Functional Clustering in Nested Designs

Abel Rodriguez

University of California, Santa Cruz, California, USA

David B. Dunson

Duke University, Durham, North Carolina, USA

Summary. We discuss functional clustering procedures for nested designs, where multiple curves are collected for each subject in the study. We start by considering the application of standard functional clustering tools to this problem, which leads to groupings based on the average profile for each subject. After discussing some of the shortcomings of this approach, we present a mixture model based on a generalization of the nested Dirichlet process that clusters subjects based on the distribution of their curves. By using mixtures of generalized Dirichlet processes, the model induces a much more flexible prior on the partition structure than other popular model-based clustering methods, allowing for different rates of introduction of new clusters as the number of observations increases. The methods are illustrated using hormone profiles from multiple menstrual cycles collected for women in the Early Pregnancy Study.

Keywords: Nonparametric Bayes; Nested Dirichlet Process; Functional Clustering; Hierarchical functional data; Hormone Profile.

1. Introduction

The literature on functional data analysis has seen a spectacular growth in the last twenty years, showing promise in applications ranging from genetics (Ramoni et al., 2002; Luan & Li, 2003; Wakefield et al., 2003) to proteomics (Ray & Mallick, 2006), epidemiology (Bigelow & Dunson, 2009) and oceanography (Rodriguez et al., 2008a). Because functional data are inherently complex, functional clustering is useful as an exploratory tool in characterizing variability among subjects; the resulting clusters can be used as a predictive tool or simply as a hypothesis-generating mechanism that can help guide further research. Some examples of functional clustering methods include Abraham et al. (2003), who use B-spline fitting coupled with k-means clustering; Tarpey & Kinateder (2003), who apply k-means clustering via the principal points of random functions; James & Sugar (2003), who develop methods for sparsely sampled functional data that employ spline representations; García-Escudero & Gordaliza (2005), where the robust k-means method for functional clustering is developed; Serban & Wasserman (2005), who use a Fourier representations for the functions along

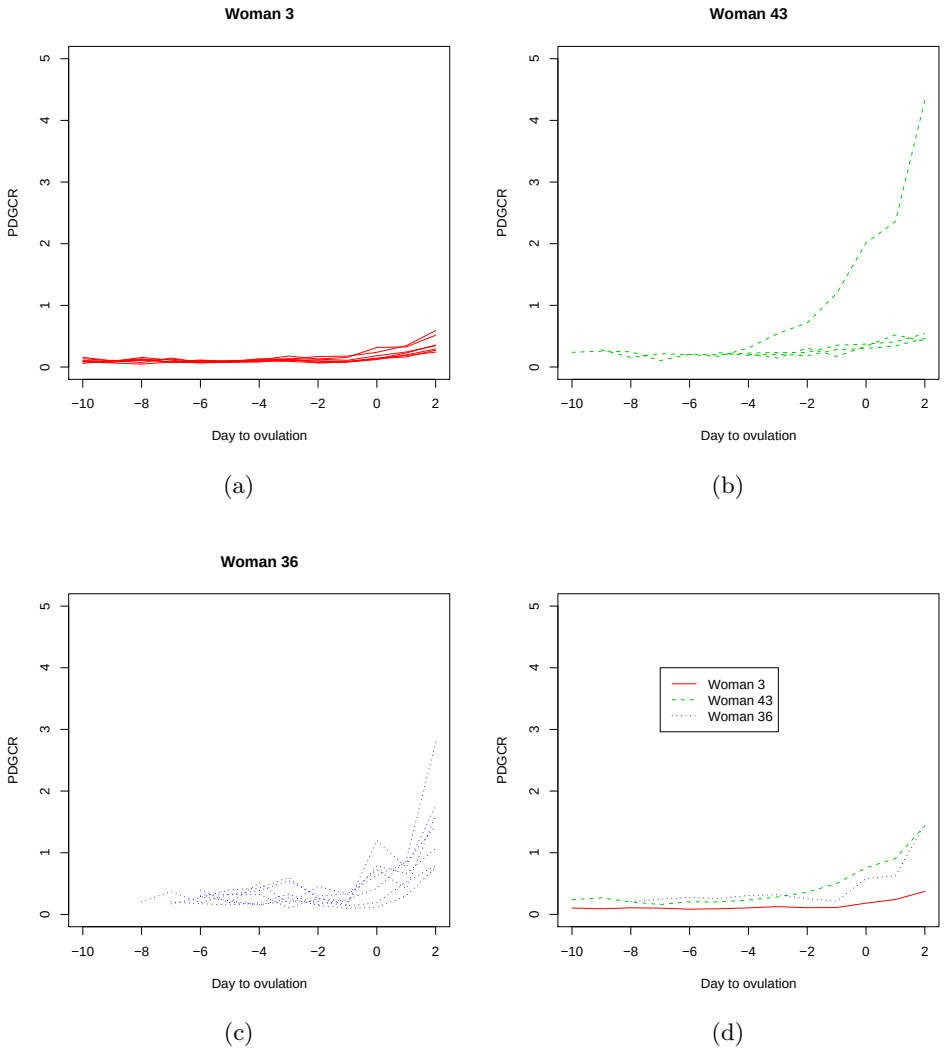


Fig. 1. Comparison of hormone profiles for three women in the Early Pregnancy Study. Frames (a) to (c) show multiple profiles for each woman, while frame the (d) shows the average profile for each woman.

with k-means clustering; Heard et al. (2006), where a Bayesian hierarchical clustering approach that relies on spline representations is proposed; Ray & Mallick (2006), who build a hierarchical Bayesian model that employs a Bayesian nonparametric mixture model on the coefficients of the wavelet representations; and Chiou & Li (2007), where a k-centers functional clustering approach is developed that relies on the Karhunen-Loève representation of the underlying stochastic process generating the curves and accounts for both the means and the modes of variation differentials between clusters.

All of the functional clustering methods described above have been designed for situations where a single curve is observed for each subject or experimental condition. Extensions to nested designs where multiple curves are collected per subject typically assume that coefficients describing subject-specific curves arise from a common parametric distribution, and clustering procedures are then applied to the parameters of this underlying distribution. The result is a procedure that generates clusters of subjects based on their average response curve, which is not appropriate in applications in which subjects vary not only in the average but also in the variability of the replicate curves. For example, in studies of trajectories in reproductive hormones that collect data from repeated menstrual cycles, the average trajectory may provide an inadequate summary of a woman's reproductive functioning. Some women have regular cycles with little variability across cycles in the hormone trajectories, while other women vary substantially across cycles, with a subset of the cycles having very different trajectory shapes. In fact, one indication of impending menopause and a decrease in fecundity is an increase in variability across the cycles. Hence, in forming clusters and characterizing variability among women and cycles in hormone trajectories, it is important to be flexible in characterizing both the mean curve and the distribution about the mean. This situation is not unique to hormone data, and similar issues arise in analyzing repeated medical images as well as other applications.

This paper discusses hierarchical Bayes models for clustering nested functional data. We motivate these models using data from the Early Pregnancy Study (EPS) (Wilcox et al., 1998), where progesterone levels were collected for both conceptive and nonconceptive women from multiple menstrual cycles. Our models use splines bases along with mixture priors to create sparse but flexible representations of the hormone profiles, and can be applied directly to other basis systems such as wavelets. We start by introducing a hierarchical random effects model on the spline coefficients which, along with a generalization of the Dirichlet process mixture (DPM) prior (Ferguson, 1973; Sethuraman, 1994; Escobar & West, 1995), allows for mean-response-curve clustering of women, in the spirit of Ray & Mallick (2006). Then, we extend the model to generate distribution-based clusters using a nested Dirichlet process (NDP) (Rodriguez et al., 2008b). The resulting model simultaneous clusters both curves and subjects, allowing us to identify outlier curves within each group of women, as well as outlying women whose distribution of profiles differs from the rest. To be best of

our knowledge, there is no classical alternative for this type of multilevel clustering.

In order to provide some insight into the challenges associated with functional clustering in nested designs, consider the hormonal profiles from the EPS depicted in Figure 1. Frames (a) to (c) depict the hormone profiles for 3 women, while frame (d) shows the mean profile corresponding to each one of them, obtained by simply averaging all available observations at a given day within the cycle. When looking at the mean profiles in (d), women 43 and 36 seem to have very similar hormonal responses, which are different from those of woman 3. However, when the individual profiles are considered, it is clear that most of the cycles of woman 43 look like those of woman 3 and that the big difference in the means is driven by the single abnormal cycle.

The use of Bayesian nonparametric mixture models for clustering has a long history (Medvedovic & Sivaganesan, 2002; Quintana & Iglesias, 2003; Lau & Green, 2007), and presents a number of practical advantages over other model-based clustering techniques. Nonparametric mixtures induce a probability distribution on the space of partitions of the data, therefore we do not need to specify in advance the number of clusters in the sample. Once updated using the data, this distribution on partitions allows us to measure uncertainty in the clustering structure (including that associated with the estimation of the curves), providing a more complete picture than classical methods. In this paper, we work with a generalized Dirichlet process (GDP) first introduced by Hjort (2000) and study some of its properties as a clustering tool. In particular, we show that the GDP generates a richer prior on data partitions than those induced by popular models such as the Dirichlet process (Ferguson, 1973) or the two parameter Poisson-Dirichlet process (Pitman, 1996), as it allows for an asymptotically bounded number of clusters in addition to logarithmic and power law rates of growth.

The paper is organized as follows: Section 2 reviews the basics of nonparametric regression and functional clustering, while Section 3 explores the design of nonparametric mixture models for functional clustering. Building on these brief reviews, Section 4 describes two Bayesian approaches to functional clustering in nested designs, while Section 5 describes Markov chain Monte Carlo algorithms for this problem. An illustration focused on the EPS is presented in Section 6. Finally, Section 7 presents a brief discusson and future research directions.

2. Model-based functional clustering

To introduce our notation, consider first a simple functional clustering problem where multiple noisy observations are collected from functions $f_1, \dots f_I$. More specifically, for subjects $i = 1, \dots, I$ and within-subject design points $t = 1, \dots, T_i$, observations

consist of ordered pairs (x_{it}, y_{it}) where

$$y_{it} = f_i(x_{it}) + \epsilon_{it}, \qquad \epsilon_{it} \sim \mathsf{N}(0, \sigma_i^2).$$

For example, in the EPS, y_{it} corresponds to the level of progesterone in the blood of subject i collected at day x_{it} of the menstrual cycle, and f_i denotes a smooth trajectory in progesterone for woman i (initially supposing a single menstrual cycle of data from each woman), and clusters in $\{f_i\}_{i=1}^I$ could provide insight into the variability in progesterone curves across women, while potentially allowing us to identify abnormal or outlying curves.

If all curves are observed at the same covariate levels (i.e., $T_i = T$ and $x_{it} = x_t$ for every i), a natural approach to functional clustering is to apply standard clustering methods to the data vectors, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$. For example, in the spirit of Ramsay & Silverman (2005), one could apply hierarchical or K-means clustering to the first few principal components (Yeung & Ruzzo, 2001). From a model-based perspective, one could instead suppose that $\mathbf{y}_i$ is drawn from a mixture of k multivariate Gaussian distributions, with each Gaussian corresponding to a different cluster (Fraley & Raftery, 2002; Yeung et al., 2001). The number of clusters could then be selected using the BIC criteria (Fraley & Raftery, 2002; Li, 2005) or a nonparametric Bayes approach could be used to bypass the need for this selection, while allowing the number of clusters represented in a sample of I individuals to increase stochastically with sample size (Medvedovic & Sivaganesan, 2002). However, in many studies, including the EPS, there are different numbers and spacings of observations on the different subjects.

More generally, we can represent the unknown function f_i as a linear combination of pre-specified basis functions $\{b_k\}_{k=1}^p$, i.e., we can write

$$f_i(x_{it}) = \theta_{i0} + \sum_{k=1}^{p} \theta_{ik} b_k(x_{it})$$

where $\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{i1}, \ldots, \theta_{ip})$ are basis coefficients specific to subject i, with variability in these coefficients controlling variability in the curves $\{f_i\}_{i=1}^I$. A common approach to functional clustering is to induce clustering of the curves through clustering of the basis coefficients (Abraham et al., 2003; Heard et al., 2006). Then the methods discussed above for clustering of the data vectors $\{\mathbf{y}_i\}_{i=1}^I$ in the balanced design case can essentially be applied directly to the basis coefficients $\{\boldsymbol{\theta}_i\}_{i=1}^I$.

Although the methods apply directly to other choices, our focus will be on splines, which have been previously used in the context of hormone profiles (Brumback & Rice, 1998; Bigelow & Dunson, 2009); given a set of knots $\tau_1, \ldots, \tau_p$, the k-th member of the basis system is defined as

$$b_k(x) = (x - \tau_i)_+^q$$

where $(\cdot)_+ = \max\{\cdot, 0\}$. Given the knot locations, inferences on $\boldsymbol{\theta}_i$ and σ_i^2 can be carried out using standard linear regression tools, however, selecting the number and location of the nodes $\tau_1, \ldots, \tau_p$ can be a challenging task. A simple solution is to use a large number of equally spaced knots, together with a penalty term on the coefficients to prevent overfitting. From a Bayesian perspective, this penalty term can be interpreted as a prior on the spline coefficients; for example, the maximum likelihood estimator (MLE) obtained under an L^2 penalty on the spline coefficients is equivalent to the maximum a posteriori estimates for a Bayesian model under a normal prior, while the MLE under an L^1 penalty is equivalent to the maximum a posterior estimate under independent double-exponential priors on the spline coefficients.

Instead of the more traditional Gaussian and double exponential priors, in this paper we focus on zero-inflated priors, in the spirit of Smith & Kohn (1996). Priors of this type enforce sparsity by zeroing out some of the spline coefficients and, by allowing us to select a subset of the knots, provides adaptive smoothing. In their simpler form, zero-inflated priors assume that the coefficients are independent from each other and that

$$\theta_{ik}|\gamma, \sigma_i^2 \sim \gamma \mathsf{N}(0, \omega_k \sigma_i^2) + (1-\gamma)\delta_0, \qquad \sigma_i^2 \sim \mathsf{IGam}(\nu_1, \nu_2), \tag{1}$$

where δ_x denotes the degenerate distribution putting all its mass at x, ω_k controls the overdispersion of the coefficients with respect to the observations and γ is the prior probability that the coefficient θ_{ik} is different from zero. In order to incorporate a priori dependence across coefficients, we can reformulate the hierarchical model by introducing 0-1 random variables $\lambda_{i1}, \ldots, \lambda_{ip}$ such that

$$\mathbf{y}_i|\boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i \sim \mathsf{N}(\mathbf{B}(\mathbf{x}_i)\boldsymbol{\Lambda}_i\boldsymbol{\theta}_i, \sigma_i^2\mathbf{I}), \quad \boldsymbol{\theta}_i|\sigma_i^2 \sim \mathsf{N}(\mathbf{0}, \sigma_i^2\boldsymbol{\Omega}), \quad \sigma_i^2 \sim \mathsf{IGam}(\nu_1, \nu_2),$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i})$ are, respectively, the vectors of responses and covariates associated with subject i, $\mathbf{B}(\mathbf{x}_i)$ is the matrix of basis functions also associated with subject i with entries $[\mathbf{B}(\mathbf{x}_i)]_{tk} = b_k(x_{it})$, and $\boldsymbol{\Lambda}_i = \mathrm{diag}\{\lambda_{i1}, \ldots, \lambda_{ip}\}$ and λ_i equals 1 independently with probability γ. Note that if $\boldsymbol{\Omega}$ is a diagonal matrix and $[\boldsymbol{\Omega}]_{ii} = \omega_i$ we recover the independent priors in (1). For the single curve case, choices for $\boldsymbol{\Omega}$ based on the regression matrix $\mathbf{B}(\mathbf{x}_i)$ are discussed in DiMatteo et al. (2001), Liang et al. (2005) and Paciorek (2006).

Although the preceding two-stage approach is simple to implement using off-the-shelf software, it ignores the uncertainty associated with the estimation of the basis coefficients while clustering the curves. In the spirit of Fraley & Raftery (2002), an alternative that deals with this issue is to employ a mixture model of the form

$$\mathbf{y}_i|\{\boldsymbol{\theta}_k^*\}, \{\sigma_k^{*2}\}, \{\boldsymbol{\Lambda}_k^*\} \sim \sum_{k=1}^{K} w_k \mathsf{N}\left(\mathbf{B}(\mathbf{x}_j)\boldsymbol{\Lambda}_k^*\boldsymbol{\theta}_k^*, \sigma_k^{*2}\mathbf{I}\right), \qquad \sum_{k=1}^{K} w_k = 1, \tag{2}$$

where $\boldsymbol{\theta}_k^*$ is the vector of coefficients associated with the k-th cluster, $\boldsymbol{\Lambda}_k^*$ is the diagonal selection matrix for the k-th cluster, σ_k^{*2} is the observational variance associated with observations collected in the k-th cluster, w_k can be interpreted as the proportion of curves associated with cluster k, and K is the maximum number of clusters in the sample. From a frequentist perspective, estimation of this model can be performed using expectation-maximization (EM) algorithms, however, such EM algorithm leaves the issue of how many mixture components to use unanswered. Alternatively, Bayesian inference can be performed for this model using Markov chain Monte Carlo (MCMC) algorithms once appropriate priors for the vector $\mathbf{w} = (w_1, \ldots, w_K)$ and the cluster-specific parameters $(\boldsymbol{\theta}_k^*, \boldsymbol{\Lambda}_k^*, \sigma_k^{*2})$ have been chosen, opening the door to simple procedures for the estimation of the number of clusters in the sample.

3. Bayesian nonparametric mixture models for functional data

Note that the model in (2) can be rewritten as a hierarchical model by introducing latent variables $\{(\boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i)\}_{i=1}^I$ so that

$$\mathbf{y}_i|\boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i \sim \mathsf{N}\left(\mathbf{B}(\mathbf{x}_i)\boldsymbol{\Lambda}_i\boldsymbol{\theta}_i, \sigma_i^2\mathbf{I}\right) \quad \boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i|G \sim G \quad G(\cdot) = \sum_{k=1}^{K} w_k \delta_{(\boldsymbol{\theta}_k^*, \sigma_k^{*2}, \boldsymbol{\Lambda}_k^*)}(\cdot). \tag{3}$$

Therefore, specifying a joint prior on $\mathbf{w}$ and $\{(\boldsymbol{\theta}_k^*, \sigma_k^{*2}, \boldsymbol{\Lambda}_k^*)\}_{k=1}^K$ is equivalent to specifying a prior on the discrete distribution G generating the latent variables $\{(\boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i)\}_{i=1}^I$. In this section we discuss strategies to specify flexible prior distribution on this mixing distribution in the context of functional clustering. In particular we concentrate on nonparametric specifications for G through the class of stick-breaking distributions.

A random probability measure G on $\mathbb{R}^p$ is said to follow a stick-breaking prior (Ishwaran & James, 2001; Ongaro & Cattaneo, 2004) with baseline measure G_0 and precision parameters $\{a_l\}_{l=1}^L$ and $\{b_l\}_{l=1}^L$ if

$$G(\cdot) = \sum_{k=1}^{K} w_k \delta_{\boldsymbol{\vartheta}_k}(\cdot) \tag{4}$$

where the atoms $\{\boldsymbol{\vartheta}_k\}_{k=1}^K$ are independent and identically distributed samples from G_0 and the weights $\{w_k\}_{l=1}^K$ are constructed as $w_k = u_k \prod_{s<k}(1-u_s)$, with $\{u_k\}_{k=1}^K$ another independent and identically distributed sequence of random variables such that $u_k \sim \mathsf{Beta}(a_k, b_k)$ for $k < K$ and $u_K = 1$. For example, taking $K = \infty$, $a_k = 1-a$ and $b_k = b + ka$ for $0 \leq a < 1$ and $b > -a$ yields the two-parameter Poisson-Dirichlet process (Pitman, 1995; Ishwaran & James, 2001), denoted $\mathsf{PY}(a, b, G_0)$, with the choice $a = 0$ resulting in the Dirichlet Process (Ferguson, 1973; Sethuraman, 1994),

denoted $\mathsf{DP}(b, G_0)$. In mixture models such as (3), G_0 acts as the common prior for the cluster-specific parameters $\{\boldsymbol{\vartheta}_k\}_{k=1}^K$, while the sequences $\{a_k\}_{k=1}^K$ and $\{b_k\}_{k=1}^K$ control the a priori expected number and size of the clusters.

The main advantage of nonparametric mixture models such as the Poisson-Dirichlet process as a clustering tool is that they allow for automatic inferences on the number of components in the mixture. Indeed, these models induce a prior probability on all possible partitions of the set of observations, which is updated based on the information contained in the data. However, Poisson-Dirichlet processes have two properties that might be unappealing in our EPS application; firstly, they place a relatively large probability on partitions that include many small clusters, and secondly, they imply that the number of clusters will tend to grow logarithmically (if $a = 0$) or as a power law (if $a > 0$) as more observations are included in the data set. However, priors that favor introduction of increasing numbers of clusters without bound as the number of subjects increase have some disadvantages in terms of interpretability and sparsity in characterizing high-dimensional data. For example, in applying DP mixture models for clustering of the progesterone curves in EPS, Bigelow and Dunson (2009) obtained approximately 32 different clusters, with half of these clusters singletons. Many of the clusters appeared similar, and it may be that this large number of clusters was partly an artifact of the DP prior. Dunson (2009) proposed a local partition process prior to reduce dimensionality in characterizing the curves, but this method does not produce easily interpretable functional clusters. Hence, it is appealing to use a more flexible global clustering prior that allows the number of clusters to instead converge to a finite constant.

With this motivation, we focus on the generalized Dirichlet process (GDP) introduced by Hjort (2000), denoted $\mathsf{GDP}(a, b, G_0)$. The GDP corresponds to a stick-breaking prior with $K = \infty$, $a_k = a$ and $b_k = b$ for all k. When compared against the Poisson-Dirichlet process, the GDP has quite distinct properties.

THEOREM 1. *Let Z_n be the number of distinct observations in a sample of size n from a distribution G, where $G \sim \mathsf{GDP}(a, b, G_0)$. The expected number of clusters $\mathsf{E}(Z_n)$ is given by*

$$\mathsf{E}(Z_n) = \sum_{i=1}^{n} \frac{ia\Gamma(a+b)\Gamma(b+i-1)}{\Gamma(b)\Gamma(a+b+i) - \Gamma(a+b)\Gamma(b+i)}$$

The proof can be seen in appendix A. Note that for $a = 1$, this expression simplifies to $\mathsf{E}(Z_n) = \sum_{i=1}^{n} \frac{b}{b+i-1} \sim o(\log n)$, a well known result for the Dirichlet process (Antoniak, 1974). Letting $W_n = Z_n - Z_{n-1}$ denote the change in the number of clusters in adding the n-th individual to a sample with $n-1$ subjects, Stirling

approximation can be used to show that

$$E(W_n) = \frac{na\Gamma(a+b)\Gamma(b+n-1)}{\Gamma(b)\Gamma(a+b+n) - \Gamma(a+b)\Gamma(b+n)} \approx C(a,b)n^{-a}.$$

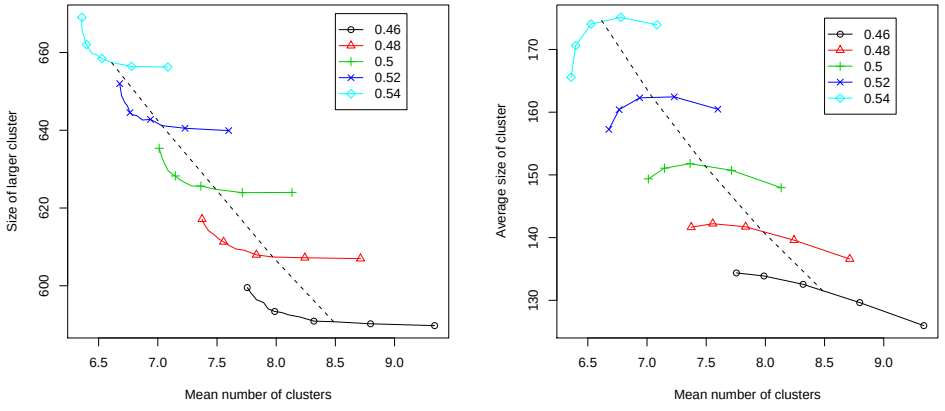
where $C(a,b) = \{a\Gamma(a+b)/\Gamma(b)\}\exp\{-2(a+1)\}$. Hence, $\mathrm{E}(W_n) \to 0$ as $n \to \infty$ and new clusters become increasingly rare as the sample size increases. Note that for $a \leq 1$, the number of clusters will grow slowly but without bound as n increases, with $\mathrm{E}(Z_n) \to \infty$. The rate of growth in this case is proportional to n^{1-a}, which is similar to what is obtained by using the Poisson Dirichlet prior (Sudderth & Jordan, 2009). However, when $a > 1$ the expected number of clusters instead converges to a finite constant, which is a remarkable difference compared with the Dirichlet and Poisson-Dirichlet process. As mentioned above, there may be a number of practical advantages to bounding the number of clusters. In addition, a finite bound on the number clusters seems to be more realistic in many applications, including the original species sampling applications that motivated much of the early development in this area (McCloskey, 1965; Pitman, 1995).

In order to gain further insight into the clustering structure induced by the $\mathsf{GDP}(a,b,G_0)$, we present in Figure 2 the relationship between the size of the largest cluster and the mean number of clusters in the partition (left panel), and the mean cluster size and the number of clusters (right panel) for a sample of size $n = 1000$. Each continuous line correspond to a combination of shape parameters such that $a/(a+b)$ is constant, while the dashed line in the plots corresponds to the combinations available under a Dirichlet process. The plots demonstrate that the additional parameter in the GDP allows us to simultaneously control the number of clusters and the relative size of the clusters, increasing the flexibility of the model as a clustering procedure.

The previous discussion focused on the impact of the prior distribution for the mixture weights on the clustering structure. Another important issue in the specification of the model is the selection of the baseline measure G_0. Note that in the functional clustering setting $\boldsymbol{\vartheta}_k = (\boldsymbol{\theta}_k^*, \sigma_k^{*2}, \boldsymbol{\Lambda}_k^*)$, and therefore a computationally convenient choice that is in line with our previous discussion on basis selection and zero-inflated priors is to write

$$G_0(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\Lambda}) = \mathsf{N}(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\boldsymbol{\Omega}) \times \mathsf{IGam}(\sigma^2|\nu_1,\nu_2) \times \prod_{s=1}^{p} \mathsf{Ber}(\lambda_s|\gamma) \tag{5}$$

A prior of this form allows differential adaptive smoothing for each cluster in the data; the level of smoothness is controlled by γ (the prior probability of inclusion for each of the spline coefficients), and therefore it is convenient to assign to it a hyperprior such as $\gamma \sim \mathsf{Beta}(\eta_1,\eta_2)$.



(a) (b)

Fig. 2. Clustering structure induced by a GDP(a, b, G_0) for a sample of size $n = 1000$. Panel (a) shows the relationship between the size of the largest cluster and the mean number of clusters for different GDPs, where each curve shares a common E$(u_k) = a/(a + b)$. Panel (b) shows the relationship between the average cluster size and the mean number of clusters. The dashed lines corresponds to the combinations available under a standard Dirichlet process.

4. Functional clustering in nested designs

Consider now the case where multiple curves are collected for each subject in the study. In this case, the observations consist of ordered pairs (y_{ijt}, x_{ijt}) where

$$y_{ijt} = f_{ij}(x_{ijt}) + \epsilon_{ijt},$$

where f_{ij} is the jth functional replicate for subject i, with $i = 1, \ldots, I, j = 1, \ldots, n_i$ and $t = 1, \ldots, T_{ij}$. For example, in the EPS, f_{ij} is the measurement error-corrected smooth trajectory in the progesterone metabolite PdG over the j-th menstrual cycle from woman i, with t indexing the sample number and x_{ijt} denoting the day within the i, j menstrual cycle relative to a marker of ovulation day.

A natural extension of (3) to nested designs arises by modeling the expected evolution of progesterone in time for cycle j of woman i as $f_{ij} = \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\theta}_{ij}$ and using a hierarchical model for the set of curve-specific parameters $\{\boldsymbol{\theta}_{ij}\}$ in order to borrow information across subjects and/or replicates. In the following subsections, we introduce two alternative nonparametric hierarchical priors that avoid parametric assumptions on the distribution of the basis coefficients, while inducing hierarchical functional clustering.

4.1. Mean-curve clustering

As a first approach, we consider a Gaussian mixture model, which characterizes the basis coefficients for functional replicate j from subject i as conditionally independent draws from a Gaussian distribution with subject-specific mean and variance, in the spirit of Booth et al. (2008):

$$\mathbf{y}_{ij}|\boldsymbol{\theta}_{ij}, \sigma_i \sim \mathsf{N}(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\theta}_{ij}, \sigma_i^2\mathbf{I}) \qquad \boldsymbol{\theta}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i, \sigma_i^2 \sim G_i \qquad G_i = \mathsf{N}(\boldsymbol{\Lambda}_i\boldsymbol{\theta}_i, \sigma_i^2\boldsymbol{\Sigma}) \tag{6}$$

where $\boldsymbol{\Lambda}_i, \boldsymbol{\theta}_i, \sigma_i^2$ are as described in expression (3). In this model, the average curve for subject i is obtained as $\mathsf{E}\{f_{ij}(x) \,|\, \boldsymbol{\Lambda}_i, \boldsymbol{\theta}_i, \sigma_i^2\} = \mathbf{B}(x)\boldsymbol{\Lambda}_i\boldsymbol{\theta}_i$, with $\boldsymbol{\Lambda}_i$ providing a mechanism for subject-specific basis selection, so that the curves from subject i only depend on the basis functions corresponding to non-zero diagonal elements of $\boldsymbol{\Lambda}_i$. The variability in the replicate curves for the same subject is controlled by $\sigma_i^2\boldsymbol{\Sigma}$, with the subject-specific multiplier allowing subjects to vary in the degree of variability across the replicates. The need to allow such variability is well justified in the hormone curve application.

In order to borrow information across women, we need a hyperprior for the woman specific parameters $\{(\boldsymbol{\Lambda}_i, \sigma_i^2, \boldsymbol{\theta}_i)\}_{i=1}^I$. Since we are interested in clustering subjects, a natural approach is to specify this hyperprior nonparametrically through a generalized Dirichlet process centered around the baseline measure in (5), just as we did for the

single curve case. This yields

$$(\boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i)|G \sim G \qquad\qquad G \sim \mathsf{GDP}(a, b, G_0)$$

with G_0 given in (5). Since the distribution G is almost surely discrete, the model identifies clusters of women with similar average curves. This is clearer if we marginalize out the curve-specific coefficients $\{\boldsymbol{\theta}_{ij}\}$ and the unknown distribution G to obtain the joint likelihood of the data from subject i

$$\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i}|\{w_k\}, \{\boldsymbol{\theta}_k^*\}, \{\sigma_k^{*2}\}, \{\boldsymbol{\Lambda}_k^*\} \sim \sum_{k=1}^{K} w_k \left\{ \prod_{j=1}^{n_i} \mathsf{N}\left(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_k^*\boldsymbol{\theta}_k^*, \sigma_k^{*2}(\mathbf{I} + \boldsymbol{\Sigma})\right) \right\} \quad (7)$$

By incorporating the distribution of the selection matrices $\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_I$ in the random distribution G, this model allows for a different smoothing pattern for each cluster of curves. This is an important difference with a straight generalization of the model in Ray & Mallick (2006), who instead treat the selection matrix as a hyperparameter in the baseline measure G_0 and therefore induce a common smoothing pattern across all clusters.

The model is completed by assigning priors for the hyperparameters. For the random effect variances we take inverse-Wishart priors.

$$\boldsymbol{\Omega} \sim \mathsf{IWis}(\nu_\Omega, \boldsymbol{\Omega}_0) \qquad\qquad \boldsymbol{\Sigma} \sim \mathsf{IWis}(\nu_\Sigma, \boldsymbol{\Sigma}_0)$$

In the spirit of the unit information priors (Paciorek, 2006), the hyper-parameters for these priors can be chosen so that $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Sigma}_0$ are proportional to

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} \mathbf{B}(\mathbf{x}_{ij})'\mathbf{B}(\mathbf{x}_{ij})$$

Finally, the concentration parameters a and b are given gamma priors $a \sim \mathsf{Gam}(\kappa_a, \tau_a)$ and $b \sim \mathsf{Gam}(\kappa_b, \tau_b)$ and the probability of inclusion γ is assigned a beta prior, $\gamma \sim \mathsf{Beta}(\eta_1, \eta_2)$.

4.2. Distribution-based clustering

Because the subject-specific distributions $\{G_i\}_{i=1}^I$ were assumed to be Gaussian and the nonparametric prior was placed on their means, the model in the previous section clusters subjects based on their average profile. However, as we discussed in Section 1, clustering based on the mean profiles might be misleading in studies such as the

EPS in which there are important differences among subjects in not only the mean curve but also the distribution about the mean. In hormone curve applications, it is useful to identify clusters of trajectories over the menstrual cycle to study variability in the curves and identify outlying cycles that may have reproductive dysfunction. It is also useful to cluster women based not simply on the average curve but on the distribution of curves. With this motivation, we generalize our hierarchical nonparametric specification to construct a model that clusters curves within subjects as well as subjects.

To motivate our nonparametric construction, consider first the simpler case in which there are only two types of curves in each cluster of women (say, normal and abnormal), so that it is natural to model the subject-specific distribution as a two-component mixture where

$$\mathbf{y}_{ij}|\varpi_i, \boldsymbol{\Lambda}_{1i}, \boldsymbol{\theta}_{1i}, \sigma^2_{1i}, \boldsymbol{\Lambda}_{2i}, \boldsymbol{\theta}_{2i}, \sigma^2_{2i} \sim \varpi_i \mathsf{N}(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{1i}\boldsymbol{\theta}_{1i}, \sigma^2_{1i}\mathbf{I}) + (1-\varpi_i)\mathsf{N}(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{2i}\boldsymbol{\theta}_{2i}, \sigma^2_{2i}\mathbf{I}) \quad (8)$$

where π_i can be interpreted as the proportion of curves from subject i that are in group 1 (say, normal), and $(\boldsymbol{\Lambda}_{1i}, \boldsymbol{\theta}_{1i}, \sigma^2_{1i})$ are the parameters that describe curves from a normal cycle and $(\boldsymbol{\Lambda}_{2i}, \boldsymbol{\theta}_{2i}, \sigma^2_{2i})$ are the parameters describing the curves from an abnormal cycle. Note that in this case we have not one but two variance parameters for each individual, which provides additional flexibility by allowing each cluster of curves to present a different level of observational noise. This feature is desirable in the EPS because, for a given woman, observational noise in abnormal cycles tends to be larger than in normal cycles.

Under this formulation, the subject-specific distribution is described by the vector of parameters $(\varpi_i, \boldsymbol{\Lambda}_{1i}, \boldsymbol{\theta}_{1i}, \sigma^2_{1i}, \boldsymbol{\Lambda}_{2i}, \boldsymbol{\theta}_{2i}, \sigma^2_{2i})$, and clustering subjects could be accomplished by clustering these vectors. We can accomplish this by using another mixture model that mimics (2) and (7), so that

$$\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i}|\{\pi_k\}, \{\varpi_k\}, \{\boldsymbol{\theta}^*_{1k}\}, \{\sigma^{*2}_{1k}\}, , \{\boldsymbol{\Lambda}^*_{1k}\}, \{\boldsymbol{\theta}^*_{2k}\}, \{\sigma^{*2}_{2k}\}, \{\boldsymbol{\Lambda}^*_{2k}\} \sim \sum_{k=1}^{K} \pi_k \prod_{j=1}^{n_i} \left\{\varpi_k \mathsf{N}(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}^*_{1k}\boldsymbol{\theta}^*_{1k}, \sigma^{*2}_{1k}\mathbf{I}) + (1-\varpi_k)\mathsf{N}(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}^*_{2l}\boldsymbol{\theta}^*_{2k}, \sigma^{*2}_{1k}\mathbf{I})\right\} \quad (9)$$

As with the simpler model-based functional clustering model we introduced at the end of Section 2, we could generate ML estimators for the parameters of this model using an EM algorithm. However, such an approach still leaves open the question of how many mixture components should be used, both at the subject and curve level. For this reason, we adopt a Bayesian perspective and generalize the model using the nonparametric priors discussed in Section 3. To do so, we start by rewriting (9) as a

general mixture model where

$$\mathbf{y}_{ij}|\boldsymbol{\theta}_{ij},\sigma^2_{ij},\boldsymbol{\Lambda}_{ij} \sim \mathsf{N}\left(\mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{ij}\boldsymbol{\theta}_{ij},\sigma^2_{ij}\mathbf{I}\right) \qquad \boldsymbol{\theta}_{ij},\sigma^2_{ij},\boldsymbol{\Lambda}_{ij}|G_i \sim G_i \tag{10}$$

and G_i is a discrete distribution which is assigned a nonparametric prior. Note that this is analogous to the formulation in (6), but by replacing the Gaussian distribution with a random distribution with a nonparametric prior we are modeling the within-subject variability by clustering curves into groups with homogeneous shape.

Now, we need to define a prior over the collection $\{G_i\}_{i=1}^I$ that induces clustering among the distributions. For example, we could use a discrete distribution *whose atoms are in turn random distributions*, for example,

$$G_i \sim \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*}$$

where $\pi_k = v_k \prod_{s<k}(1-v_s)$, $v_k \sim \mathsf{Beta}(a_1,b_1)$ and $G_k^* \sim \mathsf{GDP}(a_2,b_2,G_0)$ independently. This implies that

$$G_k^* = \sum_{l=1}^{\infty} \varpi_{lk} \delta_{(\boldsymbol{\theta}^*_{lk},\sigma^{2*}_{lk},\boldsymbol{\Lambda}^*_{lk})} \qquad (\boldsymbol{\theta}^*_{lk},\sigma^{2*}_{lk},\boldsymbol{\Lambda}^*_{lk}) \sim G_0.$$

with $\varpi_{lk} = u_{lk}\prod_{s<l}(1-u_{sk})$ and $u_{lk} \sim \mathsf{Beta}(a_2,b_2)$ and G_0 as in (5). Therefore, if we were to replace the collection $\{G_k^*\}_{k=1}^{\infty}$ with random discrete distributions with only two atoms, and we were to integrate over the random distributions $\{G_i\}_{i=1}^I$, this model would be equivalent to (9) with $K=\infty$.

This model on the collection $\{G_i\}_{i=1}^I$ is a generalization of the nested Dirichlet process introduced in Rodriguez et al. (2008b) and, as with other models based on nested nonparametric processes, interesting special cases can be obtained by considering the limit of the precision parameters. For example, letting $b_2 \to 0$ while keeping a_2 fixed induces a model where all menstrual cycles within a woman are assumed to have the same profile, and subjects are clustered according to their mean cycle. Such a model is equivalent to the one obtained by taking $\boldsymbol{\Sigma} \to 0$ in (6). On the other hand, by letting $b_1 \to \infty$ while keeping a_1 constant, we obtain a model where all subjects are treated as different and menstrual cycles are clustered within each women. In this case, information is borrowed across the menstrual cycles of each women, but not across women.

Again, the model is completed by specifying prior distributions on the remaining parameters. As before, we let $\boldsymbol{\Omega} \sim \mathsf{IWis}(\nu_\Omega,\boldsymbol{\Omega}_0)$, $\nu_2 \sim \mathsf{Gam}(\rho,\psi)$ and $\gamma \sim \mathsf{Beta}(\eta_1,\eta_2)$, providing a conditionally conjugate specification amenable for simple computational implementation. Finally, for the precision priors of the GDPs we set

$$a_1 \sim \mathsf{Gam}(\kappa_{a_1},\tau_{a_1}) \qquad b_1 \sim \mathsf{Gam}(\kappa_{b_1},\tau_{b_1})$$
$$a_2 \sim \mathsf{Gam}(\kappa_{a_2},\tau_{a_2}) \qquad b_2 \sim \mathsf{Gam}(\kappa_{b_2},\tau_{b_2})$$

5. Computation

As is commonplace in Bayesian inference, we resort to Markov chain Monte Carlo (MCMC) algorithms (Robert & Casella, 1999) for computation in our functional clustering models. Given an initial guess for all unknown parameters in the model, the algorithms proceed by sequentially sampling blocks of parameters from their full conditional distributions. In particular, we design our algorithms using truncated versions of the GDP and the nested GDP, where a large but finite number of atoms is used to approximate the nonparametric mixture distributions; the well known results on the convergence of truncations as the number of atoms grows that were originally presented in Ishwaran & James (2001) and Rodriguez et al. (2008b) can be directly extended to this problem (see Appendix B). In this section we briefly describe the blocked Gibbs sampling algorithms associated with the two models discussed in Section 4, further details can be seen in Appendix C.

5.1. Sampling in the mean-based clustering model

For the purpose of sampling the parameters of the mean-based clustering model we truncate the stick-breaking construction of the random distribution G so that it has K atoms, i.e,

$$G(\cdot) = \sum_{k=1}^{K} w_k \delta_{(\boldsymbol{\theta}_k^*, \sigma_k^{*2} \boldsymbol{\Lambda}_k^*)}$$

where $w_k = u_k \prod_{s<k}(1-u_s)$ with $u_k \sim \mathsf{Beta}(a,b)$ for $k < K$ and $u_k = 1$, which ensures that the weights of all components add up to 1. Also, we introduce a sequence of latent indicator variables $\zeta_1, \ldots, \zeta_I$, where $\zeta_i = k$ if and only if subject i is assigned to cluster k, $(\boldsymbol{\theta}_i, \sigma_i^2, \boldsymbol{\Lambda}_i) = (\boldsymbol{\theta}_{\zeta_i}^*, \sigma_{\zeta_i}^{*2}, \boldsymbol{\Lambda}_{\zeta_i}^*)$. After introducing these latent indicators, the posterior distribution of the model parameters can be written as:

$$\begin{aligned} p(\{\boldsymbol{\theta}_{ij}\}, \{\boldsymbol{\theta}_k^*\}, \{\sigma_k^{*2}\}, \{\boldsymbol{\Lambda}_k^*\}, \{\zeta_i\}, \{w_k\}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}, a, b, \gamma \,|\{\mathbf{y}_{ij}\}) \propto \\ p(\{\mathbf{y}_{ij}\}|\{\boldsymbol{\theta}_{ij}\}, \{\sigma_k^{*2}\}, \{\zeta_i\}) p(\{\boldsymbol{\theta}_{ij}\}|\{\boldsymbol{\theta}_k^*\}, \{\boldsymbol{\Lambda}_k^*\}, \{\sigma_k^{*2}\}, \boldsymbol{\Sigma}, \{\zeta_i\}) p(\boldsymbol{\Sigma}) \\ p(\{\boldsymbol{\theta}_k^*\}, \{\boldsymbol{\Lambda}_k^*\}, \{\sigma_k^{*2}\}|\boldsymbol{\Omega}, \nu_2, \gamma) p(\{\zeta_i\}|\{w_k\}) p(\{w_k\}|a,b) p(\boldsymbol{\Omega}) p(\nu_2) p(\gamma) p(a,b) \end{aligned}$$

The indicators $\{\zeta_i\}$ and the cluster specific parameters $\{\boldsymbol{\theta}_k^*\}$, $\{\boldsymbol{\Lambda}_k^*\}$ and $\{\sigma_k^{*2}\}$ can be sampled easily after integrating out the curve-specific parameters $\{\boldsymbol{\theta}_{ij}\}$. Specifically, ζ_i can be sampled conditionally on all other parameters in the model from a

multinomial distribution where

$$\mathsf{Pr}(\zeta_i = k|\cdots) \propto w_k \sigma_{\zeta_i}^{-\sum_{j=1}^{n_i} T_{ij}} \times \exp\left\{ -\frac{1}{2\sigma_{\zeta_i}^2} \sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} - \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{\zeta_i}^* \boldsymbol{\theta}_{\zeta_i}^*\right)' (\boldsymbol{\Sigma} + \mathbf{I})^{-1} \left(\mathbf{y}_{ij} - \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{\zeta_i}^* \boldsymbol{\theta}_{\zeta_i}^*\right) \right\}$$

and $k = 1, \ldots, K$, while the cluster specific parameters are sampled by first integrating out $\boldsymbol{\theta}_k^*$ and σ_k^{*2} and sampling $\boldsymbol{\Lambda}_k^*$, and then sampling $\boldsymbol{\theta}_k^*$ and σ_k^{*2} conditionally on $\boldsymbol{\Lambda}_k^*$ (see Appendix C). The component weights are simulated through the stick-breaking ratios, whose full conditional distribution is given by

$$u_k|\cdots \sim \mathsf{Beta}\left(a + r_k, b + \sum_{s=k+1}^{K} r_s\right)$$

where $r_k = \sum_{i=1}^{I} \mathbf{1}_{(\zeta_i = k)}$ is the number of observations assigned to component k of the mixture. Given the indicators and class specific parameters, we can sample the parameters $\{\boldsymbol{\theta}_{ij}\}$ independently from their full conditional distribution.

$$\boldsymbol{\theta}_{ij}|\cdots \sim \mathsf{N}\Big(\left[\mathbf{B}(\mathbf{x}_{ij})'\mathbf{B}(\mathbf{x}_{ij}) + \boldsymbol{\Lambda}_{\zeta_i}^{*\prime} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}_{\zeta_i}^*\right]^{-1} \left[\mathbf{B}(\mathbf{x}_{ij})'\mathbf{y}_{ij} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_{\zeta_i}^* \boldsymbol{\theta}_{\zeta_i}^*\right], \sigma_{\zeta_i}^{*2} \left[\mathbf{B}(\mathbf{x}_{ij})'\mathbf{B}(\mathbf{x}_{ij}) + \boldsymbol{\Lambda}_{\zeta_i}^{*\prime} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}_{\zeta_i}^*\right] \Big)$$

The rest of the steps are relatively straightforward. The full conditional random effects variance $\boldsymbol{\Sigma}$ can be sampled from an inverse Wishart distribution,

$$\boldsymbol{\Sigma}|\cdots \sim \mathsf{IWis}\left(\nu_\Sigma + \sum_{i=1}^{I}\sum_{j=1}^{n_i} T_{ij}, \boldsymbol{\Sigma}_0 + \sum_{i=1}^{I}\sum_{j=1}^{n_i} \frac{1}{\sigma_{\zeta_i}^{*2}} (\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}_{\zeta_i})(\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}_{\zeta_i})' \right),$$

and the parameters of the baseline measure $\boldsymbol{\Omega}$, ν_2 and γ are respectively sampled from another inverse Wishart, a Gamma and a beta distribution,

$$\boldsymbol{\Omega}|\cdots \sim \mathsf{IWis}\left(\nu_\Omega + Kp, \boldsymbol{\Omega}_0 + \sum_{k=1}^{K} \frac{1}{\sigma_k^{*2}} \boldsymbol{\theta}_k^* \boldsymbol{\theta}_k^{*\prime} \right)$$

$$\nu_2|\cdots \sim \mathsf{Gam}\left(\rho + K\nu_1, \psi + \sum_{k=1}^{K} \sigma_k^{*2} \right)$$

$$\gamma|\cdots \sim \mathsf{Beta}\left(\eta_1 + \sum_{k=1}^{K} \mathrm{tr}\, \boldsymbol{\Lambda}_k^*, \eta_2 + Kp - \sum_{k=1}^{K} \mathrm{tr}\, \boldsymbol{\Lambda}_k^* \right)$$

where $\operatorname{tr}\mathbf{D}$ denotes the trace of the matrix $\mathbf{D}$. Finally, sampling for the precision parameters a and b is done using a random-walk Metropolis-Hastings algorithm with log normal proposals.

5.2. Sampling in the distribution-based clustering model

Again, in this case we truncate the stick-breaking constructions to generate a blocked Gibbs sampler. Hence,

$$G_i \sim \sum_{k=1}^{K} \pi_k \delta_{G_k^*} \qquad\qquad G_k^* = \sum_{l=1}^{L} \varpi_{lk} \delta_{(\boldsymbol{\theta}_{lk}^*, \sigma_{lk}^{*2}, \boldsymbol{\Lambda}_{lk}^*)}$$

Also, we introduce two sequences of latent indicator variables, $\zeta_1, \ldots, \zeta_I$ and $\{\xi_{1j}\}_{j=1}^{n_1}, \ldots, \{\xi_{Ij}\}_{j=1}^{n_I}$ such that$(\boldsymbol{\theta}_{ij}, \sigma_{ij}^2, \boldsymbol{\Lambda}_{ij}) = (\boldsymbol{\theta}_{\xi_{ij},\zeta_i}^*, \sigma_{\xi_{ij},\zeta_i}^{*2}, \boldsymbol{\Lambda}_{\xi_{ij},\zeta_i}^*)$. After introducing these latent indicators, the posterior distribution of the model parameters can be written as:

$$\begin{aligned} p(\{\zeta_i\}, \{\xi_{ij}\}, \{\boldsymbol{\theta}_{lk}^*\}, \{\sigma_{lk}^{*2}\}, \{\boldsymbol{\Lambda}_{lk}^*\}, \{\pi_k\}, \{\varpi_{lk}\}, \boldsymbol{\Omega}, a_1, b_1, a_2, b_2, \gamma \,|\{\mathbf{y}_{ij}\}) \propto \\ p(\{\mathbf{y}_{ij}\}|\{\zeta_i\}, \{\xi_{ij}\}, \{\boldsymbol{\theta}_{lk}^*\}, \{\sigma_{lk}^{*2}\}, \{\boldsymbol{\Lambda}_{lk}^*\}) p(\{\boldsymbol{\theta}_{lk}^*\}, \{\sigma_{lk}^{*2}\}, \{\boldsymbol{\Lambda}_{lk}^*\}|\boldsymbol{\Omega}, \nu_2, \gamma) \\ p(\{\zeta_i\}|\{\pi_k\}) p(\{\pi_k\}|a_1, b_1) p(\{\xi_{ij}\}|\{\varpi_{lk}\}) p(\{\varpi_{lk}\}|a_2, b_2) p(\boldsymbol{\Omega}) p(\nu_2) \\ p(\gamma) p(a_1, b_1) p(a_2, b_2) \end{aligned}$$

The indicators $\{\zeta_i\}$ can be sampled from a multinomial distribution with weights

$$\begin{aligned} \mathsf{Pr}(\zeta_i = k|\cdots) \propto \pi_k \Bigg\{ \prod_{j=1}^{n_i} \Bigg[\sum_{l=1}^{L} \varpi_{lk} \left(2\pi\sigma_{lk}^{*2}\right)^{-T_{ij}/2} \times \\ \exp\left\{ -\frac{1}{2\sigma_{lk}^{*2}} \left(\mathbf{y}_{ij} - \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{lk}^*\boldsymbol{\theta}_{lk}^*\right)' \left(\mathbf{y}_{ij} - \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{lk}^*\boldsymbol{\theta}_{lk}^*\right) \right\} \Bigg] \Bigg\} \end{aligned}$$

for $k = 1, \ldots, K$. Similarly the indicators $\{\xi_{ij}\}$ are sampled from another multinomial where

$$\begin{aligned} \mathsf{Pr}(\xi_{ij} = l|\cdots) \propto \varpi_{l\zeta_i} \left(\sigma_{l\zeta_i}^{*}\right)^{-T_{ij}} \times \\ \exp\left\{ -\frac{1}{2\sigma_{l\zeta_i}^{*2}} \left(\mathbf{y}_{ij} - \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{l\zeta_i}^*\boldsymbol{\theta}_{l\zeta_i}^*\right)' \left(\mathbf{y}_{ij} - \mathbf{B}(\mathbf{x}_{ij})\boldsymbol{\Lambda}_{l\zeta_i}^*\boldsymbol{\theta}_{l\zeta_i}^*\right) \right\} \end{aligned}$$

and $l = 1, \ldots, L$. The rest of the parameters are sampled almost identically as in the mean-based clustering algorithm. The cluster specific parameters are sampled

by first integrating out $\boldsymbol{\theta}^*_{lk}$ and σ^{*2}_{lk} and sampling $\boldsymbol{\Lambda}^*_{lk}$, and then sampling $\boldsymbol{\theta}^*_{lk}$ and σ^{*2}_{lk} condtionally on $\boldsymbol{\Lambda}^*_{lk}$ (see Appendix C). The mixture weights are sampled again through the stick-breaking weights $\{v_k\}$ and $\{u_{lk}\}$ so that

$$v_k \sim \mathsf{Beta}\left(a_1 + r_k, b_1 + \sum_{s=k+1}^{K} r_s\right), \quad u_{lk} \sim \mathsf{Beta}\left(a_2 + m_{lk}, b_2 + \sum_{s=l+1}^{L} m_{ls}\right),$$

where $r_k = \sum_{i=1}^{I} \mathbf{1}_{(\zeta_i=k)}$ and $m_{lk} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \mathbf{1}_{(\zeta_i=k,\xi_{ij}=l)}$. The parameters of the baseline measure $\boldsymbol{\Omega}$, ν_2 and γ are sampled from their full conditional distributions,

$$\begin{aligned}
\boldsymbol{\Omega}|\cdots &\sim \mathsf{IWis}\left(\nu_\Omega + Kp, \boldsymbol{\Omega}_0 + \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{1}{\sigma^{*2}_{lk}} \boldsymbol{\theta}^*_{lk}\boldsymbol{\theta}^{*\prime}_{lk}\right) \\
\nu_2|\cdots &\sim \mathsf{Gam}\left(\rho + K\nu_1, \psi + \sum_{k=1}^{K}\sum_{l=1}^{L} \sigma^{*2}_{lk}\right) \\
\gamma|\cdots &\sim \mathsf{Beta}\left(\eta_1 + \sum_{k=1}^{K}\sum_{l=1}^{L} \operatorname{tr}\boldsymbol{\Lambda}^*_{lk}, \eta_2 + Kp - \sum_{k=1}^{K}\sum_{l=1}^{L} \operatorname{tr}\boldsymbol{\Lambda}^*_{lk}\right)
\end{aligned}$$

Finally, sampling for the precision parameters a_1, b_1, a_2 and b_2 is done as before using a random-walk Metropolis-Hastings algorithm with log normal proposals.

6. An illustration: The Early Pregnancy Study

Progesterone plays a crucial role in controlling different aspects of reproductive function in women, from fertilization to early development and implantation. Therefore, understanding the variability of hormonal profiles across the menstrual cycle and across subjects is important in understanding mechanisms of infertility and early pregnancy loss, as well as for developing approaches for identifying abnormal menstrual cycles and women for diagnostic purposes. Our data, extracted from the Early Pregnancy Study (Wilcox et al., 1998), consists of daily creatinine-corrected concentrations of pregnanediol-3-glucuronide (PdG) for 60 women along multiple menstrual cycles, measured in micrograms per milligram of creatinine (μg/ml Cr). We focus on a 13-day intervals extending from 10 days before ovulation to 2 days after ovulation. According to the results in Dunson et al. (1999), this interval should include the fertile window of the menstrual cycle during which non-contracepting intercourse has a non-neglible probability of resulting in a conception. Also, for this illustration we considered only non-conceptive cycles and women with at least four cycles in record. Therefore, the number of curves per woman varies between 4 and 9.

We analyzed the EPS data using both the mean-based clustering model described in Section 4.1 and the distribution-based clustering model of Section 4.2 using the algorithms from Section 5. In the mean-based clustering algorithm, the GDP was truncated so that $K = 40$, while in the distribution-based algorithm the nested GDP was truncated so that $K = 40$ and $L = 30$. Although these numbers might seem large given the sample sizes involved, a large number of empty components is helpful in improving the mixing of the algorithms. In both cases, we used piecewise linear splines ($q = 1$) and $p = 13$ knots, corresponding to each of the days considered in the study.

Prior distributions in the mean-based clustering algorithm were set as follows. For the concentration parameters, we used proper priors $a \sim \mathsf{Gam}(3, 3)$ and $b \sim \mathsf{Gam}(3, 3)$, for the observational variance, we set $\sigma^2 \sim \mathsf{IGam}(2, 0.04)$, so that $\mathsf{E}(\sigma^2) = 0.04$. To allow uncertainty in the probability of basis selection within the base measure, we let $\gamma \sim \mathsf{Beta}(2, 4)$, implying that we expect about one third of the spline basis functions to be used in any given cluster. Priors for the distribution-based clustering algorithm were chosen in a similar way, with $a_1 \sim \mathsf{Gam}(3, 3)$, $b_1 \sim \mathsf{Gam}(3, 3)$, $a_2 \sim \mathsf{Gam}(3, 3)$ and $b_2 \sim \mathsf{Gam}(3, 3)$, while for the baseline measure we picked a prior the inclusion probabilities $\gamma \sim \mathsf{Beta}(2, 4)$ and the prior on the group specific variances as given by $\mathsf{IGam}(2, 0.04)$.

All inferences presented in this section are based on 100,000 samples obtained after a burn-in period of 10,000 iterations. Results seemed robust to reasonable changes in the parameter values, and no convergence issues where detected when reviewing trace plots for model parameters.

We start by comparing the clustering structure generated by the mean-based and distribution-based models considered in Section 4. For this purpose, we show in Figures 3 and 4 heatmaps of the average pairwise clustering probability matrix under these two models. Entry (i, j) of the matrix contains the posterior probability that observations i and j are assigned to the same cluster. The black squares in the plots correspond to point estimates of the clustering structure obtained through the method described in Lau & Green (2007). In our case, the point estimate is obtained by minimizing a loss function that assigns equal weights to all pairwise misclassification errors. Therefore, the resulting plots provide information about the optimal clustering structure for the data as well as the uncertainty associated with it.

Figures 3 and 4 show that, as our descriptive analysis suggested, under the mean-based clustering model, subjects 36 and 43 are assigned to a common cluster while subject 3 is assigned to a different cluster. In contrast, under the distribution-based clustering model it is subjects 3 and 43 who are placed in a common cluster, while subject 36 is assigned to another. In addition, note that in spite of the difference in the composition of the main clusters, the outlier subjects (corresponding to the small clusters at the top and right of both heatmaps) are very similar under both methods.

More generally, posterior estimates of the precision parameters on the GDP suggest that a logarithmic rate of growth for the number of clusters might be reasonable for this data. For the mean-based clustering, the posterior mean for a was 1.06 and the 95% posterior symmetric credible interval was $(0.65, 1.49)$, while the posterior mean for b was 0.77 with 95% credible interval $(0.17, 1.72)$. For the distribution based clustering, the corresponding estimates are 1.03 $(0.62, 1.56)$ and 0.72 $(0.20, 1.73)$ for a_1 and b_1, and 1.12 $(0.71, 1.43)$ and 0.27 $(0.15, 1.51)$ for a_2 and b_2.

Figure 5 shows reconstructed profiles under the distribution-based clustering model for some representative women in each of the main four groups. Most profiles are flat before ovulation, when hormone levels start to increase. Also, in most clusters the profiles tend to be relatively consistent for any single woman. However, we can see some outliers, typically corresponding to elevated post-ovulation levels and/or early increases in the hormone levels. Cluster 3 corresponds to women with very low hormonal levels, even after ovulation. This group has few outliers, and those present are characterized by a slightly larger increase in progesterone after ovulation, which is still under 1 μg/ml Cr. Group 2 shows much more diversity in the hormonal profiles, as well as a slightly higher baseline level in progesterone level and an earlier rise in progesterone than group 3. Group 1 tends to show few outliers, and otherwise differs from the previous ones in a higher baseline level and an early and very fast increase in progesterone. Finally, group 4 presents "normal" cycles with the highest baseline level of progesterone (1 μg/ml Cr) and the fastest increase in progesterone after ovulation, along with "abnormal" cycles with even higher baseline levels and very extreme levels of progesterone after ovulation (close to 5 μg/ml Cr).

7. Discussion

We have presented two approaches to functional clustering in nested designs. These approaches look into different features of the nested samples, and are therefore applicable in different circumstances. Our mean-based clustering approach is easier to interpret and provides an excellent alternative when within-subject samples are homogenous. However, when within-subject curves are heterogeneous, mean-based clustering can lead to biased results. Therefore, in studies such as the EPS, distribution-based models such as the one described here provide a viable alternative that acknowledges the heterogeneity in the function replicates from a subject..

One interesting insight that can be gathered from the results of the EPS data is that, for small numbers of functional replicates per subject and rare outliers, the effect of the distribution-based clustering is to perform clustering based on the modal, rather than the mean profile. That is, the distribution-based clustering model is able to automatically discount the abnormal curves, leading to more appropriate clustering patterns if the effect of outliers needs to be removed.

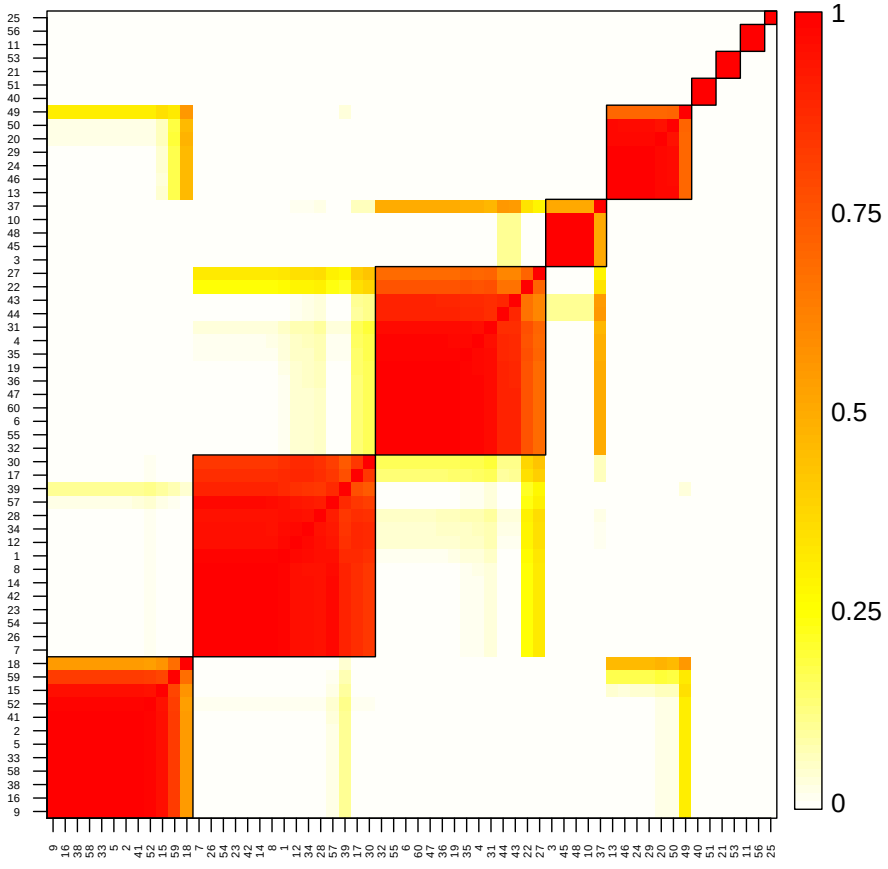


Fig. 3. Average incidence matrix, illustrating probabilities of joint pairwise classification for the 60 women in the EPS under the mean-curve clustering procedure described in Section 4.1. White corresponds to zero probability, while red corresponds to 1. The squares correspond to a point estimate of the cluster structure in the data.

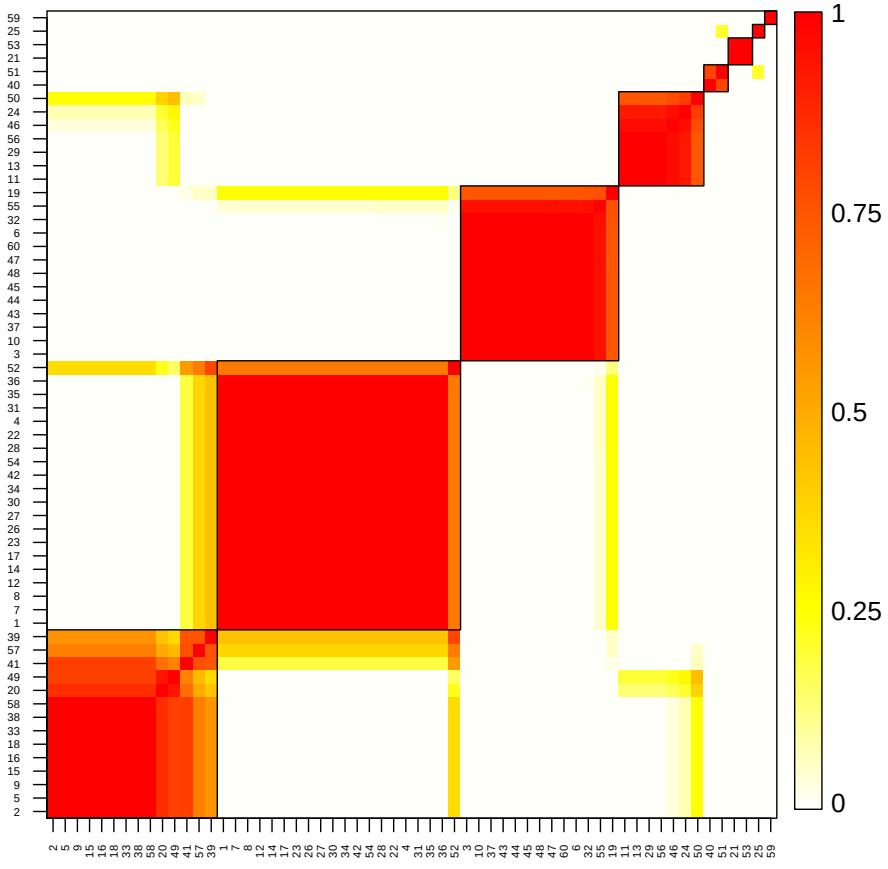


Fig. 4. Average incidence matrix, illustrating probabilities of joint pairwise classification for the 60 women in the EPS under the distribution-based clustering procedure described in Section 4.2. White corresponds to zero probability, while red corresponds to 1. The squares correspond to a point estimate of the cluster structure in the data.

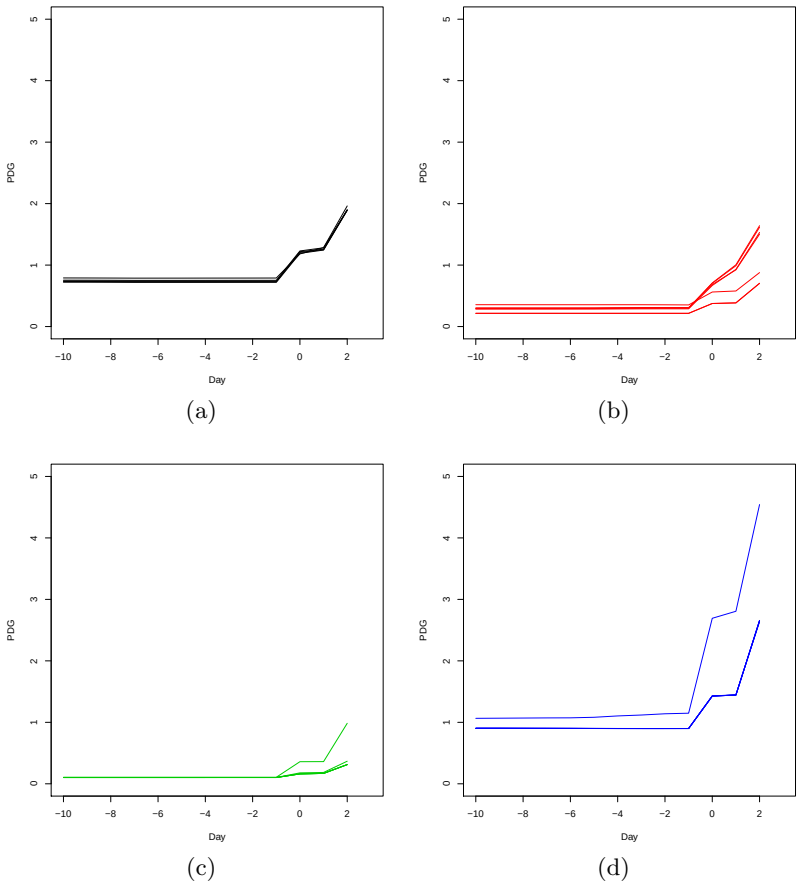


Fig. 5. Reconstructed profiles for some representative subjects in the study. Panel (a) corresponds to patient 9 (who was chosen from cluster 1, counting from the bottom left), panel (b) to patient 36 (who was chosen from cluster 2), panel (c) to patient 45 (who was chosen from cluster 3), and panel (d) corresponds to patient 13 (who was chosen from cluster 4).

8. Acknowledgement

We would like to thank Nils Hjort for his helpful comments and Allen Wilcox for providing access to the EPS data. AR was partially supported by grant R01GM090201-01 from the National Institute of General Medical Sciences of the National Institutes of Health. DBD was partially supported by grant R01 ES017240-01 from the National Institute of Environmental Health Sciences of the National Institutes of Health.

A. Proof of theorem 1

Let $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots$ be a sequence of independent and identically distributed samples from a random distribution G, which follows a $\mathsf{GDP}(a, b, G_0)$ distribution. Also, let W_i be 1 if $\boldsymbol{\theta}_i^*$ is different from every $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*$, and zero otherwise. Clearly, $Z_n = \sum_{i=1}^n W_i$ is the number of distinct values among the first n samples form a $\mathsf{GDP}(a, b, G_0)$. Hjort (2000) shows that

$$\begin{aligned}
\mathsf{E}(W_i) &= i\frac{\mathsf{E}\{u(1-u)^{i-1}\}}{1-\mathsf{E}\{(1-u)^i\}} \\
&= i\frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b+i-1)}{\Gamma(a+b+i)}}{1-\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a)\Gamma(b+i)}{\Gamma(a+b+i)}} \\
&= \frac{ia\Gamma(a+b)\Gamma(b+i-1)}{\Gamma(b)\Gamma(a+b+i)-\Gamma(a+b)\Gamma(b+i)}
\end{aligned}$$

which completes the proof.

B. Truncations of Generalized Dirichlet processes

THEOREM 2. *Assume that samples of n observations have been collected for each of J distributions and are contained in vector $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_J')$. Also, let*

$$\begin{aligned}
P^{\infty\infty}(\boldsymbol{\theta}) &= \int\int P(\boldsymbol{\theta}|G_j)P^{\infty}(dG_j|Q)P^{\infty}(dQ) \\
P^{LK}(\boldsymbol{\theta}) &= \int\int P(\boldsymbol{\theta}|G_j)P^{L}(dG_j|Q)P^{K}(dQ)
\end{aligned}$$

be, respectively, the prior distribution of the model parameters under the nested GDP model and its corresponding truncation after integrating out the random distributions, and $P^{\infty\infty}(\mathbf{y})$ and $P^{LK}(\mathbf{y})$ be the prior predictive distribution of the observations derived from these priors. Then

$$\int \left|P^{LK}(\mathbf{y}) - P^{\infty\infty}(\mathbf{y})\right| d\mathbf{y} \leq \int \left|P^{LK}(\boldsymbol{\theta}) - P^{\infty\infty}(\boldsymbol{\theta})\right| \leq \epsilon^{LK}(\alpha, \beta)$$

where

$$\epsilon^{LK}(\alpha,\beta)=\begin{cases}4\left(1-\left[1-\left(\frac{b_1}{a_1+b_1}\right)^{K-1}\right]^{J}\right) & \text{if } L=\infty, K<\infty\\ 4\left(1-\left[1-\left(\frac{b_2}{a_2+b_2}\right)^{L-1}\right]^{nJ}\right) & \text{if } L<\infty, K=\infty\\ 4\left(1-\left[1-\left(\frac{b_1}{a_1+b_1}\right)^{K-1}\right]^{J}\left[1-\left(\frac{b_2}{a_2+b_2}\right)^{L-1}\right]^{nJ}\right) & \text{if } L<\infty, K<\infty\end{cases}$$

The proof is a direct extension of results in Ishwaran & James (2001), Ishwaran & James (2002) and Rodriguez et al. (2008b) and it is omitted for reasons of space. This result is particularly important since it justifies the use of computational algorithms based on finite mixtures and allows us to choose adequate truncation levels.

C. Details on the computational algorithm

Here we provide details on the sampling of the component-specific parameters. For the mean-curve clustering model, the probability of the model associated with $\mathbf{\Lambda}_k^*$ is proportional to

$$\mathsf{Pr}(\mathbf{\Lambda}_k^*=\mathbf{\Lambda}|\cdots)\propto(2\pi)^{-s_k/2}\left|\mathbf{\Omega}\right|^{-1/2}\left|\mathbf{\Omega}^{-1}+\mathbf{E}_k(\mathbf{\Lambda})\right|^{1/2}\frac{\Gamma(\nu_1+s_k/2)}{\Gamma(\nu_1)}\nu_2^{\nu_1}$$
$$\left\{\nu_2+\left[A_k(\mathbf{\Lambda})-\mathbf{d}_k(\mathbf{\Lambda})'\{\mathbf{\Omega}^{-1}+\mathbf{E}_k(\mathbf{\Lambda})\}\mathbf{d}_k(\mathbf{\Lambda})\right]/2\right\}^{-(\nu_1+s_k/2)}$$

with

$$s_k=\sum_{\{i:\zeta_i=k\}}\sum_{j=1}^{n_i}T_{ij}$$

$$A_k(\mathbf{\Lambda})=\sum_{\{i:\zeta_i=k\}}\sum_{j=1}^{n_i}\mathbf{y}_{ij}'\left[\mathbf{\Sigma}_{\mathbf{\Lambda}}+\mathbf{I}_{p_{\mathbf{\Lambda}}}\right]^{-1}\mathbf{y}_{ij}$$

$$\mathbf{d}_k(\mathbf{\Lambda})=\sum_{\{i:\zeta_i=k\}}\sum_{j=1}^{n_i}\mathbf{B}_{\mathbf{\Lambda}}(\mathbf{x}_{ij})'\left[\mathbf{\Sigma}_{\mathbf{\Lambda}}+\mathbf{I}\right]^{-1}\mathbf{y}_{ij}$$

$$\mathbf{E}_k(\mathbf{\Lambda})=\sum_{\{i:\zeta_i=k\}}\sum_{j=1}^{n_i}\mathbf{B}_{\mathbf{\Lambda}}(\mathbf{x}_{ij})'\left[\mathbf{\Sigma}_{\mathbf{\Lambda}}+\mathbf{I}\right]^{-1}\mathbf{B}_{\mathbf{\Lambda}}(\mathbf{x}_{ij})$$

where $\mathbf{\Sigma}_{\mathbf{\Lambda}}$ and $\mathbf{B}_{\Lambda}(\mathbf{x}_{ij})$ correspond to the restrictions of matrices $\mathbf{\Sigma}$ and $\mathbf{B}(\mathbf{x}_{ij})$ to the entries where the diagonal elements of $\mathbf{\Lambda}$ are different from 0. In our case, since

the number of possible values for $\mathbf{\Lambda}_k^*$ is small ($2^{13} = 8192$), we explicitly compute the full-conditional probability distribution for each of the possible models and perform exact sampling from this posterior distribution. More generally, when the number of nodes is large we can use a random-walk Metropolis-Hasting algorithm as described in George & McCulloch (1997). Now, conditionally on $\mathbf{\Lambda}_k^*$, we can sample σ_k^* and $\boldsymbol{\theta}_k^*$ from

$$\sigma_k^*|\mathbf{\Lambda}_k^* \sim \mathsf{IGam}\left(\nu_1 + s_k/2, \nu_2 + \left[A_k(\mathbf{\Lambda}_k^*) - \mathbf{d}_k(\mathbf{\Lambda}_k^*)'\{\mathbf{\Omega}^{-1} + \mathbf{E}_k(\mathbf{\Lambda}_k^*)\}\mathbf{d}_k(\mathbf{\Lambda}_k^*)\right]/2\right)$$
$$\boldsymbol{\theta}_k^*|\sigma_k^*, \mathbf{\Lambda}_k^* \sim \mathsf{N}\left(\left[\mathbf{\Omega}^{-1} + \mathbf{E}_k(\mathbf{\Lambda}_k^*)\right]^{-1}\mathbf{d}_k(\mathbf{\Lambda}_k^*), \sigma_k^*\left[\mathbf{\Omega}^{-1} + \mathbf{E}_k(\mathbf{\Lambda}_k^*)\right]^{-1}\right)$$

For the distribution-based clustering algorithm the expressions are similar, but we replace s_k, $A_k(\mathbf{\Lambda})$, $\mathbf{d}_k(\mathbf{\Lambda})$ and $\mathbf{E}_k(\mathbf{\Lambda})$ by

$$\begin{aligned}
s_{lk} &= \sum_{\{(i,j):\zeta_i=k,\xi_{ij}=l\}} T_{ij} \\
A_{lk}(\mathbf{\Lambda}) &= \sum_{\{(i,j):\zeta_i=k,\xi_{ij}=l\}} \mathbf{y}_{ij}'\mathbf{y}_{ij} \\
\mathbf{d}_{lk}(\mathbf{\Lambda}) &= \sum_{\{(i,j):\zeta_i=k,\xi_{ij}=l\}} \mathbf{B}_{\mathbf{\Lambda}}(\mathbf{x}_{ij})'\mathbf{y}_{ij} \\
\mathbf{E}_{lk}(\mathbf{\Lambda}) &= \sum_{\{(i,j):\zeta_i=k,\xi_{ij}=l\}} \mathbf{B}_{\mathbf{\Lambda}}(\mathbf{x}_{ij})'\mathbf{B}_{\mathbf{\Lambda}}(\mathbf{x}_{ij})
\end{aligned}$$

References

Abraham, C., Corillon, P. A., Matzner-Løber, E. & Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics* **30**, 581–595.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Bigelow, J. L. & Dunson, D. B. (2009). Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association* **To apeear**.

Booth, G., Casella, G. & Hobert, J. P. (2008). Clustering using objective functions and stochastic search. *Journal of Royal Statistical Society, Series B* **70**, 119–139.

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

Chiou, J.-M. & Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society, Series B.* **69**, 679–699.

DiMatteo, I., Genovese, C. & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–1071.

Dunson, D. B., Weinberg, C., Wilcox, A. J. & Baird, D. D. (1999). Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. *Human Reproduction* **14**, 1835–1839.

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

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

Fraley, C. & Raftery, A. E. (2002). Model-based Gaussian, discriminant analysis and density estimation. *Journal of American Statistical Association* **97**, 611–631.

García-Escudero, L. A. & Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification* **22**, 185–201.

George, E. I. & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Heard, N. A., Holmes, C. C. & 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. *Journal of the American Statistical Association* **101**, 18–29.

Hjort, N. (2000). Bayesian analysis for a generalized Dirichlet process prior. Technical report, University of Oslo.

Ishwaran, H. & James, L. (2002). Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

James, G. & Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of American Statistical Association* **98**, 397–408.

Lau, J. W. & Green, P. (2007). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

Li, J. (2005). Clustering based on multilayer mixture models. *Journal of Computational and Graphical Statistics* **14**, 547–568.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2005). Mixtures of g-priors for Bayesian variable selection. Technical report, Institute of Statistics and Decision Sciences, Duke University.

Luan, Y. & Li, H. (2003). Clustering of time-course gene expression data using a mixed effects model with b-splines. *Bioinformatics* **19**, 474–482.

McCloskey, J. W. (1965). *A Model for the Distribution of Individuals by Species in an Environment*. Ph.D. thesis, Michigan State University.

Medvedovic, M. & Sivaganesan, S. (2002). Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

Ongaro, A. & Cattaneo, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics and Probability Letters* **67**, 33–45.

Paciorek, C. J. (2006). Misinformation in the conjugate prior for the linear model with implications for free-knot spline modelling. *Bayesian Analysis* **1**, 375–383.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Eds. T. S. Ferguson, L. S. Shapeley & J. B. MacQueen, pp. 245–268. Hayward, CA:IMS.

Quintana, F. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B.* **65**, 557–574.

Ramoni, M., Sebastiani, P. & Kohane, P. (2002). Cluster analysis of gene expression dynamics. *Procedings of the National Academi of Sciences* **99**, 9121–9126.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edition.

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

Robert, C. P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer - Verlag.

Rodriguez, A., Dunson, D. B. & Gelfand, A. E. (2008a). Bayesian nonparametric functional data analysis through density estimation. *Biometrika To appear*.

Rodriguez, A., Dunson, D. B. & Gelfand, A. E. (2008b). The nested Dirichlet process, with discussion. *Journal of American Statistical Association* **103**, 1131–1144.

Serban, N. & Wasserman, L. (2005). CATS: clustering after transformation and smoothing. *Journal of American Statistical Association* **100**, 990–999.

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

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–343.

Sudderth, E. B. & Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems 21*, Eds. D. Koller, D. Schuurmans, Y. Bengio & L. Bottou.

Tarpey, T. & Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification* .

Wakefield, J., Zhou, C. & Self, S. (2003). Modelling gene expression over time: curve clustering with informative prior distributions. In *In Bayesian Statistics 7*, Eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, pp. 721–732. Oxford University Press.

Wilcox, A. J., Weinberg, C., O'Connor, J. F., Baid, D. D., Schlatterer, J., E., C. R., Armstrong, E. G. & Nisula, B. C. (1998). Incidence of early loss of pregnancy. *New England Journal of Medicine* **319**, 189–194.

Yeung, K. Y., Fraley, C., Muruan, A., Raftery, A. E. & Ruzzo, W. L. (2001). Model-based clustering and data transformation for gene expression data. *Bioinformatics* **17**, 977–987.

Yeung, K. Y. & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763–774.