chosen as in (5),  $\mathcal{H}(\phi_1 = \phi_2) = 1/(1+\alpha)$ . I let  $\boldsymbol{\theta}_i \sim P$ ,  $P \sim KLPP(\alpha, \beta, \psi, P_0)$  as shorthand notation for the kernel local partition process specified by expressions (3) - (5), with  $K_\psi(\mathbf{z}, \mathbf{z}') = \exp(-\psi \|\mathbf{z} - \mathbf{z}'\|^2)$ . To allow the data to inform about the local partitioning structure, I recommend choosing hyperpriors for  $\alpha, \beta, \psi$ . In simulating from the prior for  $\rho_{\boldsymbol{\lambda}}(\mathbf{z}_j) = [\rho_{\boldsymbol{\lambda}}(\mathbf{z}_j, \mathbf{z}_1^*), \dots, \rho_{\boldsymbol{\lambda}}(\mathbf{z}_j, \mathbf{z}_r^*)]'$ , for moderately large  $r$  and a variety of choices of  $\alpha, \beta, \psi$ , it is clear that the local dependence structure is highly-flexible, allowing a rich variety of shapes. Figure 1 provides some representative draws.

**Proposition 3.** It follows from (3) - (5) that

$$\begin{aligned} & \Pr(\gamma_{ij} = \gamma_{i'j} | \gamma_{ij-l} = \gamma_{i'j-l}, \boldsymbol{\lambda}, \alpha, \psi) \\ & = \kappa_j \kappa_{j-l} + \left( \frac{2}{2 + \alpha} \right) (\kappa_j \bar{\kappa}_{j-l} + \bar{\kappa}_j \kappa_{j-l}) + \left( \frac{6 + \alpha}{6 + 5\alpha + \alpha^2} \right) \bar{\kappa}_j \bar{\kappa}_{j-l}, \end{aligned}$$

where  $\kappa_j = K(\mathbf{z}_j)' \boldsymbol{\Lambda}^2 K(\mathbf{z}_j) / \{\boldsymbol{\lambda}' K(\mathbf{z}_j)\}^2$  and  $\bar{\kappa}_j = 1 - \kappa_j$  as shorthand notation.

Proposition 3 provides some insight into the borrowing of information that occurs in local partitioning under the KLPP prior. Conditionally on the information that subject's

$i$  and  $i'$  have the same coefficient for basis function  $j - l$ , the probability of these subjects having the same coefficient for basis function  $j$  will be increased by an amount that depends on the hyper-parameter  $\alpha$  and on the lag  $l$ , with the impact of the lag dependent on the weights  $\lambda$  and kernel precision  $\psi$ . The result is a highly-flexible, adaptive local dependence structure. Figure 2 provides a flavor of the flexibility through simulating from the prior for selected hyperparameter values. By changing the hyperparameter values, one obtains widely different shapes for the conditional clustering probability curves, with the dependence potentially non-monotone in the lag  $l$  and changing according to the location  $j$ . This allows long-range dependence, and the occurrence of critical windows in which the cluster allocation at an important early time predicts the cluster allocation at a much later time. Such critical windows are common in epidemiology and other applications. Note that the local partition process (Dunson, 2008) and matrix stick-breaking process (Dunson et al., 2008) would produce horizontal lines in Figure 2 regardless of the hyperparameter values.

### 3. Posterior Computation

For posterior computation, I propose a Markov chain Monte Carlo (MCMC) algorithm, which is a hybrid of data augmentation, the exact block Gibbs sampler of Papaspiliopoulos (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 (Papaspiliopoulos and Roberts, 2008) and the slice sampler (Walker, 2007).

For concreteness, I focus on a functional data analysis model with  $y_{it} \sim t_\nu(f_i(x_{it}), \tau)$ , where  $t_\nu(\mu, \tau)$  denotes the  $t$ -density centered on  $\mu$ , with  $\nu$  degrees of freedom and scale parameter  $\tau$ . In addition,  $f_i(x)$  follows (1) with  $\theta_i \sim P$  and  $P \sim KLPP(\alpha, \beta, \psi, P_0)$ , where  $P_0 = \otimes_{j=1}^p P_0^*$  and  $P_0^*$  denotes a Cauchy prior centered on zero. The Cauchy prior is an

appealing choice for robust shrinkage of the basis coefficients, as small coefficients will tend to be shrunk to very close to zero, while larger coefficients fall in the tails (Bhuiyan et al., 2007). To complete a Bayes specification, let  $\alpha \sim \text{gamma}(a_\alpha, b_\alpha)$ ,  $\beta \sim \text{gamma}(a_\beta, b_\beta)$ ,  $\psi \sim \text{gamma}(a_\psi, b_\psi)$ ,  $\nu \sim \text{gamma}(a_\nu, b_\nu)$ , 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  $y_{it} \sim N(\mathbf{b}'_{it}\boldsymbol{\theta}_i, \tau^{-1}v_{it}^{-1})$ ,  $\mathbf{b}_{it} = \mathbf{b}(x_{it})$ ,  $v_{it} \sim \text{gamma}(\nu/2, \nu/2)$ ,  $P_0^* = N(0, \kappa^{-1})$  and  $\kappa \sim \text{gamma}(1/2, 1/2)$ .

Letting  $S_{ij} \in \{1, \dots, r\}$  denote the location  $\gamma_{ij}$  is allocated to, the algorithm proceeds as follows:

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

$$\Pr(S_{ij} = h | -) = \frac{\pi_{jh} \prod_{t=1}^{n_i} N(y_{it}; \mathbf{b}'_{it} \boldsymbol{\Theta} \boldsymbol{\gamma}_i(S_{ij}=h), \tau^{-1}v_{it}^{-1})}{\sum_{l=1}^r \pi_{jl} \prod_{t=1}^{n_i} N(y_{it}; \mathbf{b}'_{it} \boldsymbol{\Theta} \boldsymbol{\gamma}_i(S_{ij}=l), \tau^{-1}v_{it}^{-1})}, \quad h = 1, \dots, r, \quad (6)$$

where  $\pi_{jh}$  is the prior probability of allocating  $\gamma_{ij}$  to location  $h$  in the top line of (4) and  $\boldsymbol{\Theta} \boldsymbol{\gamma}_i(S_{ij}=h)$  equals  $(\Theta_{\gamma_{i1}}, \dots, \Theta_{\gamma_{ip}})'$  but with the current value of  $\gamma_{ij}$  set to  $\phi_{ih}$ .

2. To update  $\lambda_h$ , for  $h = 1, \dots, r$ , use a data augmentation approach related to Holmes and Held (2006) and Dunson, Pillai and Park (2007). Letting  $K_{jh} = K_\psi(\mathbf{z}_j, \mathbf{z}_h^*)$  and  $K_{jh}^* = K_{jh} / (\sum_{l \neq h} \lambda_l K_{jl})$ , the conditional likelihood for  $\lambda_h$  is

$$L(\lambda_h) = \prod_{j=1}^p \left( \frac{\lambda_h K_{jh}^*}{1 + \lambda_h K_{jh}^*} \right)^{\sum_i 1(S_{ij}=h)} \left( \frac{1}{1 + \lambda_h K_{jh}^*} \right)^{\sum_i 1(S_{ij} \neq h)},$$

which can be obtained via  $1(S_{ij} = h) = 1(Z_{ijh}^* > 0)$ , with  $Z_{ijh}^* \sim \text{Poisson}(\lambda_h \xi_{ijh} K_{jh}^*)$  and  $\xi_{ijh} \sim \text{exp}(1)$ . Update  $\{Z_{ijh}^*, \xi_{ijh}\}$  and  $\{\lambda_h\}$  in Gibbs steps:

- (a) Let  $Z_{ijh}^* = 0$  if  $S_{ij} \neq h$  and otherwise  $Z_{ijh}^* \sim \text{Poisson}(\lambda_h \xi_{ijh} K_{jh}^*) 1(Z_{ijh}^* > 0)$ .
- (b)  $\xi_{ijh} \sim \text{gamma}(1 + Z_{ijh}^*, 1 + \lambda_h K_{jh}^*)$ .
- (c)  $\lambda_h \sim \text{gamma}(\beta/r + \sum_{i,j} Z_{ijh}^*, 1 + \sum_{i,j} \xi_{ijh} K_{jh}^*)$ .

3. Update  $\psi, \beta, \nu$  using Metropolis steps.

4. Update  $v_{it}$ ,  $i = 1, \dots, n$ ,  $t = 1, \dots, n_i$ , and  $\tau$  by sampling from gamma full conditionals

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

5. Implement exact block Gibbs sampler steps:

(a) Sample  $u_{ih} \sim \text{uniform}(0, \omega_{\phi_{ih}})$ , for  $i = 1, \dots, n$ , with  $\omega_l = V_l \prod_{m < l} (1 - V_m)$ .

(b) Sample the stick-breaking random variables,

$$V_l \sim \text{beta}\left(1 + \sum_{h=1}^r \sum_{i=1}^n 1(\phi_{ih} = l), \alpha + \sum_{h=1}^r \sum_{i=1}^n 1(\phi_{ih} > l)\right),$$

for  $l = 1, \dots, \phi^*$ , with  $\phi^*$  the minimum value satisfying  $\omega_1 + \dots + \omega_{\phi^*} > 1 - \min\{u_{ih}\}$ .

(c) Update  $\boldsymbol{\Theta}_l$ , for  $l = 1, \dots, \phi^*$ , from  $N_p(\widehat{\boldsymbol{\Theta}}_l, \Sigma_{\boldsymbol{\Theta}_l})$  with

$$\widehat{\boldsymbol{\Theta}}_l = \Sigma_{\boldsymbol{\Theta}_l} \sum_{i=1}^n \sum_{t=1}^{n_i} \tau v_{it} \Gamma_{il} \mathbf{b}_{it} (y_{it} - \mathbf{b}'_{it} \Gamma_{i(-l)} \boldsymbol{\theta}_i), \quad \Sigma_{\boldsymbol{\Theta}_l} = \left( \kappa \mathbf{I}_p + \sum_{i=1}^n \sum_{t=1}^{n_i} \tau v_{it} \Gamma_{il} \mathbf{b}_{it} \mathbf{b}'_{it} \Gamma_{il} \right),$$

where  $\mathbf{I}_p$  is the  $p \times p$  identity matrix,  $\Gamma_{il} = \text{diag}(1(\gamma_{i1} = l), \dots, 1(\gamma_{ip} = l))$  and  $\Gamma_{i(-l)} = \text{diag}(1(\gamma_{i1} \neq l), \dots, 1(\gamma_{ip} \neq l))$ .

(d) Update  $\phi_{ih}$ ,  $i = 1, \dots, n$ ,  $h = 1, \dots, r$  from the multinomial conditional with

$$\Pr(\phi_{ih} = l | -) \propto 1(u_{il} < w_l) \prod_{t=1}^{n_i} N\left(y_{it}^{(h)}; \sum_{j=1}^p 1(S_{ij} = h) b_{itj} \boldsymbol{\theta}_{lj}, \tau^{-1} v_{it}^{-1}\right), \quad l = 1, \dots, \phi^*,$$

where  $y_{it}^{(h)} = y_{it} - \sum_{j=1}^p 1(S_{ij} \neq h) b_{itj} \boldsymbol{\theta}_{lj}$ .

6. Update  $\alpha$  by sampling from the conditional posterior

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

7. Update  $\kappa$  by sampling from the conditional posterior

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

This algorithm is straightforward to program and has exhibited good rates of convergence and mixing in simulated and real data applications I have considered. 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 start by considering a simulation example. I assumed  $n = 100$  and simulated data under the functional data analysis model described in Section 3, with  $n = 100$ ,  $\nu = 10$ ,  $\tau = 20$ ,  $\mathcal{X} = (0, 1)$ ,  $b_j(x) = 1$ ,  $b_{j+1}(x) = \exp\{-25(x - x_j^*)^2\}$  and  $x_j = (j - 1)/19$ , for  $j = 1, \dots, 19$ . In addition, I let  $n_i = 10$  plus a discrete uniform random variable on  $[1, \dots, 10]$ , for  $i = 1, \dots, n$ , 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  $\Theta_h = (\Theta_{h1}, \dots, \Theta_{h20})'$ , 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  $j = 1, \dots, p$ , I simulated the elements of  $\gamma_i = (\gamma_{i1}, \dots, \gamma_{ip})'$  from a Markov chain, with  $\gamma_{i1}$  set to a random element of  $\{1, 2, 3\}$ ,  $\gamma_{ij} = \gamma_{i,j-1}$  with probability 0.9 and  $\gamma_{ij}$  otherwise set to a random element of  $\{1, 2, 3\}$ .

To implement our Bayesian analysis, the approach of Section 3 was applied after choosing priors for each of the parameters by letting  $a_\alpha = b_\alpha = 1$ ,  $a_\beta = b_\beta = 1$ ,  $a_\psi = 25$ ,  $b_\psi = 1$ ,  $a_\nu = 4$ ,  $b_\nu = 1$ , 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. In the absence of prior knowledge of the scale of  $y$ , one can standardize the data. In addition, I let  $K_\psi(z, z') = \exp\{-\psi(z - z')^2\}$  and  $x_j^* = z_j^*$ , for  $j = 1, \dots, 19$ . Note that these values of  $x_j^*$ ,  $z_j^*$ , 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  $\kappa$  and  $\beta$ . 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, I observed rapid rates of apparent convergence and mixing. Note that it is important to avoid diagnosing convergence based on examination of the unique coefficient vectors,  $\Theta_h$ , due to the well-known label switching issue, which 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. The focus of this article is on using local partitioning as a highly-flexible approach for sparse modeling of unknown random effects distributions underlying functional data and not on clustering as a focus of the analysis.

Figure 3 shows the data, estimated posterior mean curves (solid lines), 95% pointwise credible intervals (dashed lines) and true curves (dotted lines) for subjects 1, . . . , 8 (top 8 panels) and subjects 51, . . . , 58 (bottom 8 panels). Recall that subjects 1, . . . , 50 have observations throughout the  $[0, 1]$  interval, while subjects 51, . . . , 100 have no observations in the  $[2/3, 1]$  interval. For subjects 1, . . . , 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, . . . , 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 (2008), the results are very good at locations close to data points. However, interpolations and predictions are not as good as for the KLPP. For example, for

the subject in the (2,3) panel there is a substantial gap in the observations. Across this gap, the 95% credible intervals are much wider in the LPP analysis. In addition, across the  $[2/3, 1]$  region for many of the subjects 51,  $\dots$ , 100, the LPP estimates are substantially farther from the truth than the KLPP-based estimates. The mean square error for the LPP is 0.183.

## 5. Application to Hormone Curve Data

I consider progesterone curve data in early pregnancy previously analyzed by Dunson (2008) 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 KLPP prior for the random effects distribution, I implemented the approach used in Section 4, with the same prior specification. As in the simulation examples, rates of apparent convergence and mixing were good.

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 the same 16 randomly selected women shown in Figures 2 and 3 of Dunson (2008). There is a small but notable improvement in fit in using the KLPP 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  $\alpha$  value suggests that a small number of unique coefficient vectors are sufficient to characterize the data. Indeed, the estimate of  $\phi^*$  was 4.72, implying that the basis coefficient vectors for all 165 women are constituted of elements selected from  $\sim 5$  unique coefficient vectors.

In addition, the small value of  $\beta$  suggests the occurrence of a few dominate locations with much higher  $\lambda_h$  values, again leading to a sparse characterization.

A primary motivation for the KLPP over the LPP is that the incorporation of information on the relative locations of the basis functions should allow substantial improvements in prediction. Although the LPP is flexible enough to provide a good fit to complex functional data, one may expect diminished predictive performance when there is no observed data available for a subject at times close to the time of interest. To assess predictive performance, I 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, I 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 KLPP, with a substantial 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 KLPP or LPP.

## 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 proposed kernel local partition process (KLPP) prior is particularly motivated by functional data analysis applications, including longitudinal data and image analysis. The KLPP has clear advantages over previous nonparametric Bayes methods based on global partitioning, such as the Dirichlet process. In particular, the KLPP favors a sparser representation of the data in allowing subjects to have identical basis coefficients without forcing their functions to be identical. Although functional clustering is a useful exploratory tool, it seems unlikely that functions for two different subjects are exactly identical, though the functions may be locally similar within particular regions. My goal is not to obtain functional clusters but to use local partitioning as a tool for sparsely characterizing complex functional data, while facilitating borrowing of information in a flexible manner. The KLPP has major advantages over recently proposed local partition processes (Dunson, 2008; Dunson et al., 2008) in incorporating information on the relative locations of the basis functions.

In addition to sparse characterization of unknown random effects distributions, there are clear applications of the KLPP to multiple changepoint detection and image segmentation. In such settings, it is useful to consider a minor modification of the formulation in expression (3) to replace the vector  $\Theta_h = (\Theta_{h1}, \dots, \Theta_{hp})'$  with a scalar  $\Theta_h$ , letting  $\theta_{ij} = \Theta_{\gamma_{ij}}$ , for  $j = 1, \dots, p$ . Then, using piecewise constant or linear basis functions, so that  $\mathbf{z} = (z_1, \dots, z_p)'$  is a vector of potential knot locations, a changepoint in  $f_i$  occurs at  $z_j$  if  $\gamma_{ij} \neq \gamma_{ij-1}$ . The KLPP 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.

## Appendix

### Proof of Propositions 1 - 3

Under (4),  $\gamma_{ij} = \gamma_{ij'}$  if  $\gamma_{ij}$  and  $\gamma_{ij'}$  are allocated to the same  $\phi_{ih}$ , which occurs with probability

$$\begin{aligned} & \sum_{h=1}^r \Pr(\gamma_{ij} = \phi_{ih} \mid \boldsymbol{\lambda}) \Pr(\gamma_{ij'} = \phi_{ih} \mid \boldsymbol{\lambda}) \\ &= \sum_{h=1}^r \left( \frac{\lambda_h K_\psi(\mathbf{z}_j, \mathbf{z}_h^*)}{\sum_{l=1}^r \lambda_l K_\psi(\mathbf{z}_j, \mathbf{z}_l^*)} \right) \left( \frac{\lambda_h K_\psi(\mathbf{z}_{j'}, \mathbf{z}_h^*)}{\sum_{l=1}^r \lambda_l K_\psi(\mathbf{z}_{j'}, \mathbf{z}_l^*)} \right), \end{aligned}$$

or if  $\gamma_{ij}$  and  $\gamma_{ij'}$  are allocated to different but equal  $\phi_{ih}$ , which occurs with probability

$$\left\{ 1 - \sum_{h=1}^r \left( \frac{\lambda_h K_\psi(\mathbf{z}_j, \mathbf{z}_h^*)}{\sum_{l=1}^r \lambda_l K_\psi(\mathbf{z}_j, \mathbf{z}_l^*)} \right) \left( \frac{\lambda_h K_\psi(\mathbf{z}_{j'}, \mathbf{z}_h^*)}{\sum_{l=1}^r \lambda_l K_\psi(\mathbf{z}_{j'}, \mathbf{z}_l^*)} \right) \right\} \mathcal{H}(\phi_1 = \phi_2).$$

Proposition 1 then follows after simple algebra. Proposition 2 is a trivial consequence of (4).

To show proposition 3, first note that there are three cases that can result in  $C : \gamma_{ij} = \gamma_{i'j}, \gamma_{ij-l} = \gamma_{i'j-l}$ : (i)  $A \cap B$ , (ii)  $(A \cap \bar{B}) \cup (\bar{A} \cap B)$ ; and (iii)  $\bar{A} \cap \bar{B}$ , where  $A : \gamma_{ij} = \gamma_{ij-l} = \phi_{ih}$  for some  $h \in \{1, \dots, r\}$ ,  $B : \gamma_{i'j} = \gamma_{i'j-l} = \phi_{i'h}$  for some  $h \in \{1, \dots, r\}$ , and  $\bar{A}$  and  $\bar{B}$  are the complements of  $A$  and  $B$ , respectively. As a straightforward consequence of properties of the Dirichlet process, the probability of  $C$  in cases (i) - (iii) are, respectively,

$$\frac{1}{1 + \alpha}, \quad \frac{2}{(1 + \alpha)(2 + \alpha)}, \quad \frac{6 + \alpha}{(1 + \alpha)(2 + \alpha)(3 + \alpha)}.$$

Summing these probabilities weighted by the probabilities of (i)-(iii), which can be obtained in a similar manner to proposition 1, and dividing by  $1/(1+\alpha)$  from proposition 2, proposition 3 is obtained.

## References

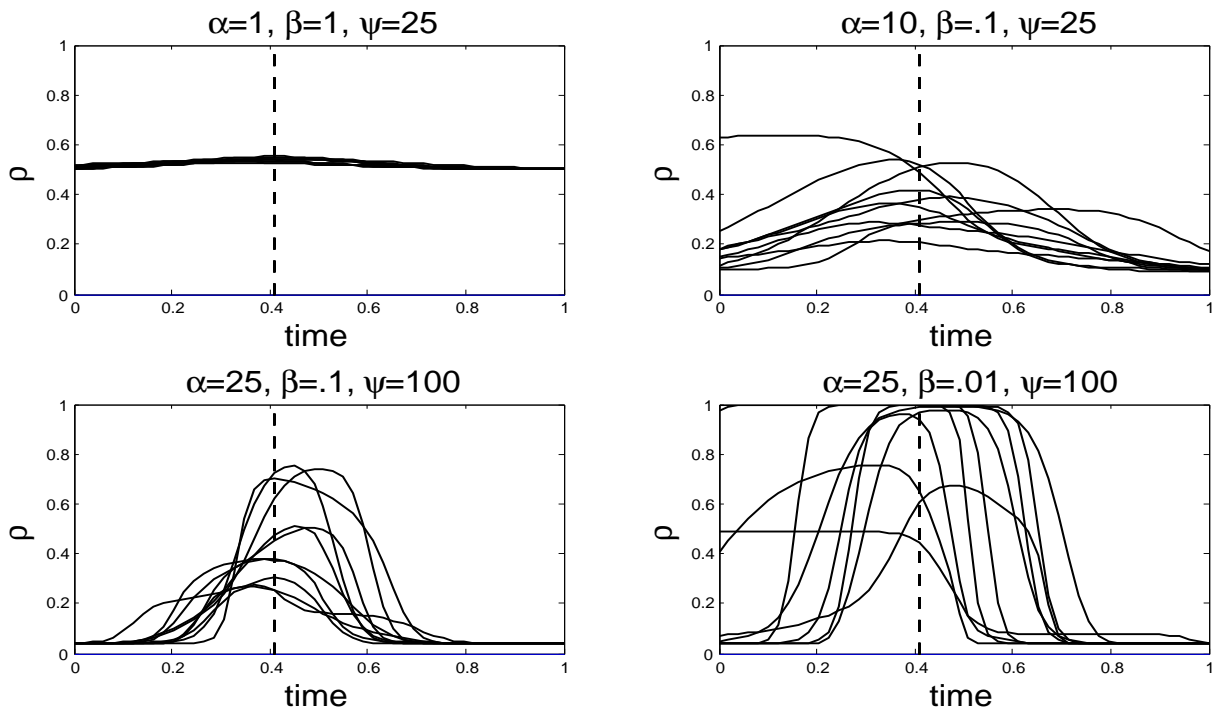
- Behseta, S., Kass, R.E. and Wallstrom, G.L. (2005). Hierarchical models for assessing variability among functions. *Biometrika*, **92**, 419-434.
- Bigelow, J.L. and Dunson, D.B. (2007). Bayesian adaptive regression splines for hierarchical data. *Biometrics*, **63**, 724-732.

- 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 Transactions on Circuits and Systems for Video Technology*, **17**, 500-507.
- Dunson, D.B. (2008). Nonparametric Bayes local partition models for random effects. *Biometrika*, revision submitted.
- Dunson, D.B., Pillai, N. and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society B*, **69**, 163-183.
- Dunson, D.B., Xue, Y. and Carin, L. (2008). The matrix stick-breaking process: Flexible Bayes meta analysis. *Journal of the American Statistical Association*, **103**, 317-27.
- 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. *Journal of the American Statistical Association*, **96**, 161-173.
- Jasra, A., Holmes, C.C. and Stephens, D.A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20**, 50-67.
- Kaufman, C. and Sain, S. (2007). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Journal of Computational and Graphical Statistics*, submitted.
- 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.

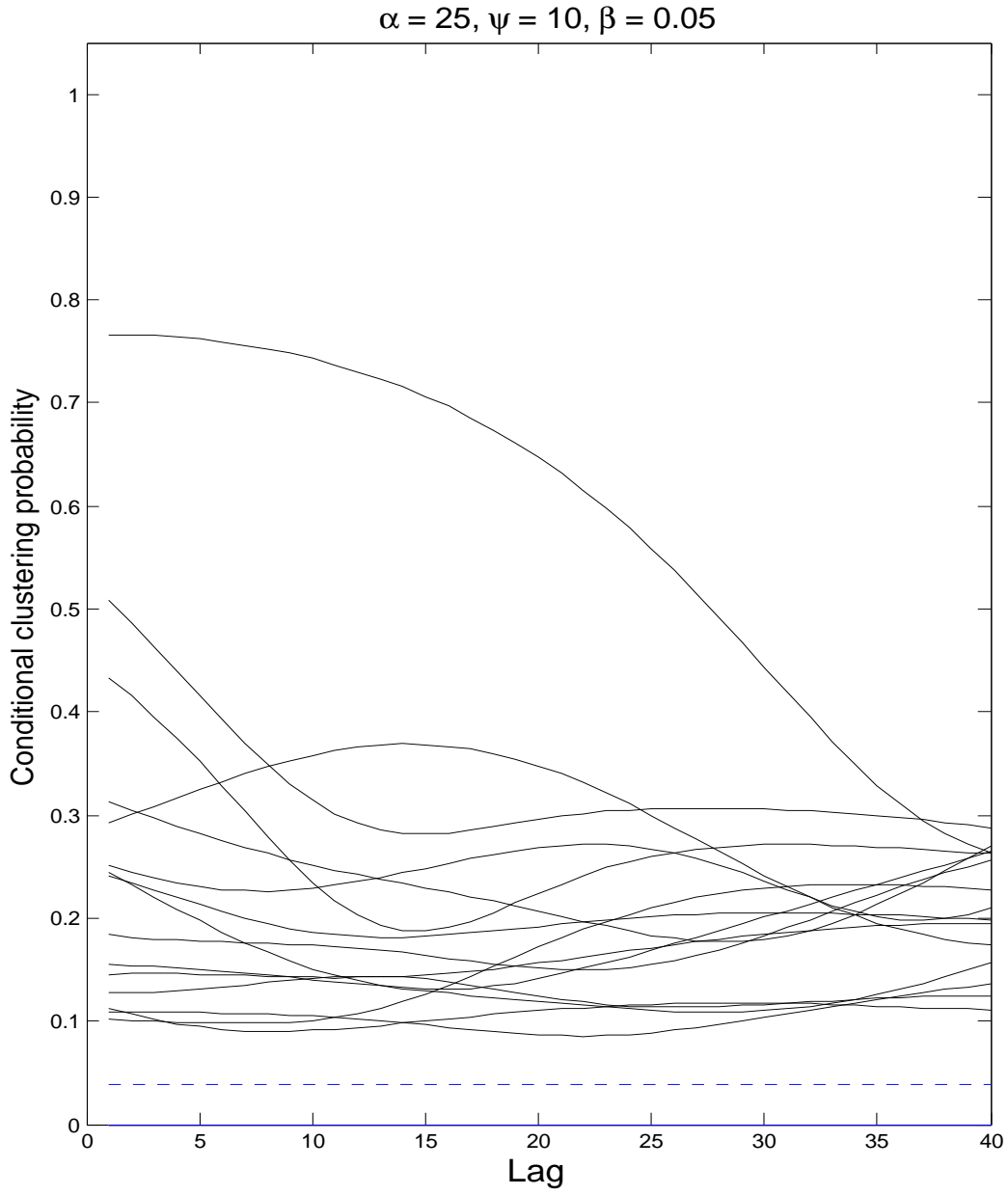
- Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika* 95, 169-186.
- Petrone, S., Guindani, M. and Gelfand, A.E. (2008). Hybrid Dirichlet processes for functional data. *Journal of the Royal Statistical Society B*, revision invited.
- Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society B*, **68**, 305-322.
- Rodriguez, A., Dunson, D.B. and Gelfand, A.E. (2008). Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, to appear.
- Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**, 211-244.
- Thompson, W. and Rosen, O. (2008). A Bayesian model for sparse functional data. *Biometrics*, **64**, 54-63.
- Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, 36, 45-54.
- West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, **74**, 646-648.

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

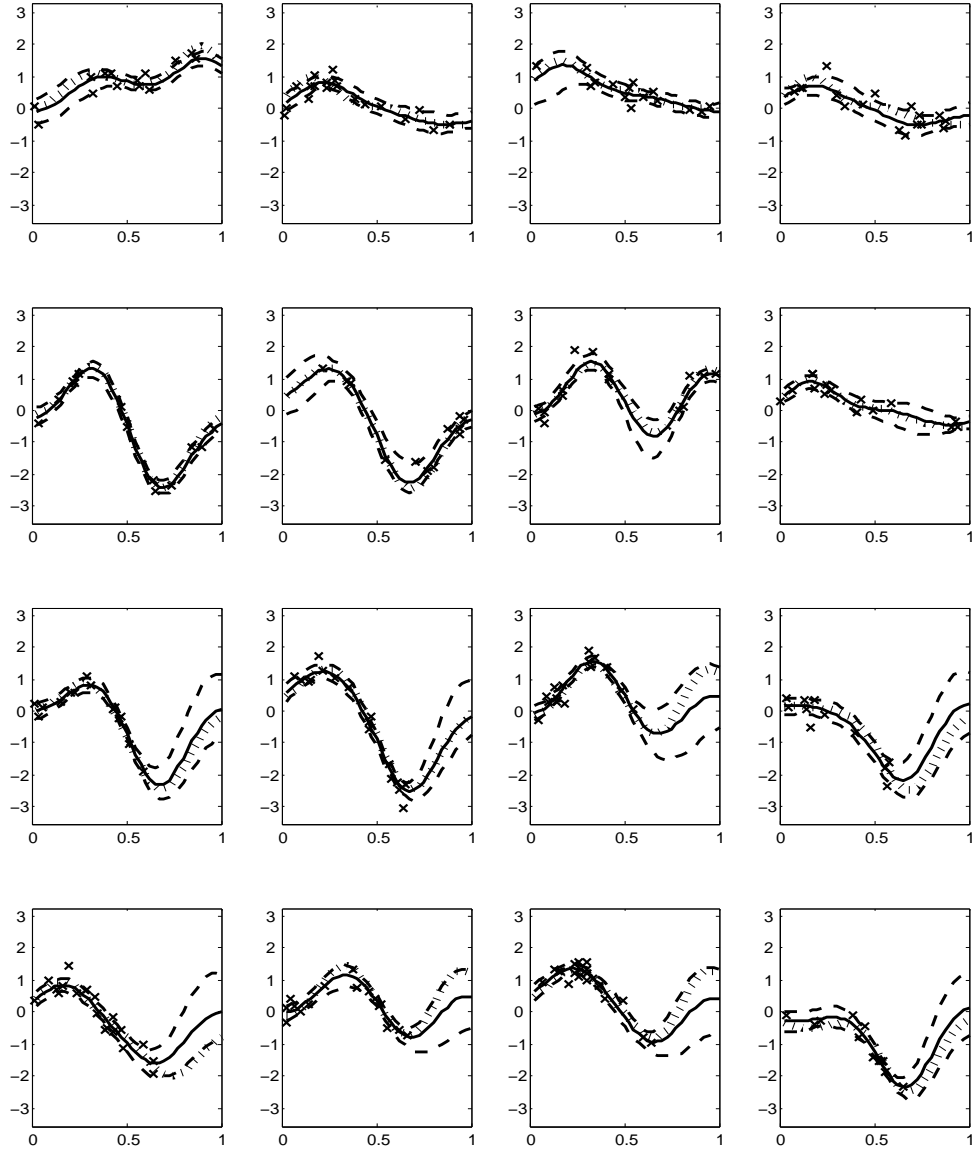
Parameter	Posterior Summary			
	Mean	Median	SD	95% CI
$\alpha$	0.43	0.38	0.26	[0.08, 1.11]
$\beta$	0.30	0.30	0.07	[0.18, 0.45]
$\tau$	12.94	12.93	0.84	[11.27, 14.71]
$\psi$	4.41	4.34	0.94	[2.81, 6.56]
$\nu$	2.06	2.06	0.02	[2.02, 2.11]
$\kappa$	30.42	25.61	17.04	[12.26, 78.63]
$\phi^*$	4.72	4.00	1.96	[3.00, 9.50]



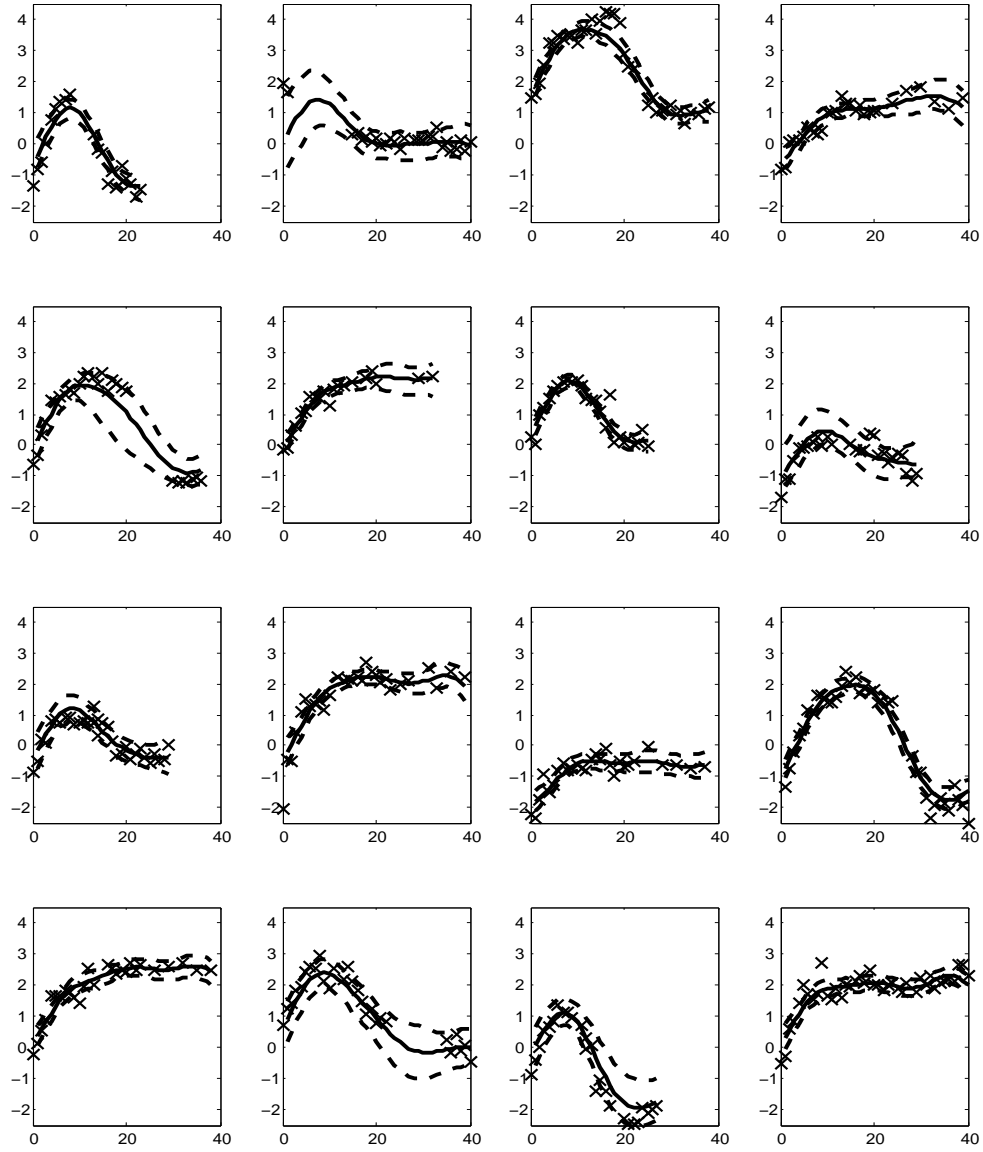
**Fig. 1.** Samples from the prior for  $\rho_{\boldsymbol{\lambda}}(\mathbf{z}_j, \mathbf{z}_{j'}) = \Pr(\gamma_{ij} = \gamma_{ij'} | \boldsymbol{\lambda})$  for  $r = 50$ , equally-spaced  $b_j$  between 0 and 1, and a variety of values of  $\alpha, \beta, \psi$ . Plots are shown holding  $j'$  fixed (dotted line) and varying  $j$ .



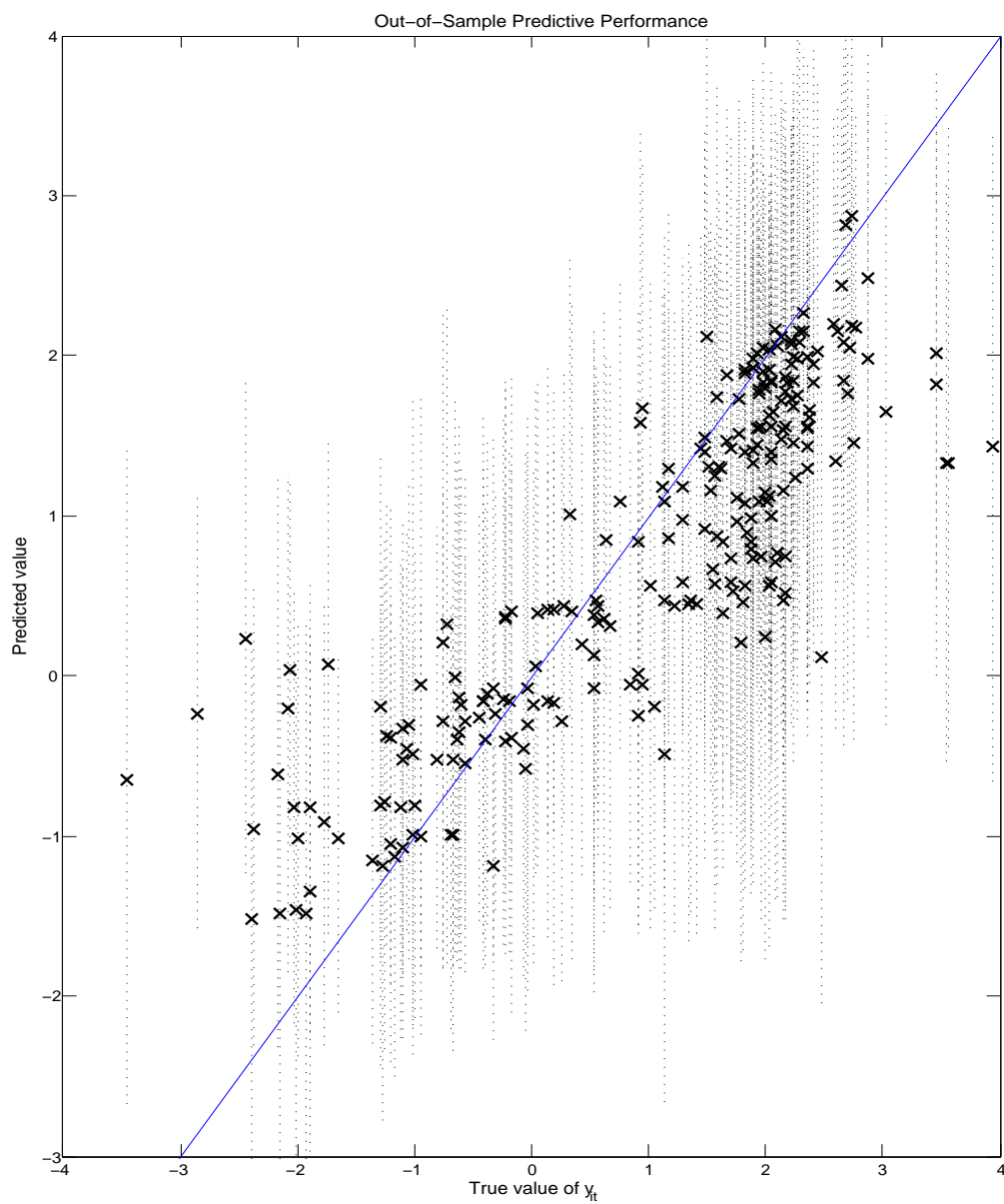
**Fig. 2.** Samples from the prior for  $\Pr(\gamma_{ij} = \gamma'_{i'j} | \gamma_{ij-l} = \gamma'_{i'j-l}, \boldsymbol{\lambda}, \alpha, \beta, \psi)$  for  $\alpha = 25, \psi = 10, \beta = 0.05$  and different choices of lag ranging from  $l = 1, \dots, 40$  with  $j = 41$ . The dashed line represents the lower bound,  $\Pr(\gamma_{ij} = \gamma'_{i'j} | \boldsymbol{\lambda}, \alpha, \beta, \psi)$ .



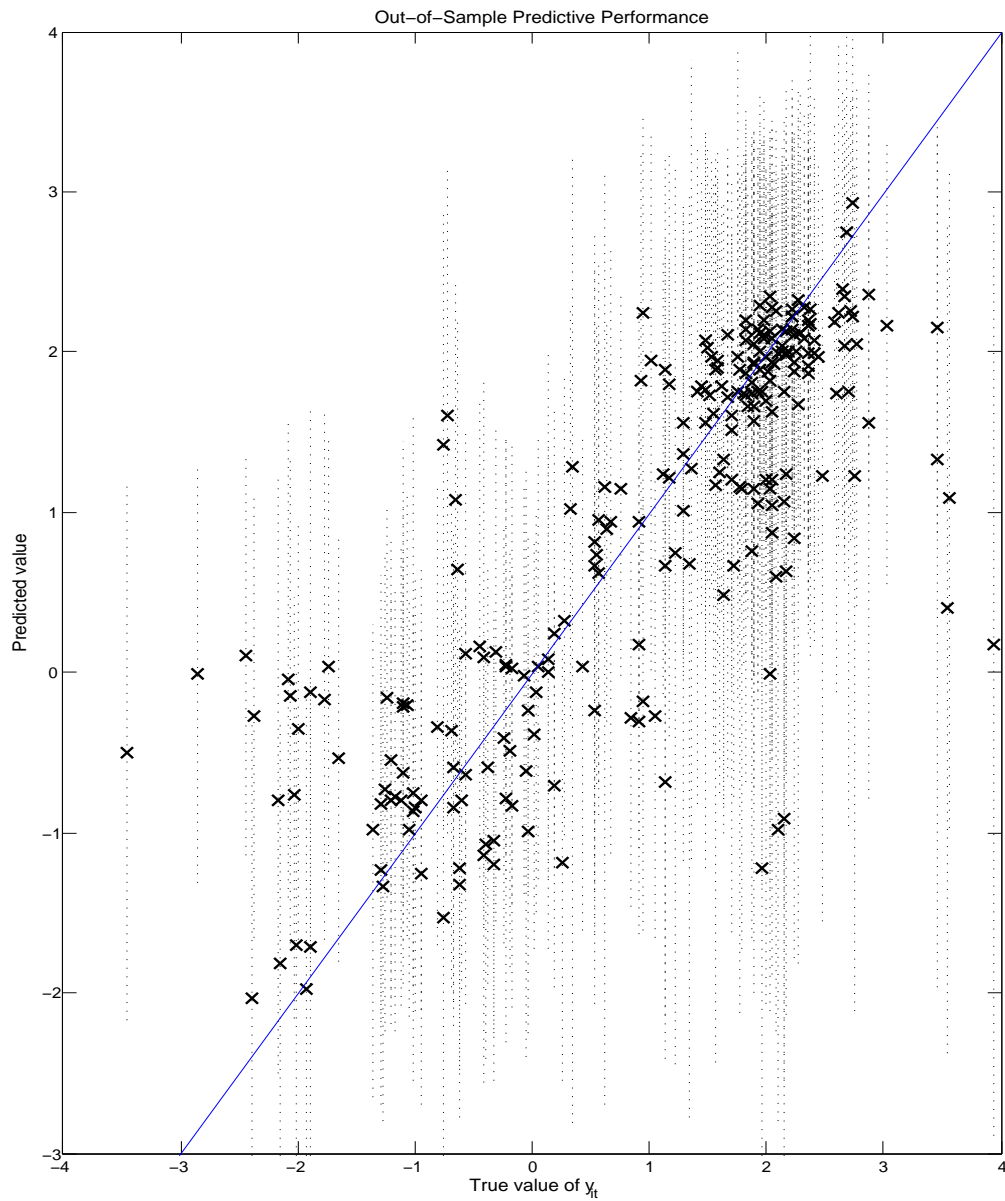
**Fig. 3.** Results for simulation example. The upper 8 plots are for subjects having observations distributed randomly in  $[0,1]$ , while the lower 8 plots are for subjects having observations in  $[2/3,1]$ . The solid lines are posterior mean curves, dashed lines are 95% pointwise credible intervals,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



**Fig. 4.** Log(PdG) data and KLPP-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.



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



**Fig. 6.** Out-of-sample predictive performance for the LPP (Dunson, 2008). The last 5  $\log(\text{PdG})$  observations for 50 women randomly selected from those with 10 or more observations were held out.