THE DIRICHLET LABELING PROCESS FOR CLUSTERING FUNCTIONAL DATA

XuanLong Nguyen and Alan E. Gelfand

University of Michigan and Duke University

Abstract: We consider problems involving functional data where we have a collection of functions, each viewed as a process realization, e.g., a random curve or surface. For a particular process realization, we assume that the observation at a given location can be allocated to separate groups via a random allocation process, which we name the Dirichlet labeling process. We investigate properties of this process and its use as a prior in a mixture model. We develop exact and approximate representations for the labeling process, analyze the global and local clustering behavior, clarify model identifiability and posterior consistency, and develop efficient inference methods for models using such priors. Performance is demonstrated with synthetic data examples, a public-health application, and an image segmentation task.

Key words and phrases: Clustering, Dirichlet process mixtures, Gaussian processes, Gibbs posterior, hybrid Dirichlet process, variational Bayes.

1. Introduction

A recurring theme in the nonparametric Bayes literature has been the development of mixture models based on Dirichlet processes (DP) (Ferguson (1973); Sethuraman (1994); Ishwaran and James (2001)). These models have proved to be useful in applications that involve clustering observations into distinct groups; the dependence of different groupings can be achieved via the formalism of dependent Dirichlet processes (e.g., MacEachern (2000); DeIorio et al. (2004); Gelfand, Kottas, and MacEachern (2005); Teh et al. (2006)).

In this paper we are interested in mixture modeling for functional data (Ramsay and Silverman (2002, 2006); Ferraty and Vieu (2006)). From the viewpoint of functional data analysis we are given a sample of n functions, surfaces or curves $Y_1, \ldots, Y_n$ over $\mathbb{R}^d$, each viewed as a realization of a stochastic process Y. The curves are observed at a common set of locations $x_1, \ldots, x_m \in D$, where D is a subset of $\mathbb{R}^d$. This setting is natural in many applications: an image is a surface of light intensity on $\mathbb{R}^2$. The ocean temperature at a location is a function of depth. The monthly progesterone level of a female subject is a function of time.

The primary objective here is to examine clustering of the set of curves. Formalizing the notion of clustering of curves raises several interesting challenges.

First, we envision Y_i as a *noisy* version of the curve θ_i. The θ_i's are assumed to be smooth at least continuous and clustering is considered with regard to these latent θ's. For instance, it is easy to ensure mean square continuous realizations using a Gaussian process with a suitable covariance function (see, e.g., Stein (1999)). Of course, introducing noise raises a trade-off issue. With too much noise, one θ can explain all of the Y_i's - one cluster; with too little noise, each Y_i requires a distinct θ_i - no clustering. Second, we can envision a notion of local clustering by clustering curve realizations $\theta_1(x), \ldots, \theta_n(x)$ at any location $x \in D$ using a DP mixture. We can envision global clustering, θ_i and $\theta_{i'}$ identical for all $x \in D$, and, possibly, attempt to formalize notions of "partial" clustering. With smoothness for these functions, the groupings at locations close to each other are expected to be more similar than those at distant locations. In other words, there is an uncountable collection of dependent DP mixtures, one for each location, with the dependence regulated by the inherent spatial structure. Such clustering would be viewed as "local". Alternatively, one can assume that the spatial dependence is regulated by, say, a Gaussian process (GP) on D. For instance, a simple approach is to allow random curves to be drawn from a Dirichlet process with a GP as base measure (Gelfand, Kottas, and MacEachern (2005)). However, this approach is limited by the discrete nature of DP realizations: conditional on the DP atoms, a random curve is either a replicate of one of a countable set of curves at *all* locations in D, or not at all. Evidently, this is "global" clustering.

Our approach assumes that the collection of curve realizations can be represented in terms of k "canonical" curves drawn from a stochastic process G_0, but each realization can be expressed as a hybrid species – random portions of the curve may belong to different species. Canonical curves provide the basis for representing a curve in terms of disjoint segments with distinct behavior (in terms of, e.g., smoothness and monotonic properties). In certain applications, such as image modeling, a canonical curve might simply be a (random) constant function that represents a corresponding level set in the image. The notion of hybrid species curves has been explored in various contexts, including text analysis and genetics (Blei, Ng, and Jordan (2003)); (Pritchard, Stephens, and Donnelly (2000)), as well as in the context of spatial and functional data (Duan, Guidani, and Gelfand (2007)); (Petrone, Guidani, and Gelfand (2009)). In particular, our approach is based on the hybrid Dirichlet process mixture model first introduced by Petrone, Guidani, and Gelfand (2009). Implicit in their modeling is a latent group allocation process, which we call the *Dirichlet labeling process*. This labeling process, which we now denote by **p**, allows random *local* allocation to one of a collection of species curves. Operating formally, we work with finite-dimensional Dirichlet processes (Ishwaran and James (2001)) where **p** is a random probability measure on $\{1, \ldots, k\}^D$ that is drawn from a Dirichlet

process via a base measure $\mathbf{q}$ (i.e., $\mathbf{p} \sim DP(\alpha\mathbf{q})$), where $\mathbf{q}$ is also a probability measure on $\{1, \ldots, k\}^D$. Explicitly, we mean that for any finite set of locations, $\{x_j, \ j = 1, \ldots, m\}$, $\mathbf{p}$ and $\mathbf{q}$ are probability distributions on a k^m dimensional simplex such that $\mathbf{p}|\mathbf{q} \sim Dir(\alpha\mathbf{q})$. To allow spatial dependence of random allocation, $\mathbf{q}$ is constructed via discretization and copula transformation of a latent Gaussian process, which essentially regulates the random allocation. Letting $k \to \infty$, it can be shown that the marginal distribution of the curve (at each location) tends to the marginal drawn from a Dirichlet process mixture (Petrone, Guidani, and Gelfand (2009)).

The novel contributions offered here are the following. First, we undertake a detailed investigation of the Dirichlet labeling process model that provides a random label for each θ curve at each location x. The labels are dependent within a realization of a curve and, through the Dirichlet process, can introduce clustering of labels across curves. We illuminate properties of this proposed process, develop both exact and approximate representations of the labeling processes, exact calculation when only two labels are allowed and approximate calculation when a large number of labels are allowed. Then, we investigate the overall mixture model. We clarify the identifiability of the mixture distribution, building upon results from Ishwaran and Zarepour (2002) that broaden the classical work of Teicher (1963). We also discuss consistency of posterior inference under the overall mixture model, extending results in Ishwaran and Zarepour (2002). Here, the key issue is how k/n behaves as $n \to \infty$. For any finite number of locations m, the needed asymptotic rate is $k^m = O(n)$. However, our practical interest resides in the case where k is small relative to n, where we can represent a large number of curves with a small number of canonical species. This, in turn, leads to analyzing the local and global clustering behavior in the overall mixture model. Lastly, statistical inference with the latent labeling process is expensive with a large number of local sites and clusters. We offer computationally efficient inference methods by proposing a model fitting strategy using Gibbs sampling that employs ideas of pseudo-likelihood and approximate variational inference in Markov random fields (Wainwright and Jordan (2003)). We provide application to curves of progesterone levels of women during the course of a menstrual cycle and, perhaps surprisingly, to an image segmentation setting.

There are several recent approaches that permit random local allocation for functional data. In Fernandez and Green (2002), the authors consider Markov random fields over lattices with Poisson distributed data where the weights in the mixture vary with locations. Closer in spirit to our framework is the nonparametric Bayesian mixtures of Hidden Markov Models Teh et al. (2006). Our labeling process is arguably more computationally tractable, especially for high-dimensional D and large m, due to the exploitation of spatial structure in the

model that yields accurate conditional probability approximation. A number of recent papers introduce various constructions based on the Sethuraman's stick-breaking representation, with varying weights assigned for different locations (Griffin and Steel (2006); Dunson and Park (2008); Duan, Guidani, and Gelfand (2007); Sudderth et al. (2008)). The work of Griffin and Steel (2006) and Dunson and Park (2008) exemplify several distinct proposals for constructing spatially dependent DP mixture marginals. In contrast with our approach, these are somewhat indirect methods for enforcing the spatial dependence – while label sharing across the collection of curves is encouraged, label sharing across nearby locations of the *same* curve is not directly possible.

A number of recent work consider Bayesian models for representing a collection of functions in terms of kernel basis functions (Pillai et al. (2006); MacLehose and Dunson (2008); Dunson (2008b,a)), $f(x) = \int K(x, u)\gamma(u)du$, where the coefficient function $\gamma(\cdot)$ is endowed with a nonparametric prior. In particular, Dunson (2008b) and Dunson (2008a) insist on sparse representations by modeling the coefficient covariates $\gamma(\cdot)$ in terms of a labeling process. In Dunson (2008b), the labeling process is modeled by independent Dirichlet processes, while Dunson (2008a) uses kernel functions to induce the spatial dependency of labels in a manner similar to that of the Dirichlet labeling process. The key distinction between these and our work is that the Dirichlet labeling process allows distributional specification of labeling realizations over continuous domain without the need for kernel basis specification. More similar to our approach is the work of Duan, Guidani, and Gelfand (2007). It also specifies a generalized DP mixture model using the view of hybrid species curves. Their approach requires a labeling process obtained by thresholding k latent Gaussian processes, resulting in a model that is computationally challenging to fit. By contrast, our approach utilizes only one latent Gaussian process to regulate spatial dependence, while allowing label sharing through the use of the Dirichlet process at the next stage. The resultant model is simpler and computationally more tractable.

Although we are taking a nonparametric Bayesian approach to the clustering of functional data, we must mention that there is a substantial non-Bayesian literature on this important topic. See, e.g., Abraham et al. (2003); Biau, Devroye, and Lugosi (2008); Chiou and Li (2007); Dabo-Niang, Ferraty, and Vieu (2006); Fraiman, Justel, and Svarc (2008); Fraiman and Muniz (2001); James and Sugar (2003); Ma and Zhong (2008); Tokushige, Yadohisa, and Inada (2007) and the references listed therein. Comparison between Bayesian and non-Bayesian approaches is lacking in the literature, but is not our objective here.

The paper is organized as follows. Section 2 provides background on the Dirichlet labeling process prior for a mixture model. Section 3 presents properties of the Dirichlet labeling process and the overall "hybrid" prior. Section

4 discusses identifiability of the mixture model we propose, as well as posterior consistency. Section 5 addresses parameter identifiability, a concern in model fitting with our flexible specification. Section 6 focuses on model fitting and inference. Section 7 offers results for experimental and data analysis. We conclude with some discussion in Section 8. All proofs and additional details are deferred to an Appendix.

2. Formalizing the Model

We define a mixture model for curve realizations $Y_1, \ldots, Y_n$ over $\mathbb{R}^D$, that are noisy versions of, respectively, $\theta_1, \ldots, \theta_n$. In particular, observations are obtained at local sites $x_1, \ldots, x_m \in D$, so $Y_i(x_j) = \theta_i(x_j) + \epsilon_i(x_j)$. Such modeling is standard in functional data analysis (see, e.g., Ramsay and Silverman (2006, p.40) where the ϵ's contribute noise, perturbation, disturbance, error, to capture roughness in the raw data). In different words, the data is assumed to be subject to pure error fluctuations relative to the process model with the process model specifying suitably smooth curves (though, in some applications, we may prefer to leave the noise in the θ_i's). This process view is in accord with the idea of local and global clustering for the collection of curves.

For a given k, we envision k "canonical" species curves θ_j^* $(j = 1, \ldots, k)$ based on which the collection of θ_i's can be represented. Indeed, each of the θ_i curves is described by the label function, $L_i(x), x \in D, L_i(x) \in \{1, \ldots, k\}$ where $L_i(x) = j$ implies $\theta_i(x) = \theta_j^*(x)$. The labels are random as are the canonical species curves, each defined over an uncountable set D so, to define a stochastic process, we specify finite dimensional distributions and verify necessary consistency conditions. In particular, for the labels, for any finite set of locations $x_1, \ldots, x_m \in D$, we specify the random distribution $\mathbf{p}_{x_1, \ldots, x_m}$ which is such that $\mathbf{p}_{x_1, \ldots, x_m}(j_1, \ldots, j_m) = P(L(x_1) = j_1, \ldots, L(x_m) = j_m)$. That is $(L(x_1), \ldots, L(x_n))$ is a realization of a multinomial trial driven by the set of probabilities, $\{\mathbf{p}_{x_1, \ldots, x_m}(j_1, \ldots, j_m)\}$. For a single site x, the marginal distribution is a multinomial over the labels, $P(L(x) = j) = \mathbf{p}_x(j)$ for $j = 1, \ldots, k$.

Below, the collection of $\mathbf{p}_{x_1, \ldots, x_m}$ is specified to consistently determine a random probability measure $\mathbf{p}$ on $\{1, \ldots, k\}^D$ by what we define as a Dirichlet labeling process. For the canonical species curves θ_j^*'s, we assume they are i.i.d. GP realizations, again characterized by the finite dimensional multivariate normals for any set of locations, $x_1, \ldots, x_m \in D$. The θ_j^*'s could be modeled as realizations from a more general process on D but this would not provide any benefit within our setting. Again, the smoothness of the θ_j^*'s can be controlled through the choice of covariance function, as noted in the Introduction. We denote the

GP by G_0. Then, formally:

$$\theta_j^* \overset{i.i.d.}{\sim} G_0, \qquad j = 1, \ldots, k,$$

$$L_i | \, \mathbf{p} \overset{i.i.d.}{\sim} \mathbf{p}, \qquad i = 1, \ldots, n,$$

$$\theta_i(x_t) | \, L, \theta^* \, = \, \theta_{L_i(x_t)}^*, \qquad i = 1, \ldots, n; \; t = 1, \ldots, m,$$

$$Y_i(x_t) | \, \theta_i(x_t) \, \sim \, N(\theta_i(x_t), \tau^2), \; i = 1, \ldots, n; \; t = 1, \ldots, m.$$

In addition, depending on the application, there may be prior distributions for G_0 and τ. Also, there may be covariate information, which can be included in the mean for $Y_i(x_t)$.

An alternative representation sacrifices the hierarchical specification through the labels and expresses the model directly through a random finite mixture distribution, $\theta_i \overset{i.i.d.}{\sim} G$ for $i = 1, \ldots, n$, where G is a random measure on $\mathbb{R}^D$ such that

$$G_{x_1,\ldots,x_m} = \sum_{(j_1,\ldots,j_m)\in\{1,\ldots,k\}^m} \mathbf{p}_{x_1,\ldots,x_m}(j_1,\ldots,j_m)\delta_{(\theta_{j_1}^*(x_1),\ldots,\theta_{j_m}^*(x_m))}. \qquad (2.1)$$

See Ishwaran and Zarepour (2002), expressions (1) and (3) in this context.

Regardless, we need to specify $\mathbf{p}$, a random probability measure on $\{1,\ldots,k\}^D$. For locations $x_1,\ldots,x_m$, $\mathbf{p}_{x_1,\ldots,x_m}$ has the k^m-dimensional Dirichlet distribution

$$(\mathbf{p}_{x_1,\ldots,x_m}(j_1,\ldots,j_m), j_i = 1,\ldots,k) \sim Dir(\alpha\mathbf{q}_{x_1,\ldots,x_m}(j_1,\ldots,j_m), j_i = 1,\ldots,k),$$
$$(2.2)$$

where the base measure $\mathbf{q}$ is a probability measure on $\{1,\ldots,k\}^D$.

The base measure $\mathbf{q}$ is constructed such that $\mathbf{q}$ has a uniform marginal distribution at every location $x \in D$, $\mathbf{q}_x(1) = \ldots = \mathbf{q}_x(k) = 1/k$. Additionally, $\mathbf{q}$ inherits the spatial dependence structure exhibited by a stochastic process F on $\mathbb{R}^D$ as we now clarify.

Denote by $F_{x_1,\ldots,x_m}$ the finite-dimensional distributions of F. Let $(\eta(x_1),\ldots,\eta(x_m)) \sim F_{x_1,\ldots,x_m}$, and consider the random vector $(F_{x_1}(\eta(x_1)),\ldots,F_{x_m}(\eta(x_m))) \in [0,1]^m$, where F_{x_t} denotes the cumulative distribution function at location x_t for F. This vector has uniform marginals and induces a joint distribution function denoted by $H_{F,x_1,\ldots,x_m}$. The collection of finite-dimensional d.f. $H_{F,x_1,\ldots,x_m}$ characterizes a probability measure H_F on $[0,1]^D$. Now, let us discretize $[0,1]^m$ into hyper-cubes

$$C_{j_1,\ldots,j_m} = \left(\frac{j_1-1}{k}, \frac{j_1}{k}\right] \times \ldots \times \left(\frac{j_m-1}{k}, \frac{j_m}{k}\right],$$

for $j_i = 1, \ldots, k$. Then, the latent labeling process $\mathbf{q}$ is defined by:

$$\mathbf{q}_{x_1, \ldots, x_m}(j_1, \ldots, j_m) = H_{F, x_1, \ldots, x_m}(C_{j_1, \ldots, j_m}).$$

Remark. (1) The overall model is characterized by a canonical curve distribution G_0 and precision parameter τ, as well as parameters specifying the labeling process $\mathbf{p}$, which is parameterized by labeling process $\mathbf{q}$.

(2) To gain some intuition about the labeling process $\mathbf{q}$, we provide an alternative representation. For each $x \in D$, let $c_1(x), \ldots, c_k(x)$ be an increasing sequence of threshold values in $\mathbb{R}$ such that $F_x(c_j(x)) = j/k$, for $j = 1, \ldots, k-1$. Complement the sequence with $c_0(x) = -\infty$ and $c_k(x) = \infty$. Conditioning on the realization $\eta = (\eta(x_1), \ldots, \eta(x_m))$, define function $Z : D \to \{1, \ldots, k\}$ such that for each $j = 1, \ldots, k$,

$$Z(x) = j \Leftrightarrow \eta(x) \in (c_{j-1}(x), c_j(x)] \Leftrightarrow F_x(\eta(x)) \in (\frac{j-1}{k}, \frac{j}{k}].$$

Hence, an η drawn from the stochastic process F yields a label $Z \sim \mathbf{q}$.

(3) In the foregoing, $\mathbf{q}$ is defined by discretizing auxiliary variables $\eta \sim F$. Then $\mathbf{p}$ is a random draw from the Dirichlet process using base measure $\mathbf{q}$, $\mathbf{p}|\mathbf{q} \sim DP(\alpha\mathbf{q})$. It is simple to show that $\mathbf{p}$ can be defined directly in terms of auxiliary variables ξ without going through the labeling function $Z \sim \mathbf{q}$. First, define a random function ξ on $\mathbb{R}^D$ such that $\xi \sim H$, where $H \sim DP(\alpha F)$ (this is called spatial Dirichlet process in Gelfand, Kottas, and MacEachern (2005). Then, discretize ξ as follows: for any $x \in D$,

$$\tilde{L}(x) = j \Leftrightarrow \xi(x) \in (c_{j-1}(x), c_j(x)] \Leftrightarrow F_x(\xi(x)) \in (\frac{j-1}{k}, \frac{j}{k}]. \qquad (2.3)$$

Marginalizing over ξ and H, we obtain a random probability distribution $\tilde{\mathbf{p}}$ generating $\tilde{L}$. It can be shown that $\mathbf{p} \overset{d}{=} \tilde{\mathbf{p}}$ and $L \overset{d}{=} \tilde{L}$. Indeed, for any $x_1, \ldots, x_m \in D$, the random vector $\tilde{L} = (\tilde{L}(x_1), \ldots, \tilde{L}(x_m))$ has to satisfy, due to the definition of the Dirichlet process,

$$\left(\tilde{\mathbf{p}}(\tilde{L} = (j_1, \ldots, j_m)), j_i = 1, \ldots, k \right)$$

$$\overset{d}{=} \left(\tilde{\mathbf{p}}(\xi(x_1) \in (c_{j_1-1}(x_1), c_{j_1}(x_1)], \ldots, \xi(x_m) \in (c_{j_m-1}(x_m), c_{j_m}(x_m)]), \right.$$

$$\left. j_i = 1, \ldots, k \right)$$

$$\sim Dir\left(\alpha F(\eta(x_1) \in (c_{j_1-1}(x_1), c_{j_1}(x_1)], \ldots, \eta(x_m) \in (c_{j_m-1}(x_m), c_{j_m}(x_m)]), \right.$$

$$\left. j_i = 1, \ldots, k \right)$$

$$= Dir(\alpha\mathbf{q}_{x_1, \ldots, x_m}(j_1, \ldots, j_m), j_i = 1, \ldots, k).$$

This implies that $\mathbf{p} \overset{d}{=} \tilde{\mathbf{p}}$ and $L \overset{d}{=} \tilde{L}$. Although we have shown that there are two equivalent characterizations of $\mathbf{p}$ in terms of latent process ξ, or in terms of latent label function Z, we shall see that the latter characterization is much more convenient to work with. They key point is that properties for the label function Z can be easily obtained and incorporated into that for L, and lead to a computationally efficient inference algorithm to be described in Section 6.

We conclude this section with some words regarding k. It is worth asking whether it is realistic to assume that k is fixed, or would one expect that identifying the number of canonical curves should be part of the problem. Perhaps the latter is more likely to be the case but, due to the complexity of the model and the challenges to fit it, even with k fixed, we decided not to pursue, for example, some sort of reversible jump algorithm to allow k to be random. In this regard, we prefer to perform model comparison to choose k or study sensitivity of clustering to the choice of k, fitting models for several fixed k's. In fact, this is what we have done with the examples in Section 7 below. We can report that, not surprisingly, a bigger k encourages more clusters but this is also mediated by the specification of the precision parameter in the labeling process. In practice, we can hope that a given application will offer some suggestion of what k's are interesting so that we can investigate model comparison for such k's. In general, as noted in the Introduction, we envision our modeling to be most useful when k is small relative to n. Moreover, the theoretical analysis in Section 4 also suggests that k should grow very slowly relatively to n to ensure strong consistency of relevant posterior distributions.

3. Properties of the Labeling Process

As is clear from the previous section, we use the label process as a prior within the hierarchical model given at the beginning of Section 2. Here we examine properties of this process: the random label functions L and Z on $\{1, \dots, k\}^D$, where $L \sim \mathbf{p}$ and $Z \sim \mathbf{q}$, as well as that of the hybrid curve realization $\theta \sim G$.

Properties of p. From (2.2), $\mathbf{p}$ and $\mathbf{q}$ are related, $\mathbf{p}|\mathbf{q} \sim DP(\alpha\mathbf{q})$. As a result, properties obtained for the labeling process Z can be easily incorporated into those for L. We start with elementary properties for $\mathbf{p}$ that are simple consequences of our use of the Dirichlet distribution:

Proposition 1.
(a) *Let $L \sim \mathbf{p}$. For any $x \in D$, the distribution for the label $L(x)$ is a k-dimensional multinomial trial with probabilities $\mathbf{p}_x \sim Dir((\alpha/k)\mathbf{1})$.*

(b) *Let $L_1, L_2|\mathbf{p} \overset{i.i.d.}{\sim} \mathbf{p}$. Then, unconditionally, $P(L_1(x) = L_2(x)) = 1/k + (1 - 1/k)/(\alpha + 1)$.*

(c) *Let $L_1, L_2 | \mathbf{p} \overset{i.i.d.}{\sim} \mathbf{p}$, and $x_1, \ldots, x_m \in D$. Then, unconditionally,*

$$P(L_1(x_1, \ldots, x_m) = L_2(x_1, \ldots, x_m)) = \frac{1}{\alpha+1} + \frac{\alpha}{\alpha+1} \sum_{j_1, \ldots, j_m} \mathbf{q}_{x_1, \ldots, x_m}(j_1, \ldots, j_m)^2.$$

(d) *Let $L \sim \mathbf{p}$ and $x_1, \ldots, x_m \in D$. Then*

$$P(L(x_1) = j_1 | L(x_2) = j_2, \ldots, L(x_m) = j_m) = \frac{\mathbf{q}_{x_1, \ldots, x_m}(j_1, j_2, \ldots, j_m)}{\mathbf{q}_{x_2, \ldots, x_m}(j_2, \ldots, j_m)}.$$

Proposition 1 shows how the clustering behavior exhibited by the label replicates $L_i \sim \mathbf{p}$ is driven by the concentration parameter α and the labeling process $\mathbf{q}$. (In particular, as $\alpha \to \infty$, $\mathbf{p}$ behaves more like the base measure $\mathbf{q}$.) It is worth noting the distinction between local and global clustering behavior implicit in the labeling process $\mathbf{p}$. Since the probabilities $\mathbf{q}_{x_1, \ldots, x_m}(\cdot)$ are of order $O(1/k^m)$ (cf., Proposition 3 and the Appendix), part (c) implies the global clustering probability $P(L_1 = L_2) \sim 1/(\alpha+1) + \alpha/(\alpha+1) \cdot 1/k^m \to 1/(\alpha+1)$ as $m \to \infty$. On the other hand, at each local site x, the probability of clustering is substantially higher:

$$P(L_1(x) = L_2(x)) = \frac{1}{\alpha+1} + \frac{\alpha}{\alpha+1} \cdot \frac{1}{k}.$$

(Due to the discreteness of L, this probability is greater than $1/(\alpha+1)$, the usual probability of a tie for a continuous variable.) However, since the probability of global clustering is still $1/(\alpha+1)$, there are evident implications regarding either the specification of α or a prior for it.

When $k \to \infty$, the distinction between global and local clusters is apparently lost: two realizations L_1 and L_2 are either identical everywhere, or nowhere at all. Although the "hard" clustering behavior is lost, the "soft" clustering behavior remains in play, being driven by $\mathbf{q}$ which is, in turn, regulated by F.

Properties of q. In the sequel, we assume that F is a mean-zero, isotropic Gaussian process $GP(0, 1, \phi_L)$ with covariance function of the form, $\rho_{12}(x_1, x_2) = cov(\eta(x_1), \eta(x_2)) = exp(-\phi_L \|x_1 - x_2\|)$ for any $x_1, x_2 \in D$, where $\phi_L > 0$ is called the decay parameter. (We can set the process variance to 1 w.l.o.g.) Under the assumptions on F, the quantile threshold functions $c_j(x)$ are constant with respect to x and the sequence $c_0, \ldots, c_k$ satisfies $\Phi(c_j) = j/k$ where Φ is the c.d.f. of the standard normal variable.

To denote the dependence of labeling process $\mathbf{q}$ on ϕ_L and k, we write $\mathbf{q}(\phi_L, k)$. Although it is easy to generate a random sample of $(Z(x_1), \ldots, Z(x_m)) \sim \mathbf{q}$, the distribution function for $\mathbf{q}$ is generally not available in closed form. In fact, the next result presents a closed form for $k = 2$ and for any two locations, but closed form expressions for $k > 2$ are not readily available.

Proposition 2. ($k = 2$). *Let $Z \sim \mathbf{q}(\phi_L, 2)$ and let $\rho_{12} = \mathrm{Cov}(\eta(x_1), \eta(x_2))$. Then*

$$P(Z(x_1) = 1, Z(x_2) = 2) = P(Z(x_1) = 2, Z(x_2) = 1)$$

$$:= \mathbf{q}_{x_1, x_2}(1, 2) = \frac{1}{\pi} \arccos\left(\frac{1}{2} + \frac{\rho_{12}}{2}\right)^{1/2},$$

$$P(Z(x_1) = 1, Z(x_2) = 1) = P(Z(x_1) = 2, Z(x_2) = 2)$$

$$:= \mathbf{q}_{x_1, x_2}(1, 1) = \frac{1}{2} - \frac{1}{\pi} \arccos\left(\frac{1}{2} + \frac{\rho_{12}}{2}\right)^{1/2}.$$

It is simple to observe that, as either $\phi_L \to \infty$ or $\|x_1 - x_2\| \to \infty$, $\rho_{12} \to 0$ so that both probabilities $\mathbf{q}_{x_1, x_2}(1, 2)$ and $\mathbf{q}_{x_1, x_2}(1, 1)$ tend to $1/4$. That is, $Z(x_1)$ and $Z(x_2)$ become independent. On the other hand, as $\phi_L \to 0$ or $\|x_1 - x_2\| \to 0$, $Z(x_1)$ and $Z(x_2)$ are equal with increasing probability.

For large k, it is possible to obtain a good approximation to the likelihood function using Riemann sum approximation.

Proposition 3.(k is large). *Let $Z \sim \mathbf{q}(\phi_L, k)$. For any $i, j \leq k$ such that c_i and c_j do not diverge to either $+\infty$ or $-\infty$ as $k \to \infty$,*

$$P(Z(x_1) = i, Z(x_2) = j) = q_{x_1, x_2}(i, j) = \frac{1}{k^2}(R_{ij}(c_i, c_j) + o(1)), \qquad (3.1)$$

$$P(Z(x_1) = i, Z(x_2) > j) = \frac{1}{k}\left(1 - \Phi\left(\frac{c_j - c_i \rho_{12}}{1 - \rho_{12}^2}\right) + o(1)\right), \qquad (3.2)$$

where the $o(1)$ terms tend to 0 uniformly for all such (i, j), and

$$R_{ij}(c_i, c_j) = \frac{1}{\sqrt{1 - \rho_{12}^2}} \exp{-\frac{(c_i^2 + c_j^2)\rho_{12}^2 - 2\rho_{12}c_i c_j}{2(1 - \rho_{12}^2)}}. \qquad (3.3)$$

Proposition 3 can also be extended to an arbitrary number of locations $x_1, \ldots, x_m$, and can be used to obtain conditional probabilities (see the Appendix).

It is useful to examine the intuitive behavior of the $\mathbf{q}$ probabilities for a fixed k as derived by Proposition 3. As $\rho_{12} \to 0$, we have $R_{ij}(c_i, c_j) \to 1$, so that $P(Z(x_1) = i, Z(x_2) = j) \to 1/k^2$, i.e., $Z(x_1)$ and $Z(x_2)$ become less dependent. On the other hand, as $\rho_{12} \to 1$, for any pair $i \neq j$, $R_{ij}(c_i, c_j) \to 0$, i.e., $Z(x_1)$ and $Z(x_2)$ take different values i and j with probability converging to 0. Accordingly, $Z(x_1) = Z(x_2)$ with probability converging to 1. Now, fixing ρ_{12} and c_i, consider $P(Z(x_2)|Z(x_1) = i) \approx (1/k)R_{ij}(c_i, c_j)$. $R_{ij}(c_i, c_j)$ achieves maximum at $c_j = \rho_{12}c_i$. In particular, when x_2 is near x_1, $\rho_{12} \approx 1$, so that $\rho_{12}c_i \approx c_i$, the conditional distribution $P(Z(x_2)|Z(x_1) = i)$ favors values that are near i. For

most of the nodes that are distant, so that $\rho_{12} \approx 0$, the conditional distribution is rather flat even though the mode $\rho_{12}c_i \approx 0$. For nodes in the middle range so that say, $\rho_{12} \approx 1/2$, there is an interesting shrinkage effect pulling $Z(x_2)$ toward the middle value (between $k/2$ and i). In addition, variable $Z(x_2)$ tends to take values that are farther away from i with decreasing probabilities.

Turning to the properties of a "hybrid" curve realization θ that is drawn from the random probability measure G (see (2.1)), we have

$$G_{x_1,\ldots,x_m} = \sum_{(j_1,\ldots,j_m)\in\{1,\ldots,k\}^m} \mathbf{p}_{x_1,\ldots,x_m}(j_1,\ldots,j_m)\delta_{(\theta_{j_1}^*(x_1),\ldots,\theta_{j_m}^*(x_m))},$$

where the randomness of G is due to the randomness of $\mathbf{p}$ and θ^*. We assume that the θ^* are very smooth curves by placing a zero-mean Gaussian process prior G_0 on θ^*, with covariance function $\rho_\theta(x_1,x_2) = \sigma_\theta^2 \exp{-\phi_\theta\|x_1 - x_2\|^2}$ (other choices of covariance function could be adopted depending on the application). It is simple to obtain that

$$\mathbb{E}[\theta(x)|\mathbf{q}, G_0] = \mathbb{E}[\theta^*(x)|G_0] = 0,$$

$$\mathbb{E}[\theta(x_1)\theta(x_2)|\mathbf{q}, G_0] = \sum_{j=1}^k \mathbf{q}_{x_1,x_2}(j,j)\mathrm{Cov}(\theta^*(x_1), \theta^*(x_2)).$$

As $\|x_1 - x_2\| \to \infty$, $\mathrm{Cov}(\theta^*(x_1), \theta^*(x_2)) \to 0$, so $\mathrm{Cov}(\theta(x_1), \theta(x_2)|\mathbf{q}, G_0) \to 0$. As $\|x_1 - x_2\| \to 0$, $\sum_{j=1}^k \mathbf{q}_{x_1,x_2}(j,j) \to 1$, so $\mathrm{Cov}(\theta(x_1), \theta(x_2)|\mathbf{q}, G_0) \to \sigma_\theta^2$. Formally, it can be shown that the hybrid species $\theta \sim G$ is mean square continuous:

Proposition 4. *Suppose that G_0 has bounded mean and variance functions, and $F(x)$ has non-atomic distribution for any $x \in D$. If both G_0 and F are mean square continuous, so is G.*

Although θ is mean square continuous, each realization is almost surely discontinuous as it is composed of multiple smooth segments of the canonical curves. Again, the Y_i's arise by the mixing with the noise or pure error process, i.i.d. random variables at locations $\epsilon(x) \sim N(0, \tau^2))$, to obtain $Y(x) = \theta(x) + \epsilon(x)$ for any $x \in D$. The joint density for $\mathbf{Y} = (Y(x_1),\ldots,Y(x_m))$ given G and τ^2 is

$$f(\mathbf{Y}|G, \tau) = \int N_m(\mathbf{Y}|\theta, \tau^2\mathbf{I}_m)G(d\theta). \tag{3.4}$$

It follows that $\mathbb{E}(\mathbf{Y}|\mathbf{q}, G_0) = \mathbb{E}(\theta^*|G_0)$ and the covariance matrix $\mathbf{\Sigma}_{\mathbf{Y}}|\mathbf{q}, G_0 = \tau^2\mathbf{I}_m + \mathbf{\Sigma}_\theta$, where $(\mathbf{\Sigma}_\theta)_{ij} = \mathrm{Cov}(\theta(x_i), \theta(x_j)|\mathbf{q}, G_0)$.

4. Model Identifiability and Posterior Consistency

The described labeling process provides a highly flexible nonparametric prior for modeling collections of curves. As is generally the case with high-dimensional mixture models, model identifiability and posterior consistency issues arise. Moreover, as we shall demonstrate, understanding of these issues can be useful for prior specification. Here, we restrict our attention to the induced distribution of the m-variate $\mathbf{Y} = (Y(x_1), \ldots, Y(x_m))$ through the associated mixing distribution $G_{x_1,\ldots,x_m}$. Treatment for the functional case is more demanding and will be pursued elsewhere, but our discussion of the multivariate case should provide some hints for the model behavior and the issues involved as $m \to \infty$.

For a fixed k, the induced distribution on $\mathbf{Y}$ can be viewed as a finite mixture of m-variate normal vectors with $N = k^m$ mixture components, where the mixing parameter $\mathbf{p}$ is endowed with a Dirichlet distribution prior, and the normal means are parameterized by the k canonical species curves. When k is unknown, an approach that has become common is to consider a prior that corresponds to the limit of the finite mixture model as $k \to \infty$. In light of the results given by Proposition 3, it can be shown that as $k \to \infty$, $G_{x_1,\ldots,x_m}$ converges in distribution to $G \sim DP(\alpha F)$ (by applying Theorem 2(a) of Petrone, Guidani, and Gelfand (2009)) In fact, as far as the marginal density of vector $\mathbf{Y}$ is concerned, the finite mixture (with N components) provides a remarkably tight approximation to the DP limit. Let $\Pi_{n,k}$ denote the marginal density of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ which is induced by our prior distributions of (G, τ) for some finite k, and $\Pi_{n,\infty}$ the marginal density of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ using prior distributions with $k \to \infty$. It is shown by Ishwaran and James (2001) that the L_1 distance $\|\Pi_{n,k} - \Pi_{n,\infty}\|_1 \sim 4n \exp(-(N-1)/\alpha)$. Because $N = k^m$ grows very fast with k, in practice the choice of k has little effect on the approximation of the marginal distribution of $\mathbf{Y}$. Rather, the choice of k hinges more on the interpretation of the canonical species vectors $\theta_1^*, \ldots, \theta_k^*$. As we see in Section 7, in an example with progesterone data analysis, we are interested in $k = 2$, whereas in another example with image analysis, $k = 8$ turns out to be sufficient for our segmentation and clustering application.

Model identifiability.

Our model is a finite mixture with mixing distribution $G_{x_1,\ldots,x_m}$. A multivariate version of Theorem 2 of Ishwaran and Zarepour (2002), shows that it is fully identified under mild conditions.

Proposition 5. *Let $\psi(\mathbf{Y}|\theta, \tau)$ denote the m-variate normal density with mean θ and covariance matrix $\tau^2 I_m$. Let $G_{x_1,\ldots,x_m}$ be a mixing distribution defined (2.1) for some fixed canonical vectors $\theta_1^*, \ldots, \theta_k^*$ for some finite k, and positive mixing proportions $\mathbf{p}_{x_1,\ldots,x_m}(\cdot)$. Given some $\tau^* > 0$, suppose there is a distribution G_1 over $\mathbb{R}^m$ and π_1 over $\mathbb{R}_+$ such that*

(a) $\int \psi(\mathbf{Y}|\theta,\tau^*)G_{x_1,\ldots,x_m}(d\theta) = \int \psi(\mathbf{Y}|\theta,\tau)G_1(d\theta)\pi_1(d\tau)$ *for almost all* $\mathbf{Y} \in \mathbb{R}^m$,

(b) *Under* $G_1 \times \pi_1$, $\mathbb{E}\exp\{1/[2(\tau^{*2} - \tau^2)]\}\sum_{r=1}^{m}\theta(x_r)^2 < \infty$,

(c) $\theta_i^*(x_r) \neq \theta_j^*(x_r)$ *for* $i \neq j; r = 1,\ldots,m$.
Then we have $G_1 = G_{x_1,\ldots,x_m}$, *and* $\pi_1(\cdot) = \delta_{\tau^*}$.

Note that the mixing distribution $G_{x_1,\ldots,x_m}$ is parameterized in terms of the canonical curves $\theta_1^*,\ldots,\theta_k^*$. Though the above result shows that G is identifiable, it does not necessarily establish that these canonical curves (m-dimensional vectors) can actually be determined. This issue of parameter determinacy can be resolved by incorporating additional assumptions on the prior distributions on the canonical curves. As a simple example, if the canonical curves are non-identical constant functions, then the identifiability of $G_{x_1,\ldots,x_m}$ trivially implies the identifiability of the each individual canonical curves. More generally, we could envision identifiability conditions for the functional case that requires "distinguishable" canonical curves based on smoothness criteria. In addition to the determination of canonical curves, label switching among the canonical curves is an issue commonly encountered in mixture models. Again, these issues can be resolved in practice by introducing additional, e.g., ordering constraints to the parameters. A more detailed discussion of parameter determinacy is deferred to the next section.

Posterior consistency. Turning to asymptotic analysis of the posterior distribution derived from our labeling process prior, we view the use of a finite dimensional Dirichlet prior in a finite mixture model as a Bayesian method of sieves by allowing k to grow with sample size n. A similar viewpoint was adopted by Ishwaran and Zarepour (2002) in their analysis of finite mixture for univariate normal variables. Indeed, with care, their results could be adapted to obtain ours.

We consider first the distribution of the m-variate $\mathbf{Y} = (\mathbf{Y}(x_1),\ldots,\mathbf{Y}(x_m))$ whose (conditional) density $f(\mathbf{Y}|G,\tau)$ is given at (3.4). Because the conditioned G and τ are random and endowed with prior distributions of their own, we can view the density f (with the conditioning notations G and τ being dropped) as a random element in a set $\mathcal{F}_k$ of densities of form (3.4) for some k and some realization of G and τ.

As before, $\Pi_{n,k}$ is used to denote the induced prior on f. The posterior distribution of density f is a random measure denoted by $\Pi_{n,k}(\cdot|\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$, and has the following form, for any measurable subset B of $\mathcal{F}_k$:

$$\Pi_{n,k}(B|\mathbf{Y}_1,\ldots,\mathbf{Y}_n) = \frac{\int_B \prod_{i=1}^n f(\mathbf{Y}_i)d\Pi_{n,k}(f)}{\int \prod_{i=1}^n f(\mathbf{Y}_i)d\Pi_{n,k}(f)}.$$

Suppose that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are i.i.d. draws from some $f_0 \in \mathcal{F}_{k^*}$, where f_0 is defined in terms of some realization of $G = G^*, \tau = \tau^*$ for some fixed but possibly unknown value of $k = k^*$. The posterior distribution of $\mathbf{Y}$ is strongly consistent if, for any $\epsilon > 0$, as n tends to infinity with k growing at an appropriate rate, we have

$$\Pi_{n,k}(\{f \in \mathcal{F}_k : \|f - f_0\|_1 < \epsilon\}|\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \to 1 \ P_{f_0} \ a.s..$$

There is a rich body of work on posterior asymptotics for nonparametric Bayesian models, and Dirichlet process mixture models in particular (see, e.g., Ghosal (2007) for an elegant exposition). Most relevant to our model are the analyses of Ghosal, Ghosh, and Ramamoorthi (1999) for Dirichlet process mixture models and Ishwaran and Zarepour (2002) for finite normal mixtures, both focusing on univariate distributions, and a more recent work extending to multivariate density estimation (Wu and Ghosal (2009)). As with these analyses we follow the now standard approach developed by Schwartz (1965); Barron, Schervish, and Wasserman (1999); Ghosal, Ghosh, and Ramamoorthi (1999) and several others, which requires meeting the two sufficient conditions: (A) The prior $\Pi_{n,k}$ is *information dense* around the true density f_0 (that is, $\Pi_{n,k}$ places positive mass on each Kullback-Leibler neighborhood of f_0), and (B) the prior $\Pi_{n,k}$ puts most of its mass around a "small" subset $\mathcal{F}_{n,k} \subset \mathcal{F}_k$, where the size of $\mathcal{F}_{n,k}$ can be measured by the entropy number. In fact, condition (A) is guaranteed by the following lemma.

Lemma 6. *Let $D(\cdot||\cdot)$ denote the Kullback-Leibler divergence between two probability densities. Assume that measure G_0 of the canonical vectors $(\theta^*(x_1), \ldots, \theta^*(x_m))$ places positive density in a rectangle containing the support of G^*. For sufficiently large k, $\Pi_{n,k}(f \in \mathcal{F}_k : D(f_0||f) < \epsilon) > 0$ for any $\epsilon > 0$.*

It is worth noting that our proof for this Lemma (see the Appendix) exploits specifically the choice of the finite Dirichlet prior for $\mathbf{p}$ as defined by (2.2) whose base measure is specified by $\mathbf{q}$. The proof, however, does not extend to the functional case (by letting m be arbitrary). Moreover, it appears that to obtain the denseness properties in the functional setting, additional assumptions on the true density f_0 are needed. For instance, one might need a condition to the effect that the majority of the hybrid curves do not switch very often, so that the technique used in our proof can be applied for arbitrary m.

Turning to condition (B), for given positive numbers δ, a_n, τ_n, we take $\mathcal{F}_{n,k}$ to consist of all densities $f \in \mathcal{F}_k$ whose associated mixing distribution $G_{x_1, \ldots, x_m}$ has support bounded within $[-a_n, a_n]^m$ with probability at least $1 - \delta$, with variance τ restricted to the interval $[\tau_n, M]$, assuming that the prior for τ is some distribution truncated to the right at M. Define the entropy number $J(\delta, \mathcal{F}_{n,k})$ to

be the logarithm of the minimum of all r such that there exists $f_1, \ldots, f_r \in \mathcal{F}_{n,k}$ with the property $\mathcal{F}_{n,k} \subseteq \cup_{i=1}^{r} \{f : \|f - f_i\|_1 < \delta\}$. The collection $\{f_1, \ldots, f_r\}$ is called a covering of $\mathcal{F}_{n,k}$.

Lemma 7. *Assume that $a_n > M/\sqrt{\delta}$. Then $J(6m\delta, \mathcal{F}_{n,k}) \leq k^m \log(1 + m\delta)/(m\delta)$ $+ km \log(1 + (2a_n)/(\tau_n m\delta^2))$.*

This lemma says that the entropy of $\mathcal{F}_{n,k}$ can be controlled by k, a_n, τ_n. Combining the last two lemmas and Theorem 2 of Ghosal, Ghosh, and Ramamoorthi (1999), the following result is immediate.

Proposition 8. *Suppose that τ has support in $[0, M]$ and the canonical curves (vectors) have prior distribution G_0. If for each $\delta > 0, \beta > 0$, there exists constants β_0, β_1, and sequences $a_n \to \infty$ and $\tau_n \to 0$, $k \to \infty$ such that*

(i) *for some β_0, $G_0(\theta^*(x_i) \in [-a_n, a_n]$ for $i = 1, \ldots, m) \geq 1 - \exp(-n\beta_0)$,*

(ii) *under the prior of τ, $P(\tau < \tau_n) \leq \exp(-n\beta_1)$,*

(iii) *$k^m \log(1 + m\delta)/(m\delta) + km \log(1 + 2a_n/(\tau_n m\delta^2)) \leq n\beta$,*

(iv) *conditions of Lemma 6 and Lemma 7 hold,*
    then the posterior distribution of $\mathbf{Y}$ is strongly consistent at f_0.

If, for instance, θ^* has a Gaussian prior distribution, and τ has an inverse gamma distribution truncated to the right, then we can allow $a_n \sim \sqrt{n}$ and $\tau_n \sim 1/\sqrt{n}$. Then, if k is allowed to grow at a rate slower than $n^{1/m}$, all conditions of the proposition hold, yielding the strong consistency of the posterior distribution of $\mathbf{Y}$. Finally, consistency results can also be extended to that of the mixing distribution $G_{x_1, \ldots, x_m}$ (see Ishwaran and Zarepour (2002), Theorem 7, and Ghosal, Ghosh, and Ramamoorthi (1999, p.151)).

5. Parameter Identifiability

The previous section focused on large sample properties of the posterior distribution of the curves, and the identifiability of the mixing distribution $G_{x_1, \ldots, x_m}$. In specific applications we are usually concerned with the identifiability (determinacy) of certain parameters and latent variables of interest, under a limited supply of data. Indeed, the foregoing discussion provides some hints on the roles of certain parameters controlling the smoothness of canonical curves θ^* and the labeling allocation probabilities $\mathbf{p}$. We examine these issues in more detail here. Section 3 discussed the roles of the concentration parameter α and the labeling decay parameter ϕ_L on both the global and local clustering behavior exhibited by the label realization $L \sim \mathbf{p}$. Here we focus on the effects of the prior of ϕ_L,

canonical curves θ^*, and the precision parameter τ on the determinacy of the labeling L and canonical curves θ^*.

Suppose that we are interested in a representation that achieves dimensionality reduction, with the goal of inferring both canonical curves θ^* and labeling $L_1, \ldots, L_n$ for observed replicates $Y_1, \ldots, Y_n$. In this scenario the canonical curves can be viewed as basis functions with the label vectors $L_1, \ldots L_n$ providing coefficients with respect to such bases. When the number of canonical curves k is small, the canonical curves are expected to represent "canonical" patterns for the whole collection of curves. As noted in the Introduction, the variance parameter τ plays an important role in the identifiability of the canonical curve θ^*. When τ is large, the learned canonical curves become very smooth but weakly distinguishable. By contrast, when τ is small, the canonical curves are less smooth and more distinguishable, as their respective posteriors cover different regions in the function space spanned by the curve collection. This phenomenon is illustrated in Section 7.

ϕ_L also plays an important role in the identifiablity of the canonical species curves θ_j^*. When ϕ_L is close to 0, as shown by Prop 2 and Prop 3, the hybridization (label switching) within each individual curve is discouraged – the model essentially insists on global clustering. If the curve collection can indeed be clustered globally in terms of canonical curves, these are strongly identifiable. On the other hand, if the curve realizations tend to switch often among the canonical curves, corresponding to large ϕ_L, or canonical curves are not very smooth, we observe that the canonical curves become more weakly identified. As we illustrate in Section 7, our model is able to recover segments of locations that admit relatively few switchings among relatively smooth canonical curves. In particular, similar locations tend to be (correctly) assigned the same labels, but it is possible that whole segment is incorrectly labeled relatively to some other segments.

Suppose, on the other hand, that we are less interested in inferring about the canonical curves, but more about the labeling realizations $L_1, \ldots, L_n$ as a means for characterizing and clustering the observed replicates $Y_1, \ldots, Y_n$. In this scenario, strong constraints can be imposed upon θ^* to improve the identifiability of labels L_i's. In the image segmentation application we present, an image can be viewed as being composed of different objects (grass, plants, buildings, animals, human faces, etc), each of which is associated with a level set corresponding to a (random) level of light intensity. Thus, canonical curves θ^* can be taken to be random constant functions. Furthermore, additional order constraints can be imposed according to label values $\{1, \ldots, k\}$. The previous discussion on properties of $\mathbf{p}$ and $\mathbf{q}$ suggests that for large k there is a natural ordering of label values $\{1, \ldots, k\}$. That is, locations near to each other have high probability of

sharing similar labels, i.e., labels j_1 and j_2 such that $|j_1 - j_2|$ is small. It is natural to assign more extreme ranges for priors to extreme labels such as 1 and k. We could even specify that $\mathbb{E}(\theta_1^*) < \mathbb{E}(\theta_2^*) < \ldots < \mathbb{E}(\theta_k^*)$. Note that such ordering constraints are not necessary to ensure model identifiability, but they would be expected to improve the mixing for simulation-based posterior inference.

6. Model Fitting and Inference

Using the bracket notation, the joint distribution associated with the model presented at the start of Section 2 is

$$\prod_{i=1}^{n}[Y_i|L_i, \theta_1^*, \ldots, \theta_k^*, \tau] \times \prod_{j=1}^{k}[\theta_j^*|\sigma_\theta, \phi_\theta] \times [L_1, \ldots, L_n|\phi_L, \alpha] \times [\phi_L] \times [\alpha] \times [\tau] \times [\phi_\theta] \times [\sigma_\theta].$$

In this expression we have implemented the usual marginalization over $\mathbf{p}$, with $\mathbf{q} \equiv \phi_L$, $[L_1, \ldots, L_n|\phi_L, \alpha] = \int \prod_{i=1}^{n}[L_i|\mathbf{p}][\mathbf{p}|\phi_L, \alpha]d\mathbf{p}$.

In this section we develop an algorithm for fitting the model and for inference regarding the parameters of interest. We use Gibbs sampling to draw from $[L_1, \ldots, L_n, \theta_1^*, \ldots, \theta_k^*, \phi_L, \alpha, \tau, \phi_\theta, \sigma_\theta|\mathbf{Y}]$. The updates of parameters $\alpha, \tau, \phi_\theta, \sigma_\theta$ are standard, see ,e.g., Duan, Guidani, and Gelfand (2007). For canonical curves, under a Gaussian process, the prior for vector $\theta_j^* = (\theta_j^*(x_1), \ldots, \theta_j^*(x_m))$ is normal with mean μ_j and covariance matrix $\Sigma_{\theta_j^*|\sigma_\theta, \phi_\theta}$. Let I_{ij} be an $m \times m$ diagonal matrix whose t-th entry is $\mathbb{I}(L_i(t) = j)$. The full conditional for θ_j^* has the form

$$[\theta_j^*|Y_1, \ldots, Y_n, L_1, \ldots, L_n, \phi_\theta, \sigma_\theta] \sim N\left(\frac{1}{\tau^2}\Lambda\sum_{i=1}^{n}I_{ij}Y_i + (\Sigma_{\theta_j^*|\sigma_\theta, \phi_\theta})^{-1}\mu_j, \Lambda\right),$$

where $\Lambda = ((\Sigma_{\theta_j^*|\sigma_\theta, \phi_\theta})^{-1} + (1/\tau^2)\sum_{i=1}^{n}I_{ij})^{-1}$.

We now turn our attention to updating label vectors $L_i, i = 1, \ldots, n$ and decay parameter ϕ_L. Due to the alternative characterization of latent labels L captured by (2.3), one simple method is to directly sample the latent variables $\xi_i \sim H$, where $H \sim DP(\alpha F_{\phi_L})$. The label vector L_i is then obtained by thresholding ξ_i. Although the full conditional distribution for ξ_i can in principle be obtained by the standard Polya urn scheme, it is simple to observe that at each iteration one has to compute an intractable sum of k^m terms. To overcome this difficulty, a simple heuristic is to introduce an auxiliary variable $\tilde{\xi}_i$, a perturbed version of ξ by a small independent noisd: $\tilde{\xi}_i = \xi + \epsilon$, where $\epsilon \sim N(0, \gamma^2 I_m)$ and I_m is an $m \times m$ identity matrix. For small γ^2, it is expected that $\tilde{\xi}_i$ and ξ_i belong to the same thresholded hypercubes with high probability. Thus, the label vector L_i can be obtained by thresholding $\tilde{\xi}_i$ instead of ξ_i. Vector ξ_i can now be updated independently of the data via Pólya's urn scheme, while $\tilde{\xi}_i$ can be updated conditionally component-by-component via truncated univariate normals.

The problem with this approach is sensitivity of the perturbation noise σ to the varying size of different thresholded hypercubes, especially when k is moderate or large. Moreover, sampling over continuous and high-dimensional latent vectors $\tilde{\xi}$ and ξ could be very inefficient and, as we shall see, is unnecessary.

Our approach relies on the characterization of L_i in terms of label vectors $_i \sim \mathbf{q}$ and the latent vector η_i for $i = 1, \ldots, n$. Furthermore, by the virtue of Proposition 3 (and its extension for any m, see the Appendix), the latent η_i can be easily marginalized so the overall mixing can be significantly improved. Thanks to our choice of the Dirichlet prior, the Gibbs sampling procedure is now straightforward by applying the Pólya urn sampling scheme. Here, we have for, say, curve 1 at x_1, that the conditional label distribution is

$$P(L_1(x_1)|L_1(x_2), \ldots, L_1(x_m), \text{the rest}) \propto \sum_{i=2}^{n} \frac{\mathbb{I}(L_1 = L_i)}{\alpha + n - 1} N(Y_1|\theta_{L_i}) +$$

$$\frac{\alpha}{\alpha + n - 1} N(Y_1|\theta_{L_1}) \mathbf{q}_{x_1, \ldots, x_m}(L(x_1), \ldots, L_1(x_m)|\phi_L, k).$$

The likelihood function under $\mathbf{q}$ is obtained via Proposition 3. This likelihood also provides means for updating ϕ_L via a standard Metropolis step. One possible issue is that the approximation of the likelihood function for $\mathbf{q}$ is not expected to be accurate for small value of k. In particular, the distribution function and relevant conditional probabilities for the labeling process $\mathbf{q}$ are not available in closed form. For the remainder of this section we develop approximate inference methods for the latent labeling process $\mathbf{q}$ for small k. We illustrate with $k = 2$.

Turning first to estimation of the ϕ's, we seek inference for ϕ_L given i.i.d. label realizations $Z_1, \ldots, Z_n$ drawn from $\mathbf{q}$, observed values at locations $x_1, \ldots, x_m$. We first consider the point estimation problem for ϕ_L. Suppose we have multiple curves, indexed by $i = 1, \ldots, n$, observed at $m = 2$ locations x_1 and x_2 only. In this scenario, one can use a maximum likelihood method to obtain a consistent estimate for ϕ_L:

$$\hat{\phi}_L = \operatorname{argmax}_{\phi_L \geq 0} \prod_{i=1}^{n} \mathbf{q}(Z_i(x_1), Z_i(x_2)),$$

where the d.f $\mathbf{q}$ for $k = 2$ is available in the closed form given in Proposition 2. The more typical scenario, however, is when m much larger than 2 and the sample size n is small. For simplicity of exposition, suppose that $n = 1$. How can one estimate ϕ_L given a *single* realization of random curve z evaluated at m locations $x_1, \ldots, x_m$: $z = (z(x_1), \ldots, z(x_m))$? An intuitive approach is to maximize a pseudo-likelihood for z that is obtained by taking the product of all pairwise likelihood functions. Simulation work indicates that this is a good estimator (see Table 1 for an illustration).

Table 1. Mean and variance of the maximum likelihood estimate (for one edge) and maximum pseudo-likelihood estimates for ϕ_L. n denotes sample size, m denotes the number of locations in a equally spaced grid in $\mathbb{R}^2$. The data is drawn from $\mathbf{q}(\phi_L, k)$ with $\phi_L = 0.5$, $k = 2$.

	OneEdge MLE	$m = 4$	$m = 36$	$m = 100$
$n = 1$	N/A	2.26 + 2.55	0.60 + 0.35	0.51 + 0.23
$n = 10$	2.03 + 8.03	0.64 + 0.43	0.48 + 0.15	0.51 + 0.06
$n = 20$	2.57 + 7.90	0.63 + 0.31	0.51 + 0.09	0.50 + 0.04
$n = 40$	0.48 + 0.30	0.53 + 0.19	0.51 + 0.05	0.51 + 0.04
$n = 60$	0.53 + 0.20	0.54 + 0.16	0.50 + 0.05	0.51 + 0.03
$n = 80$	0.52 + 0.33	0.51 + 0.16	0.50 + 0.04	0.50 + 0.02
$n = 100$	0.49 + 0.16	0.50 + 0.14	0.49 + 0.03	0.50 + 0.02

Proposition 9. *Suppose that $Z = (Z(x_1), \ldots, Z(x_m))$ is drawn from $\mathbf{q}(\phi_L^*, 2)$ via $F_{\phi_L^*}$ for some $\phi_L^* > 0$. Let r_m be the number of pairs of (x_i, x_j) s.t. $\|x_i - x_j\| \leq d_0$ for some $d_0 > 0$. Then, for*

$$\hat{\phi}_L = \text{argmax}_{\phi \geq 0} \prod_{1 \leq i < j \leq m} \mathbf{q}(Z(x_i), Z(x_j))|\phi),$$

we have that $|\hat{\phi}_L - \phi_L^*| = O(\sqrt{m/r_m})$ in probability.

Though the proof is provided for $k = 2$ it can be easily extended for $k > 2$.

Suppose now that ϕ_L is endowed with a prior distribution $\pi(\phi_L)$ on a bounded interval $[\phi_1, \phi_0]$. We are interested in sampling the posterior distribution for ϕ_L given values of the label $Z = (z(x_1), \ldots, z(x_m))$. We propose to use the aforementioned pseudo-likelihood to obtain what we term a "Gibbs posterior" distribution Zhang (2006) for ϕ_L as

$$P_\lambda(\phi_L|Z) \propto \prod_{1 \leq i < j \leq m} \mathbf{q}(Z(x_i) = z(x_i), Z(x_j) = z(x_j))|\phi_L)^\lambda \pi(\phi_L). \qquad (6.1)$$

Here $\lambda > 0$ is an arbitrary parameter that controls the dispersion of the Gibbs posterior. It can be shown that the Gibbs posterior is very close to the "true" posterior in the sense of Kullback-Leibler divergence.

Proposition 10. *Suppose that $Z = (Z(x_1), \ldots, Z(x_m))$ is drawn from $\mathbf{q}$, equivalently $F_{\phi_L^*}$, for some $\phi_L^* > 0$, and that for any sufficiently small neighborhood (u, v) of ϕ_L^*, $\pi(u, v) > |u - v|^r$ for some $r > 0$, then under the true marginal generating Z,*

$$\mathbb{E}_{P_\lambda} \frac{1}{m(m-1)/2} \log \left(\frac{P_\lambda(\phi_L^*|Z)}{P_\lambda(\phi|Z)} \right) = O_P(\frac{1}{m}).$$

Next, we introduce a variational Bayes approach for inference about $\mathbf{q}$. In particular, the proposed sampling method for the decay parameter ϕ_L via the Gibbs posterior provides a direct motivation for approximating the distribution $\mathbf{q}$ using variational inference techniques for Markov random fields (cf., e.g., Wainwright and Jordan (2003)). Let E be a subset of pairs $\{(i,j)|\ 1 \le i < j \le m\}$. E could be viewed as a collection of edges connecting the vertices $x_1, \ldots, x_m \in D$ to form a graphical structure. Our strategy is to approximate the multivariate distribution $\mathbf{q}(Z(x_1), \ldots, Z(x_m))$ by a graphical model distribution $\tilde{\mathbf{q}}$ defined as

$$\tilde{\mathbf{q}}_E(Z(x_1), \ldots, Z(x_m)) \propto \prod_{(i,j) \in E} \mathbf{q}(Z(x_i), Z(x_j)). \qquad (6.2)$$

Then the conditional probability distribution for the labels is approximated by $\tilde{\mathbf{q}}_E(Z(x_1)|Z(x_2), \ldots, Z(x_m)) \propto \prod_{j \ne 1} \mathbf{q}(Z(x_1), Z(x_j))$.

The following result shows that $\tilde{\mathbf{q}}$ is the best possible approximation within a restricted class of graphical models in the sense of Kullback-Leibler divergence $D(\cdot||\cdot)$.

Lemma 11. *Consider a class of probability distributions of $(Z(x_1), \ldots, Z(x_m)) \in \{1,2\}^m$:*

$$\mathcal{Q}_E = \left\{ Q : Q(Z(x_1), \ldots, Z(x_m)) \propto \prod_{(i,j) \in E} q_{ij}(Z(x_i), Z(x_j)) \right\},$$

where q_{ij}'s are any function on $\{1,2\}^2$. Then the distribution $\tilde{\mathbf{q}}_E$ defined in (6.2) satisfies

$$\tilde{\mathbf{q}}_E = \mathrm{argmin}_{Q \in \mathcal{Q}_E} D(\mathbf{q}||Q).$$

From the above lemma, the more edges added to set E, the better the approximation $\tilde{\mathbf{q}}$ is for $\mathbf{q}$, but it is also more difficult to estimate the log-partition function

$$A(E) = \log \sum_Z \prod_{(i,j) \in E} \mathbf{q}(Z(x_i), Z(x_j)).$$

Indeed, for a tree-structured graph, $A(E)$ is a known constant while, in general, we can only obtain upper and lower bounds.

Proposition 12.

(a) *The marginal distribution under $\tilde{\mathbf{q}}_E$ is uniform.*

(b) *If E forms a spanning tree, then $A(\theta_E) = -(m-2)\log 2$, and $\tilde{\mathbf{q}}_E(Z(x_i), Z(x_j)) = \mathbf{q}(Z(x_i), Z(x_j))$ for any $(i,j) \in E$.*

(c) *Suppose E forms a connected graph, and $E_0 \subseteq E$ is a spanning tree, then*

$$-(|E| - 1)\log 2 + U \leq A(\theta_E) \leq -(|E| - 1)\log 2 + V, \; where$$

$$U = \sum_{(i,j)\in E-E_0} \Bigg(\tilde{\mathbf{q}}_{E_0}(Z(x_i) \neq Z(x_j)) \log \mathbf{q}(Z(x_i) \neq Z(x_j))$$

$$+ \tilde{\mathbf{q}}_{E_0}(Z(x_i) = Z(x_j)) \log \mathbf{q}(Z(x_i) = Z(x_j)) \Bigg),$$

$$V = \sum_{(i,j)\in E-E_0} \Bigg(\tilde{\mathbf{q}}_E(Z(x_i) \neq Z(x_j)) \log \mathbf{q}(Z(x_i) \neq Z(x_j))$$

$$+ \tilde{\mathbf{q}}_E(Z(x_i) = Z(x_j)) \log \mathbf{q}(Z(x_i) = Z(x_j)) \Bigg).$$

For a one-dimensional domain D, we conveniently employ a tree-structured approximation for $\mathbf{q}$ in which the set of (x_t, x_{t+1}) pairs form the collection of edges for $t = 1, \ldots, m - 1$, assuming that $x_1 < \ldots < x_m$. For domains of two or higher dimensions, we also apply a minimum spanning tree approximation, although more sophisticated methods can be employed (see Wainwright and Jordan (2003)).

7. Applications

We demonstrate the behavior of the Dirichlet label process prior using simulated data in Section 7.1. Sections 7.2 and 7.3 look at a collection of progesterone curves and a collection of images, respectively.

7.1. Synthetic data

First we illustrate the fitting of the mixture model described in Section 2, where the species samples are obtained by random switching among k species curves that are drawn from a known Gaussian process on the real line. In particular, we specify $m = 20$ locations $[x_1, \ldots, x_m] = [1, 3, \ldots, 39]$ while leaving out 20 other locations $2, 4, \ldots, 40$ for validation purposes. θ_j^* for $j = 1, \ldots, k$ are independently drawn from a Gaussian process $GP(\mu_j, \phi_\theta, \sigma_\theta)$ at locations $x_1, \ldots, x_m$, where $\mu_j = -1 + 2(j-1)/(k-1)$. The label vectors $L_1, \ldots, L_n$ are drawn from label process $\mathbf{q}$, which is drawn by known ϕ_L. Species $\theta_1, \ldots, \theta_n$ are constructed by letting $\theta_i(x_t) = \theta_{L_i(t)}^*(x_t)$. Finally, the data collection $Y_1, \ldots, Y_n$ is obtained by mixing θ_i with an independent error process drawn from $N(0, \tau^2 I_m)$. We generated $n = 100$ sample curves using $k = 4$ canonical species curves. Parameter values for data generation were $\phi_\theta = 0.01$, $\sigma_\theta = 1$, $\phi_L = 0.05$, $\tau = 0.1$. For inference, we placed an uniform prior on the label switching parameter $\phi_L \sim Uni[0.0001, 1]$, while keeping $\phi_\theta, \sigma_\theta$ and τ fixed. Posterior distributions for latent labels and
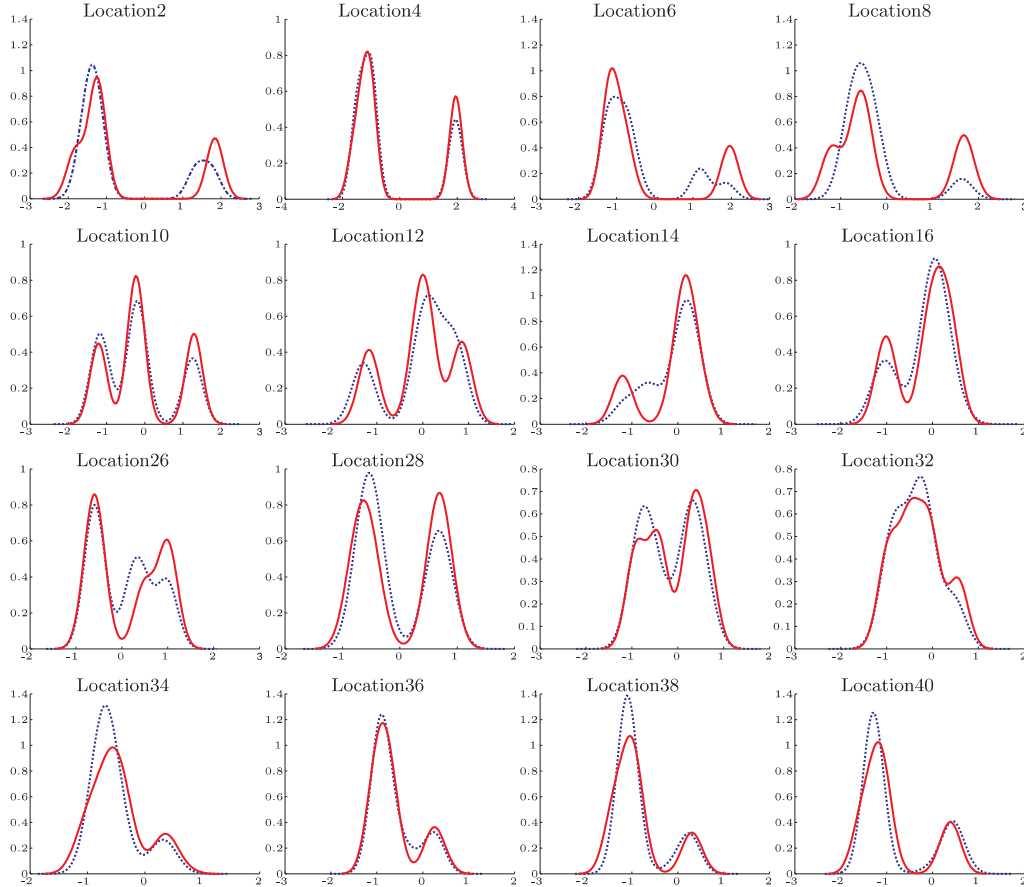
Figure 1. Evolution of posterior distributions at held-out locations $x = 2, 4, \ldots, 40$. Solid plots are true distributions. Dashed plots are predictive distributions learned from the model.

canonical species curves were obtained by running the MCMC algorithm for 4,000 iterations after a burn-in period of 1,000 iterations. An examination of running traces suggested that the sampling algorithm mixed well.

Figure 1 illustrates the evolution of the posterior distributions (in solid lines) at the held-out locations, as we move from location 2 to 40, the estimated densities obtained from our sampling. It is interesting to observe how the clusters initially "move" toward each other, then split into more clusters, and merge again. The estimated densities (in dashed lines) approximate the true densities well. The dependence in these distributions is driven by the smoothness of the k canonical species curves θ_j $(j = 1, \ldots, k)$ serving as the bases for our curve collection, as well as the label switching parameter ϕ_L.

With ϕ_θ fixed, ϕ_L plays a central role in the identifiablity of the canonical
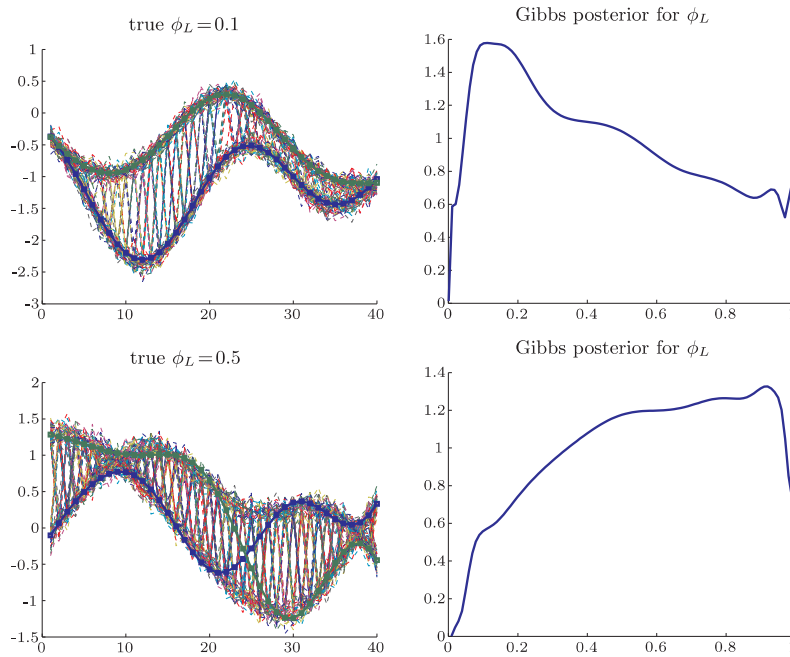
Figure 2. Illustration of canonical curve samples generated from the posterior in solid lines with squares. Figures to the right describe the corresponding Gibbs posterior for ϕ_L.

species curves θ_j. When ϕ_L is close to 0, the curves hardly switch their labels, the curve collections can be globally clustered by the canonical curves that are strongly identifiable. On the other hand, when ϕ_L is large, the curves tend to switch often among the canonical species curves which become more weakly identified. In general, our model is able to always recover segments of locations that admit relatively few switchings. Figure 2 illustrates this phenomenon with data generated from $k = 2$ canonical curves, with the true ϕ_L set to be 0.1 (top) and 0.5 (bottom figures). Note the corresponding Gibbs posterior for ϕ_L which was obtained from our sampling algorithm. In both cases, a uniform distribution prior $Uni[0.0001, 1]$ was placed on parameter ϕ_L, while $\phi_\theta = 0.005$, $\sigma_\theta = 1$, $\tau = 0.1$ were fixed. For smaller value of true ϕ_L (top figures), the posterior was well-concentrated around the true value. For larger ϕ_L (bottom figures), the posterior mass shifted to the right, because the canonical species curve estimates (due to weak identifiability) tended to over-switch between the modes.

7.2. Progesterone modeling

We turn to an application of the Dirichlet labeling process for modeling Progesterone data (cf., Brumback and Rice (1998)). This data set records the

natural logarithm of the progesterone metabolite, measured by urinary hormone assay, during a monthly cycle for 51 female subjects. Each cycle ranges from -8 to 15 (8 days pre-ovulation to 15 days post-ovulation). There are a total of 88 cycles; the first 66 cycles belong to non-contraceptive group, the remaining 22 cycles belong to the contraceptive group. This grouping is of course *unknown* to our analysis. See Figure 3 for the illustration. This data set is interesting as it allows us to compare our model to a more simplistic global clustering approach. To appreciate the noise and overlap of the two groups, we also consider a modified data set in which the curves belong to the contraceptive group are down-shifted by 2 (see Figure 6).

We focus our analysis to the case $k = 2$. We envision that there are two canonical curves providing bases for random label selection (switching). Due to the apparent noise and overlap of the two groups, we placed a prior on the switching parameter $\phi_L \sim Gam(5, 2)$ so as to allow possible duplication of canonical curves in certain local segments. Canonical curves were drawn from mean-0 Gaussian process with a covariance matrix using decay parameter $\phi_\theta = 0.005$ and $\sigma_\theta = 1$. We fixed the precision parameters $\tau = 1$, $\alpha = 1$. A discussion of the sensitivity of these parameters is included in the sequel. Samples from posterior distribution were collected from 5,000 MCMC iterations (discarding the first 1,000). An examination of running traces suggested very fast mixing. Figure 3 shows the mean estimate for the canonical curves. (The quantiles are not plotted because the posterior distribution for canonical curves are tightly concentrated around their means). It appears difficult to cluster the data for individual locations without taking into account the global smoothness of the whole curves. With our model the estimated canonical curves appear to match the general behavior of the two groups fairly well. We observe that the two canonical curves are virtually indistinguishable in the early part of the cycle. In fact, the behavioral patterns between the two curves become more distinguishable only in the post-ovulation period. Figure 4 shows the label mean for the whole monthly period for each of the 88 individual cycles. The last 22 cycles (contraceptive group) register generally higher label means than the first 66 cycles. This is also demonstrated by heatmaps in Figure 5, which illustrate the proportion of equal labels for pairs of curve replicates. Although global clustering is apparently not possible, one can observe the local clustering effect by zooming in to the curve segment corresponding to the last 5 days of the menstrual cycle. We also applied our analysis (using the same prior specification and parameter initial values) to a modified data set in which the curves belong to the second group were down-shifted by 2. Global clustering was now easily achievable (see Figure 6).

We now turn to a discussion of the effects of several parameters of interest on the identifiability of the canonical curves (see Figure 7 for illustrative results). We
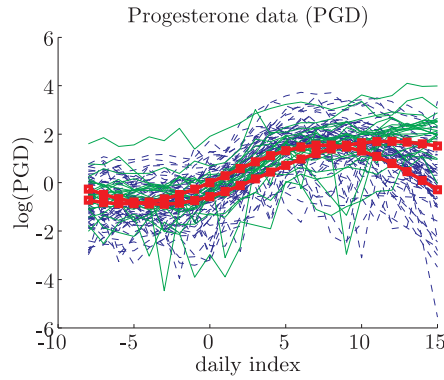
Figure 3. Monthly PGD cycle for contraceptive group (solid lines) and non-contraceptive group (dashed lines). Solid lines with squares are the mean estimate of canonical curves.
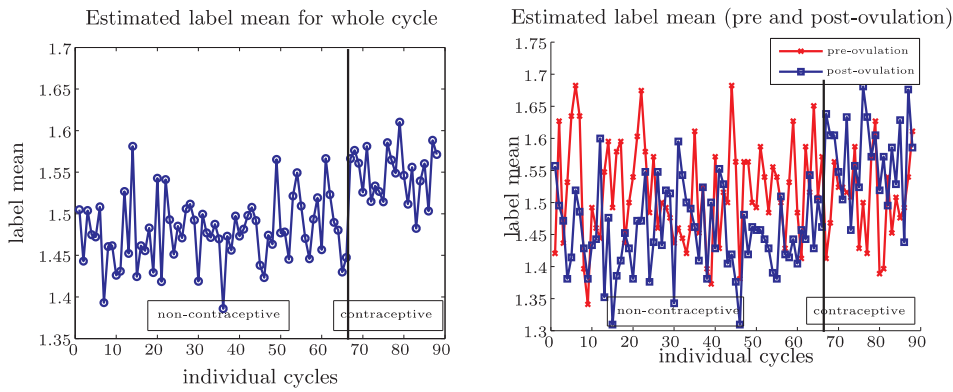

Figure 4. Left: Mean of estimated labels during the whole monthly cycle. Right: label means for pre and post-ovulation periods for 88 individuals (plots with x's and squares, resp.).

observe that, as ϕ_L gets smaller, the model insists on increasingly global clusters (and less label switching for each replicate) resulting in separable canonical curves that do not intersect. For this data set, these separable curve estimates do not reflect the behavioral pattern for each of the two groups, but act rather as a pair of basis curves for representing the curve collection. On the other hand, large ϕ_L offers more flexibility by allowing more complex canonical curve interaction. For instance, it is possible to obtain well separated clustering effects in one local segment and almost duplicates in another local segment. Turning to τ, as τ gets smaller, the canonical curves become more distinct (and less smooth) to expand the coverage of the function space. On the other hand, large τ results in weakly distinguishable canonical curves. The role of ϕ_θ is to dictate the smoothness of
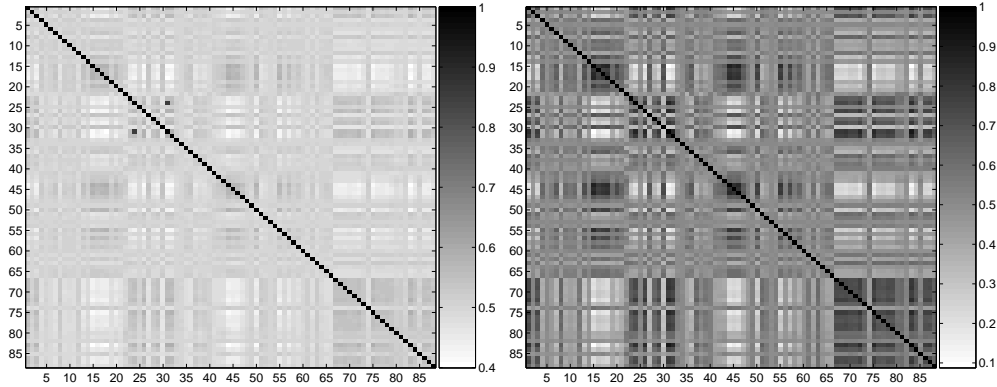
Figure 5. Heatmap illustrating propotion of equal labels for pairs of repli-
cates for the whole curve (left), and a curve segment $[20, 24]$ (i.e., last 5 days
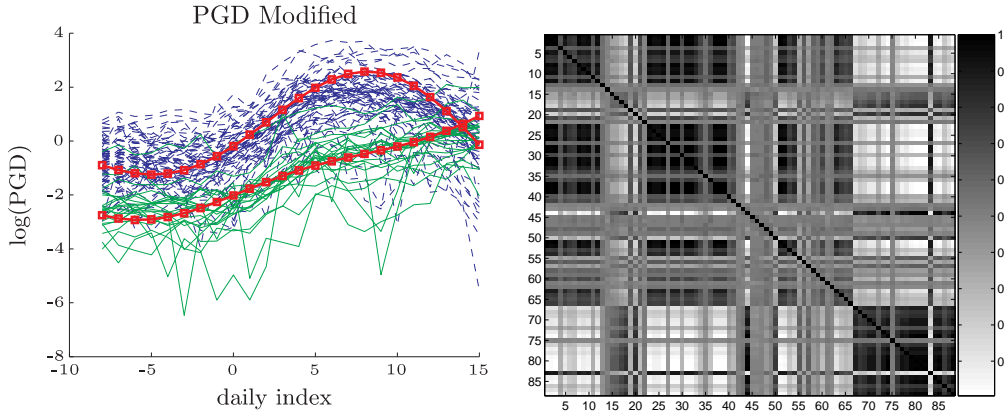of the monitored cycle) (right).

Figure 6. Analysis applied to the modified PG data set. Left: Mean estimate
for the canonical curves. Right: Heatmap illustrating proportion of equal
labels for pairs of replicates for the whole curve.

canonical curves. Finally, the influence of α (not shown here) on the number of
clusters induced by label realizations is less pronounced than that of τ and ϕ_L
for this data set.

7.3 Image modeling

In this section we demonstrate a possibly surprising application of our model
to an image segmentation task. Our data set consists of 80 color images from
a Microsoft image database Winn, Criminisi, and Minka (2005). These images
are of size 26×40. Although these images can be loosely grouped into different
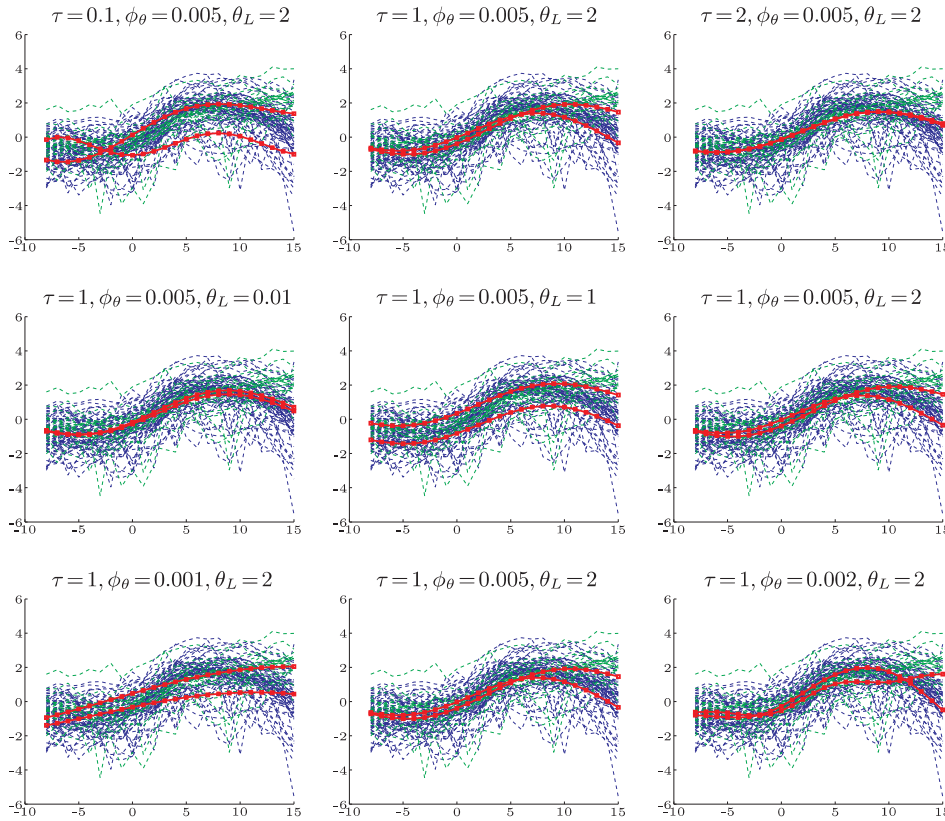categories (grass fields, plants, buildings, planes, etc), there are often multiple

Figure 7. Top row: effects of $\tau = 0.1, 1, 2$. Second row: effects of $\phi_L = 0.01, 1, 2$. Third row: effects of $\phi_\theta = 0.001, 0.005, 0.02$.

objects of different types in the same image. It is thus very natural to view them as hybrid species curves. Each image is represented by a surface realization Y_i, for $i = 1, \ldots, 80$, where $Y_i(x)$ is the color intensity of the location $x \in D = \{1, \ldots, 26\} \times \{1, \ldots, 40\}$ in the i-th image. The color intensity consists of three numbers in $[0, 255]$ (corresponding to the red, green and blue scales). Accordingly, we write $Y_i(x) = [Y_i^1(x), Y_i^2(x), Y_i^3(x)]$. We introduce $k = 8$ canonical species curves, each of which is represented by three constant random functions ranging in $[0, 255]$. (Introducing more canonical species, which we did for instance with $k = 12$, almost always yielded more than one duplicate canonical species curves). We write $\theta_j^*(\cdot) = [\theta_j^{*1}(\cdot), \theta_j^{*2}(\cdot), \theta_j^{*3}(\cdot)]$ for $j = 1, k$. The three dimensions are treated independently, by letting $Y_i^r(x) = \theta_{L_i(x)}^{*}{}^r(x) + \epsilon_{i,r,x}$, where $\epsilon_{i,r,x}$ is independent zero-mean normal noise with variance τ^2 for any $i = 1, \ldots, n$; $r = 1, 2, 3$; $x \in D$.

As described in Section 4 we introduce additional constraints into the prior structure for the canonical curves θ^*. In particular, we placed a (truncated)
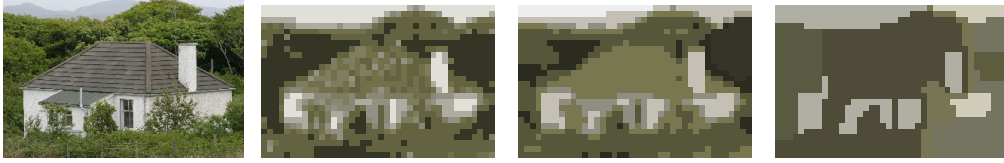
Figure 8. Effects of $\phi_L = 0.5, 0.05, 0.01$ on segmentation for the leftmost image.

Figure 9. Examples of segmented images (using $\phi_L = 0.1$).

normal prior with mean [10 10 10] and variance $10^2 I_3$ on θ_1^*, and a normal prior with mean [240 240 240] and the same variance on θ_k^*. That is, we anchored the two extreme labels with the two extremes of the color scale (black and white colors). All remaining canonical species were given a relatively non-informative prior; for $j = 2, \ldots, k-1$, $\theta_j^{*r} \overset{i.i.d.}{\sim} N(\mu_j, \sigma_j) \cdot \mathbb{I}_{[0,255]}$, where $\mu_j = 128$. For all

$j = 1, \ldots, k$, we fixed $\sigma_j = \sigma_\theta$, where $\sigma_\theta^{-2} \sim Gam(a_\sigma, b_\sigma)$. We set $a_\sigma = 0.4$ and $b_\sigma = 0.001$. For precision parameter τ we let $\tau^{-2} \sim Gam(a_\tau, b_\tau)$, where $a_\tau = 0.1$ and $b_\tau = 0.025$. We set the concentration parameter $\alpha = 1$.

To complete the prior specification, let us turn to the latent labeling processes $\mathbf{p}$ and $\mathbf{q}$. One possible approach is to endow $\mathbf{p}$ with a single Dirichlet labeling process prior for the entire domain (as in the previous applications). For the image data set, global clustering is generally not of interest (because it is unlikely that two images have exactly the same labeling everywhere). On the other hand, label sharing at smaller scales (not to mention the pixel-level scale) is much more likely due to the occurrence of similar objects in similar scenes. To encourage this sharing we decomposed each image into fixed and disjoint patches of size $r \times r$. Conditionally on $\mathbf{q}$, the labeling processes $\mathbf{p}$ defined for disjoint patches are mutually independent and follow the Dirichlet labeling process specification as before. We experimented with the choices $r = 4, 6, 8$ and received comparable results. Finally, the latent labeling process $\mathbf{q}(\phi_L, k = 8)$ was specified for the whole domain using different choices of $\phi_L = 0.5, 0.1, 0.05, 0.01$.

The MCMC algorithm was run for 200 iterations. Samples obtained from the last 150 iterations were used for image segmentation. The segmented images were obtained by assigning to each image location the light intensity of the MAP estimate of the canonical curve at the same location. Figure 9 provides examples of representative segmentation results. Figure 8 illustrates the effects of ϕ_L on the segmentation results. For ϕ_L large, the group allocation at each location is highly independent, resulting in fragmented segmentation. As ϕ_L decreases, the segments become increasingly coherent. As ϕ_L becomes too small, however, nearby locations are forced to share the same group. Furthermore, patches from different images are also encouraged to cluster, resulting in increasingly "abstract" segments.

8. Conclusions

The Dirichlet labeling process provides a highly flexible prior for modeling collections of functions (curves, surfaces). Though driven by just a few parameters, the inter-relationships between these parameters are complex with regard to process behavior. We are currently exploring multivariate extensions of the labeling process, the modeling of label clustering at random spatial scales, as well as the incorporation of prior knowledge of canonical curves. From a practical viewpoint, it is also worth comparing the functional clustering arising from the Dirichlet labeling process prior with that of non model-based approaches such as those mentioned in the Introduction. Investigation to illuminate advantages/diadvantages offers potential future work.

Acknowledgement

Appendix

Proof of Proposition 1. This result is straightforward using standard properties of Dirichlet distribution.

Proof of Proposition 2. We derive the result for a stochastic process $F \sim GP(0, \sigma_L, \phi_L)$ (the Proposition states the result for $\sigma_L = 1$). From the definition, $\mathbf{q}_{x_1, x_2}(1,1) = P(\eta(x_1) > 0, \eta(x_2) > 0)$. Note that $(\eta(x_1), \eta(x_2)) \sim N([0 \ \ 0], \begin{bmatrix} \sigma_L^2 & \rho_{12} \\ \rho_{12} & \sigma_L^2 \end{bmatrix})$. By a change of variables, $\tilde{\eta} = (1/\sqrt{2}) \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} [\eta(x_1) \ \eta(x_2)]^T$, we obtain that $\tilde{\eta} \sim N([0 \ \ 0], \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix})$, where $\lambda_1 = \sigma_L^2 + \rho_{12}$ and $\lambda_2 = \sigma_L^2 - \rho_{12}$. Then

$$P(\eta(x_1) > 0, \eta(x_2) > 0) = \frac{1}{2\pi\sqrt{\lambda_1 \lambda_2}} \int_{\eta(x_1), \eta(x_2) > 0} \exp - \frac{(\tilde{\eta}_1^2/\lambda_1 + \tilde{\eta}_2^2/\lambda_2)}{2} d\tilde{\eta}_1 \tilde{\eta}_2.$$

Another change of variables ($\tilde{\eta}_1 = r\sqrt{\lambda_1}\cos\alpha$ and $\tilde{\eta}_2 = r\sqrt{\lambda_2}\sin\alpha$) and some elementary calculus yields the desired result.

Proof of Proposition 3. We derive the result for a stochastic process $F \sim GP(0, \sigma_L, \phi_L)$ (the Proposition states the result for $\sigma_L = 1$). The proposition is concerned with an arbitrary collection of indices i, j such that $i, j \in (\alpha_1 k, \alpha_2 k)$ for given $0 < \alpha_1 < \alpha_2 < 1$. By definition, $c_j = \sigma_L \Phi^{-1}(j/k)$. By Taylor approximation, it is simple to obtain that

$$c_j = c_{j-1} + \frac{1}{k}(\sqrt{2\pi}\sigma_L e^{c_j^2/2\sigma_L^2} + o(1)). \tag{A.1}$$

Furthermore, $c_j = -c_{k-j}$ and $c_{k/2} = 0$ if k is even. Using a Riemann sum

approximation, we have:

$$
\begin{aligned}
&P(Z_1 = i, Z_2 = j) \\
&\quad = P(\eta_1 \in (c_{i-1}, c_i), \eta_2 \in (c_{j-1}, c_j)) \\
&\quad = \frac{1}{2\pi\sqrt{\sigma_L^4 - \rho_{12}^2}} \int_{c_{i-1}}^{c_i} \int_{c_{j-1}}^{c_j} \exp-\frac{\sigma_L^2(\eta_1^2 + \eta_2^2) - 2\rho_{12}\eta_1\eta_2}{2(\sigma_L^4 - \rho_{12}^2)} d\eta_1\eta_2 \\
&\quad = \frac{1}{2\pi\sqrt{\sigma_L^4 - \rho_{12}^2}}(c_i - c_{i-1})(c_j - c_{j-1}) \exp-\frac{\sigma_L^2(c_i^2 + c_j^2) - 2\rho_{12}c_ic_j}{2(\sigma_L^4 - \rho_{12}^2)}(1 + o(1)) \\
&\quad \overset{\text{(A.1)}}{=} \frac{1}{k^2}\frac{\sigma_L^2}{\sqrt{\sigma_L^4 - \rho_{12}^2}} \exp-\frac{(c_i^2 + c_j^2)\rho_{12}^2 - 2\rho_{12}\sigma_L^2c_ic_j}{2(\sigma_L^4 - \rho_{12}^2)\sigma_L^2}(1 + o(1)).
\end{aligned}
$$

By properties of the multivariate Gaussian, $\eta_2|\eta_1 = u_1 \sim N(u_1\rho_{12}/\sigma_L^2, \sigma_L^2 - \rho_{12}^2/\sigma_L^2)$. So

$$
\begin{aligned}
P(\eta_1 = i, \eta_2 \geq j) &= \int_{c_{i-1}}^{c_i} \frac{e^{-u_1^2/2\sigma_L^2}}{\sigma_L\sqrt{2\pi}} \left(1 - \Phi(\frac{c_j - u_1\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2})\right) du_1 \\
&= (c_i - c_{i-1})(\frac{e^{-c_i^2/2\sigma_L^2}}{\sigma_L\sqrt{2\pi}} \left(1 - \Phi(\frac{c_j - c_i\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2})\right)) \\
&\overset{\text{(A.1)}}{=} \frac{1}{k}\left(1 - \Phi\left(\frac{c_j - c_i\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2}\right) + o(1)\right).
\end{aligned}
$$

The above result can be used to obtain conditional probabilities (e.g., for interpolation). For instance, for $i, j_1, j_2 \in \{1, \ldots, k\}$,

$$
P(Z(x_2) \in (j_1, j_2]|Z(x_1) = i) = \Phi\left(\frac{c_{j_2} - c_i\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2}\right) - \Phi\left(\frac{c_{j_1} - c_i\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2}\right) + o(1).
$$

Finally, it is worth mentioning that as $k \to \infty$, $o(1) \to 0$ uniformly for all i, j in the specified interval, where $o(1)$ does depend on σ_L and ϕ_L.

Extension of Proposition 3 to $m > 2$ locations. The results of Proposition 2 can be easily extended to an arbitrary collection of locations $x_1, \ldots, x_m$. Let A denote the inverse covariance matrix for random vector $(\eta(x_1), \ldots, \eta(x_m))$. For any m-tuple $(j_1, \ldots, j_m) \in \{1, \ldots, k\}^m$ such that none of c_{j_i} diverges to ∞ or $-\infty$,

$$
P(Z(x_1) = j_1, \ldots, Z(x_m) = j_m) = \frac{1}{k^m}(R_{j_1,\ldots,j_m}(c_{j_1}, \ldots, c_{j_m}) + o(1)), \quad \text{(A.2)}
$$

where $o(1) \to 0$ *uniformly* for all such tuple $(j_1, \ldots, j_m)$, and

$$R_{j_1,\ldots,j_m}(c_{j_1}, \ldots, c_{j_m}) = \sigma_L^m (\det A)^{1/2} \exp \sum_{i=1}^{m} c_{j_i}^2 \left(\frac{1}{2\sigma_L^2} - \frac{A_{ii}}{2}\right) - \sum_{s<t} c_{j_s} c_{j_t} A_{st}.$$
(A.3)

Given the d.f. for $\mathbf{q}$, it is simple to obtain conditional probabilities, e.g., for label $Z(x_1)$ at location x_1 given remaining labels $Z(x_2), \ldots, Z(x_m)$. Letting $\tilde{A}$ denote the inverse of the covariance matrix for $Z(x_2), \ldots, Z(x_m)$,

$$P(Z(x_1) = j_1 | Z(x_2) = j_2, \ldots, Z(x_m) = j_m)$$
$$= \frac{\sigma_L (\det A)^{1/2}}{k (\det \tilde{A})^{1/2}} \exp\left\{ c_{j1}^2 \left(\frac{1}{2\sigma_L^2} - \frac{A_{11}}{2}\right) - c_{j_1} \sum_{t \neq 1} c_{j_t} A_{1t} \right.$$
$$\left. - \frac{1}{2} [c_{j2} \ldots c_{jm}]^T (A - \tilde{A}) [c_{j2} \ldots c_{jm}] \right\} + o\left(\frac{1}{k}\right).$$
$$P(Z(x_1) > j_1 | Z(x_2) = j_2, \ldots, Z(x_m) = j_m)$$
$$= 1 - \Phi\left(\frac{c_{j1} - [\rho_{12} \ldots \rho_{1m}] \tilde{A} [c_{j2} \ldots c_{jm}]^T}{\sigma_L^2 - [\rho_{12} \ldots \rho_{1m}]^T \tilde{A} [\rho_{12} \ldots \rho_{1m}]} \right) + o(1).$$

Proof of Proposition 4. Since θ^* and L are independent, we have

$$\mathbb{E}(\theta(x_1) - \theta(x_2))^2 = \mathbb{E} \sum_{j_1, j_2 = 1, \ldots, k} \mathbf{p}_{x_1, x_2}(j_1, j_2)(\theta_{j_1}^*(x_1) - \theta_{j_2}^*(x_2))^2$$
$$= \sum_{j_1 \neq j_2} \mathbf{q}_{x_1, x_2}(j_1, j_2)(\mathbb{E}\theta^*(x_1)^2 + \mathbb{E}\theta^*(x_2)^2 - \mathbb{E}\theta^*(x_1)\mathbb{E}\theta^*(x_2))$$
$$+ \sum_{j=1}^{k} \mathbf{q}_{x_1, x_2}(j, j)\mathbb{E}(\theta^*(x_1) - \theta^*(x_2))^2.$$

The second summand goes to 0 because θ^* is mean square continuous. It remains to show that for any j_1, j_2 such that $j_1 \neq j_2$, $\mathbf{q}_{x_1, x_2}(j_1, j_2) \to 0$ as $x_2 \to x_1$. Note that if F is a Gaussian process $GP(0, \sigma_L, \phi_L)$ and $k = 2$, this probability is available in a closed form given by Proposition 2: For $j_1 \neq j_2$, $\mathbf{q}_{x_1, x_2}(j_1, j_2) = (1/\pi) \arccos(1/2 + \rho_{12}/2\sigma_L^2)^{1/2} \to 0$ as $x_2 - x_1 \to 0$. More generally, suppose that $j_1 < j_2$. Recall the construction of Z via the auxiliary random function $\eta \sim F$. Fix an arbitrary $\epsilon > 0$. $P(Z(x_1) = j_1, Z(x_2) = j_2 | \eta(x_1) < c_{j_1} - \epsilon) \leq P(\eta(x_2) - \eta(x_1) > \epsilon) \leq \mathbb{E}(\eta(x_1) - \eta(x_2))^2/\epsilon^2 \to 0$ as $x_2 \to x_1$, due to the mean square continuity of η. Now letting $\epsilon \to 0$, we obtain $P(\eta(x_1) \in [c_{j_1} - \epsilon, c_{j_1}]) \to 0$. So, as $x_2 \to x_1$

$$\mathbf{q}_{x_1, x_2}(j_1, j_2) \leq P(\eta(x_1) \in [c_{j_1} - \epsilon, c_{j_1}]) + P(Z(x_1) = j_1, Z(x_2) = j_2 | \eta(x_1) < c_{j_1} - \epsilon)$$
$$\to 0.$$

Proof of Proposition 5. To simplify the notation in the proof we omit the subscripts $x_1, \ldots, x_m$. The equality in distribution implies the equality of moment generating functions. Thus, for any $(t_1, \ldots, t_m) \in \mathbb{R}^m$, we have

$$\sum_{(j_1,\ldots,j_m)} \mathbf{p}(j_1, \ldots, j_m) \exp\left(\sum_{r=1}^{m} \theta_{j_r}^*(x_r) t_r + \frac{1}{2} \tau^{*2} \sum t_r^2 \right)$$

$$= \int \exp\left(\sum_{r=1}^{m} \theta(x_r) t_r + \frac{1}{2} \tau^2 \sum t_r^2 \right) G_1(d\theta) \pi_1(d\tau). \tag{A.4}$$

Let $j_1^*, \ldots, j_m^*$ be the indices of the canonical curves that have the maximum values at locations $x_1, \ldots, x_m$, respectively. Divide both sides of the above equation by $\exp[\sum_r \theta_{j_r^*}^* t_r + (1/2)\tau^{*2} \sum t_r^2]$ to obtain

$$\mathbf{p}(j_1^*, \ldots, j_m^*) + \sum_{(j_1,\ldots,j_m)} \mathbf{p}(j_1, \ldots, j_m) \exp\left\{ \sum_{r=1}^{m} (\theta_{j_r}^*(x_r) - \theta_{j_r^*}^*) t_r \right\}$$

$$= \int \exp\left\{ \sum_{r=1}^{m} (\theta(x_r) - \theta_{j_r^*}^*) t_r + \frac{1}{2}(\tau^2 - \tau^{*2}) \sum t_r^2 \right\} G_1(d\theta) \pi_1(d\tau).$$

Note that for any $r = 1, \ldots, m$, letting $t_r \to +\infty$ leaves the left hand side positive and bounded while the right hand side tends to 0 or ∞, unless under the prior $G_1 \times \pi_1$ the only events with strictly positive probabilities are either $\mathcal{Y}_1 = \{\tau = \tau^*, \theta(x_r) = \theta_{j_r^*}^*(x_r)$ for all $r = 1, \ldots, m\}$ or $\mathcal{Y}_2 = \{\tau < \tau^*\}$. It follows that the RHS can be written as

$$P_{G_1 \times \pi_1}(\mathcal{Y}_1)$$
$$+ \int_{\mathcal{Y}_2} \exp\left\{ \sum_{r=1}^{m} -\theta_{j_r^*}^* t_r - \frac{1}{2}(\tau^{*2} - \tau^2)(t_r - \frac{\theta(x_r)}{\tau^{*2} - \tau^2})^2 \right\} R(\theta, \tau) G_1(d\theta) \pi_1(d\tau),$$

where $R(\theta, \tau) = \exp[1/2(\tau^{*2} - \tau^2)] \sum_{r=1}^{m} \theta(x_r)^2$. Let $t_r \to \infty$ for $r = 1, \ldots, m$, and use the assumption on function R, to obtain that $P_{G_1 \times \pi_1}(\mathcal{Y}_1) = \mathbf{p}_{j_1^*, \ldots, j_m^*} \times \delta_{\tau^*}$. Now substract the tuple $(j_1^*, \ldots, j_m^*)$ from the moment equality equation and carry the same argument to the remaining terms to obtain $G_1 = G_{x_1, \ldots, x_m}$, and $\pi_1 = \delta_{\tau^*}$.

Proof of Lemma 6. This lemma is a multivariate version of a result of Ishwaran and Zarepour (Lemma 2). The key to the proof is to exploit the distribution of the weight vector $(\mathbf{p}_{x_1, \ldots, x_m}(\cdot))$, which is a k^m-dimensional Dirichlet distribution parameterized by $\mathbf{q}_{x_1, \ldots, x_m}(j_1, \ldots, j_m)$ where $(j_1, \ldots, j_m) \in \{1, \ldots, k\}^m$. It can be shown that, by examining the proof of Proposition 3, there exists a constant $c > 0$ such that for any k there are at least $(k/2)^m$ tuples of the form $(j_1, \ldots, j_m)$

whose associated $\mathbf{q}$ probabilities are greater than c/k^m. Using this fact, and choosing $k > 2k^*$, where k^* is the true number of canonical curves used to generate f_0, it can be verified that Ishwaran and Zarepour's proof goes through here as well.

Proof of Lemma 7. First, let $\mathcal{F}^{a_n}_{n,k,\tau}$ be a subset of $\mathcal{F}_{n,k}$ where the associated τ is fixed to a constant in $[\tau_n, M]$, and the support of the associated $G_{x_1,\dots,x_m}$ lies entirely in $[-a_n, a_n]^m$. We obtain the bound of this density class first, and then relate it to the entropy of the bigger class $\mathcal{F}_{n,k}$.

We note the following bound of L_1 distance for two univariate normal densities (from Lemma 1 of Ghosal, Ghosh, and Ramamoorthi (1999)):

$$\|\psi(\theta(x_1), \tau) - \psi(\theta(x_2), \tau)\|_1 \le \frac{(\theta(x_1) - \theta(x_2))}{\tau}.$$

To extend this bound to the product of normal densities, we exploit the following bound between L_1 and the Hellinger distance h (which is defined as $2h^2(f, g) = \int (\sqrt{f} - \sqrt{g})^2$):

$$\frac{1}{2}\|f - g\|_1^2 \le 2h^2(f, g) \le \|f - g\|_1.$$

For product of densities, we have

$$h^2\left(\prod_{i=1}^m f_i, \prod_{i=1}^m g_i\right) = 1 - \int \left(\prod_{i=1}^m f_i g_i\right)^{1/2} = 1 - \prod_{i=1}^m \int (f_i g_i)^{1/2} = 1 - \prod_{i=1}^m (1 - h^2(f_i, g_i)).$$

As a result, if $|\theta(x_i) - \tilde{\theta}(x_i)| \le \delta$ for all $i = 1, \dots, m$,

$$\left\|\prod_{i=1}^m \psi(\theta(x_i), \tau) - \prod_{i=1}^m \psi(\tilde{\theta}(x_i), \tau)\right\|_1^2 \le 4\left(1 - \prod_{i=1}^m \left(1 - \frac{(\theta(x_i) - \tilde{\theta}(x_i))}{2\tau}\right)\right) \le \frac{2m\delta}{\tau}.$$

If $\|\theta - \tilde{\theta}\|_\infty \le \delta^2 \tau_n / 2m$ and $\tau \ge \tau_n$ then $\|\prod_{i=1}^m \psi(\theta(x_i), \tau) - \prod_{i=1}^m \psi(\tilde{\theta}(x_i), \tau)\|_1 \le \delta$.

Next, note that the hybrid species curves (vectors) θ is parameterized in terms of only k canonical vectors θ^*. That is, instead of needing k^m mean variables of the k^m mixture components, we need only km mean variables. This can reduce the size of the covering for $\mathcal{F}^{a_n}_{n,k,\tau}$, which we now specify: Let N_0 be the smallest integer greater than $2a_n/(\delta^2 \tau_n/2m) = 4a_n m/(\tau_n \delta^2)$. Divide $[-a_n, a_n]$ into N_0 equal intervals and collect N_0 midpoints. At each location x_i, $i = 1, \dots, m$, there are only k possible values for $\theta(x_i)$. Combining across m locations, there are $\binom{N_0}{k}^m \le N_0^{km}$ ways of choosing k^m hybrid curves (vectors) for the k^m mixture components using the midpoints constructed. Thus, we have obtained a δ-covering for the set of density products of the form

$\mathcal{P}_Y = \{\prod_{i=1}^m \psi(\mathbf{Y}(x_i)|\theta(x_i), \tau)|\theta \in [-a_n, a_n]^m\}$ that has no more than N_0^{km} elements, and $J(\delta, \mathcal{P}_Y) \le km \log N_0$.

Let $N = k^m$ and $\mathcal{P}_N = \{(P_1, \ldots, P_N) : P_i \ge 0, \sum_{i=1}^N P_i = 1\}$. As proved in Lemma 1 of Ghosal, Ghosh, and Ramamoorthi (1999) (GGR), under the l_1 metric, that $J(\delta, \mathcal{P}_N) \le N(1 + \log(1+\delta)/\delta)$. Furthermore, it is simple to observe that one can construct a 2δ-covering for $\mathcal{F}_{n,k,\tau}^{a_n}$ by combining each element of the δ-covering for $\mathcal{P}_N$ with an element of the δ-covering for $\mathcal{P}_Y$. This implies that:

$$J(2\delta, \mathcal{F}_{n,k,\tau}^{a_n}) \le J(\delta, \mathcal{P}_N) + J(\delta, \mathcal{P}_Y). \tag{A.5}$$

The final step is to relate $J(2\delta, \mathcal{F}_{n,k,\tau}^{a_n})$ to $J(2\delta, \mathcal{F}_{n,k})$. Following GGR we use another intermediate class $\mathcal{F}_{n,k,\tau}^{a_n,\delta}$ that consists of all $f \in \mathcal{F}_{n,k}$ where the support of the associated $G_{x_1,\ldots,x_m}$ lies within $[-a_n, a_n]^m$ with probability at least $1 - \delta$, and the variance τ is fixed. From GGR's Lemma 2, $J(3\delta, \mathcal{F}_{n,k,\tau}^{a_n,\delta}) \le J(\delta, \mathcal{F}_{n,k,\tau}^{a_n})$. Following GGR's Lemma 3, for $a_n > M/\sqrt{\delta}$, one can obtain that $\mathcal{F}_{n,k} \subset \mathcal{F}_{n,k,\tau_n}^{2a_n,2m\delta}$. Combining these with (A.5)

$$J(6m\delta, \mathcal{F}_{n,k}) \le J(6m\delta, \mathcal{F}_{n,k,\tau_n}^{2a_n,2m\delta}) \le J(2m\delta, \mathcal{F}_{n,k,\tau_n}^{2a_n})$$

$$\le k^m \log \frac{1 + m\delta}{m\delta} + km \log(1 + \frac{2a_n}{\tau_n m\delta^2}).$$

Proof of Proposition 9.

Let $\mathcal{D} = \{\|x_i - x_j\| | 1 \le i, j \le m\}$ and $p_\phi(d) = (1/\pi) \arccos \sqrt{1/2 + e^{-\phi d}/2}$. Let $n^+(d)$ ($n^-(d)$, resp.) be the number of (x_i, x_j) pairs such that $\|x_i - x_j\| = d$ and $\eta(x_1)\eta(x_2) \ge 0$ ($\eta(x_1)\eta(x_2) < 0$, resp.) Let $n(d) = n^+(d) + n^-(d)$. Note that $n(d)$ is independent of η. The maximum pseudo-likelihood estimator can be written as

$$\hat{\phi}_L = \text{argmax}_{\phi \ge 0} \sum_{d \in \mathcal{D}} n^+(d) \log p_\phi(d) + n^-(d) \log(\frac{1}{2} - p_\phi(d)).$$

For any $\phi, d \ge 0$, $0 \le p_\phi(d) \le 1/4$. From the definition of $\hat{\phi}_L$,

$$\sum_{d \in \mathcal{D}} n^+(d) \log p_{\hat{\phi}_L}(d) + n^-(d) \log(\frac{1}{2} - p_{\hat{\phi}_L}(d))$$

$$\ge \sum_{d \in \mathcal{D}} n^+(d) \log p_{\phi_L^*}(d) + n^-(d) \log(\frac{1}{2} - p_{\phi_L^*}(d)).$$

Due to the concavity of logarithm, $\log(u + v)/2 \ge (\log u + \log v)/2$ by Jensen's inequality. This implies

$$\sum_{d \in \mathcal{D}} n^+(d) \log \frac{p_{\hat{\phi}_L}(d) + p_{\phi_L^*}(d)}{2p_{\phi_L^*}(d)} + n^-(d) \log \frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)} \ge 0. \tag{A.6}$$

It is simple to see that both $\log[p_\phi(d) + p_{\phi_L^*}(d)]/[2p_{\phi_L^*}(d)]$ and $\log[1 - p_\phi(d) - p_{\phi_L^*}(d)]/[1 - 2p_{\phi_L^*}(d)]$ are absolutely bounded by some constant $M > 0$ for any $\phi \geq 0$. From Proposition 2, $\mathbb{E}n^+(d) = 2n(d)p_{\phi_L^*}(d)$ and $\mathbb{E}n^-(d) = 2n(d)(1/2 - p_{\phi_L^*}(d))$. Applying McDiarmid's inequality, for any $\epsilon > 0$ we obtain

$$
P\Bigg(\sup_{\phi \geq 0}\bigg| \sum_{d \in \mathcal{D}} \Big(n^+(d) - 2n(d)p_{\phi_L^*}(d)\Big) \log \frac{p_\phi(d) + p_{\phi_L^*}(d)}{2p_{\phi_L^*}(d)}
$$
$$
+ \Big(n^-(d) - 2n(d)(\frac{1}{2} - p_{\phi_L^*}(d)\Big) \log \frac{1 - p_\phi(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)}\bigg| \geq \epsilon\Bigg)
$$
$$
\leq 2\exp\frac{-4\epsilon^2}{m(m-1)M^2}. \tag{A.7}
$$

Combining (A.7) and (A.6),

$$
P\Bigg(\bigg| \sum_{d \in \mathcal{D}} -2n(d)p_{\phi_L^*}(d) \log \frac{p_{\hat{\phi}_L}(d) + p_{\phi_L^*}(d)}{2p_{\phi_L^*}(d)}
$$
$$
- 2n(d)(\frac{1}{2} - p_{\phi_L^*}(d) \log \frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)}\bigg| \geq \epsilon\Bigg) \leq 2\exp\frac{-4\epsilon^2}{m(m-1)M^2}.
$$

Thus,

$$
\sum_{d \in \mathcal{D}} n(d)\Bigg(p_{\phi_L^*}(d) \log \frac{2p_{\phi_L^*}(d)}{p_{\hat{\phi}_L}(d) + p_{\phi_L^*}(d)} + (\frac{1}{2} - p_{\phi_L^*}(d)) \log \frac{1 - 2p_{\phi_L^*}(d)}{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}\Bigg)
$$
$$
= O_P(m)
$$

in **q**-probability. Note that for $\log x \leq 2(\sqrt{x} - 1)$,

$$
p_{\phi_L^*}(d) \log \frac{2p_{\phi_L^*}(d)}{p_{\hat{\phi}_L}(d) + p_{\phi_L^*}(d)} + (\frac{1}{2} - p_{\phi_L^*}(d)) \log \frac{1 - 2p_{\phi_L^*}(d)}{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}
$$
$$
\geq -2p_{\phi_L^*}(d)\Bigg(\sqrt{\frac{p_{\phi_L^*}(d) + p_{\hat{\phi}_L}(d)}{2p_{\hat{\phi}_L}(d)}} - 1\Bigg)
$$
$$
-2(\frac{1}{2} - p_{\phi_L^*}(d))\Bigg(\sqrt{\frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)}} - 1\Bigg)
$$
$$
= \Bigg(\sqrt{\frac{p_{\phi_L^*}(d) + p_{\hat{\phi}_L}(d)}{2}} - \sqrt{p_{\phi_L^*}(d)}\Bigg)^2 + \Bigg(\sqrt{\frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{2}} - \sqrt{\frac{1}{2} - p_{\phi_L^*}(d)}\Bigg)^2
$$
$$
\geq \frac{1}{8}(p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d))^2.
$$

For any $\phi \in [0, \phi_0]$ and $d \leq d_0$, it is simple to verify that there exists a constant $C_0 > 0$ that depends only on ϕ_0 and d_0 such that $|p_\phi(d) - p_{\phi_L^*}(d)| \geq C_0|\phi - \phi_L^*|$.

As a result, $|\hat{\phi}_L - \phi_L^*| = O_P(\sqrt{m/r_m})$, where r_m is the number of pairs (x_i, x_j) such that $\|x_i - x_j\| \leq d_0$.

Proof of Proposition 10. Let $Z = (z(x_1), \ldots, z(x_m))$. Denote by P^* the joint distribution $P_Z \times P_\lambda$, where P_λ denotes the Gibbs posterior given Z, and P_Z the "true" distribution generating Z (i.e., under true ϕ_L^*). By Markov's inequality, for any $\epsilon_m > 0$,

$$
\begin{aligned}
P_\lambda(&\log P_1(\phi_L^*|) - \log P_1(\phi|) \geq \epsilon_m) \\
&= P_\lambda(\exp(\lambda(\log P_1(\phi_L^*|) - \log P_1(\phi|)) \geq \exp(\lambda\epsilon_m))) \\
&\leq \exp(-\lambda\epsilon_m)\mathbb{E}_{P_\lambda}\left(\frac{P_1(\phi_L^*|Z)}{P_1(\phi|Z)}\right)^\lambda \\
&= \exp(-\lambda\epsilon_m)\int\left(\frac{P_1(\phi_L^*|Z)}{P_1(\phi|Z)}\right)^\lambda \frac{\prod \mathbf{q}(z(x_i), z(x_j))^\lambda \pi(\phi)d\phi}{\int \prod \mathbf{q}(z(x_i), z(x_j))^\lambda \pi(\phi)d\phi} \\
&= \exp(-\lambda\epsilon_m)\frac{\prod \mathbf{q}(z(x_i), z(x_j)|\phi_L^*)^\lambda \int (\pi(\phi_L^*)/\pi(\phi))^\lambda d\phi}{\int \prod \mathbf{q}(z(x_i), z(x_j))^\lambda \pi(\phi)d\phi} \\
&= \exp(-\lambda\epsilon_m)\frac{\prod \mathbf{q}(z(x_i), z(x_j)|\phi_L^*)^\lambda C_1}{\int \prod \mathbf{q}(z(x_i), z(x_j))^\lambda \pi(\phi)d\phi},
\end{aligned}
$$

where $C_1 = \int (\pi(\phi_L^*)/\pi(\phi))^\lambda d\phi$ is a constant. Let

$$
A_m(\epsilon) = \{Z : \sup_{\phi \geq \phi_1}\left|\log \prod \mathbf{q}(z(x_i), z(x_j)) - \mathbb{E}_Z \log \prod \mathbf{q}(z(x_i), z(x_j))\right| \geq \epsilon\}.
$$

By McDiarmid's inequality, $P_Z(A_m(\epsilon)) \leq 2\exp(-4\epsilon^2)/[m(m-1)M^2]$ for some constant $M > 0$. Applying union bounds, under the joint distribution P^* we have, for any $\delta_m > 0$,

$$
\begin{aligned}
P^*(&\log P_1(\phi_L^*|) - \log P_1(\phi|) \geq \epsilon_m) \\
&\leq P_Z(A_m(\delta_m/4)) + \mathbb{E}_Z\left[\exp(-\lambda\epsilon_m)\frac{C_1 \prod \mathbf{q}(z(x_i), z(x_j)|\phi_L^*)^\lambda}{\int \prod \mathbf{q}(z(x_i), z(x_j))^\lambda \pi(\phi)d\phi}\bigg| A_m^C(\frac{\delta_m}{4})\right] \\
&\leq P_Z(A_m(\frac{\delta_m}{4})) \\
&\quad + \mathbb{E}_Z\left[\frac{C_1 \exp -\lambda(\epsilon_m - \delta_m)}{\pi(\phi : \log \prod \mathbf{q}(z(x_i), z(x_j)) \geq \log \prod \mathbf{q}(z(x_i), z(x_j)|\phi_L^*) - \delta_m)}\bigg| A_m^C(\frac{\delta_m}{4})\right] \\
&\leq P_Z(A_m(\frac{\delta_m}{4})) \\
&\quad + \frac{C_1 \exp -\lambda(\epsilon_m - \delta_m)}{\pi(\phi : \mathbb{E}_Z \log \prod \mathbf{q}(z(x_i), z(x_j)) \geq \mathbb{E}_Z \log \prod \mathbf{q}(z(x_i), z(x_j)|\phi_L^*) - \delta_m/2)} \\
&= P_Z(A_m(\frac{\delta_m}{4})) + \frac{C_1 \exp -\lambda(\epsilon_m - \delta_m)}{\pi(\phi : h(\phi) \geq h(\phi_L^*) - \delta_m/2)},
\end{aligned}
$$

where we define $h(\phi) := \mathbb{E}_Z \log \prod \mathbf{q}(z(x_i), z(x_j))$. Let n_d be the number of pairs (x_i, x_j) such that $\|x_i - x_j\| = d$, and $\mathcal{D}$ be the set of such d. For any $\phi \geq \phi_1$ and $d \geq d_1$,

$$|h(\phi_L^*) - h(\phi)| \leq C_2 |\phi - \phi_L^*| \sum_{d \in \mathcal{D}} n_d d$$

for some constant $C_2 > 0$ that depends on ϕ_1, d_1. From the assumption on the prior π,

$$\pi(\phi : h(\phi_L^*) - h(\phi) \leq \frac{\delta_m}{2}) \geq \pi\left(\phi : |\phi - \phi_L^*| \leq \frac{\delta_m}{2C_2 \sum_d n_d d}\right) \geq \left(\frac{\delta_m}{2C_2 \sum_d n_d d}\right)^r.$$

Thus we obtain

$$P^*(\log P_1(\phi_L^*|) - \log P_1(\phi|) \geq \epsilon_m)$$
$$\leq 2 \exp \frac{-\delta_m^2}{4m(m-1)M^2} + \frac{C_1 \exp(-\lambda(\epsilon_m - \delta_m))(2C_2 \sum_d n_d d)^r}{\delta_m^r}.$$

Let $\delta_m = \epsilon_m/2$ and $\epsilon_m \sim m$, it follows that under P^*, $\log P_1(\phi_L^*|) - \log P_1(\phi|) = O_P(m)$, which means $\{1/[m(m-1)/2]\} \log P_1(\phi_L^*|)/P_1(\phi|) = O_P(1/m)$.

Proof of Lemma 11. (sketch) Using standard calculations for exponential families, for each pair of values $(u, v) \in \{1, \ldots, k\}^2$, taking the derivative of $D(\mathbf{q} \| Q)$ with respect to $q_{ij}(Z(x_1) = u, Z(x_2) = v)$ and setting to 0 we can easily obtain the desired result.

Proof of Proposition 12. (sketch)
(a) The proof proceeds by induction. The result clearly holds for $m = 2$. For $m > 2$, assume that x_0 corresponds to a leaf node and let $E' = E - \{x_0\}$. It is simple to show that the marginal distribution generating the remaining $m - 1$ nodes follow the form

$$\tilde{\mathbf{q}}_{E'}(Z(x_2), \ldots, Z(x_m)) = \sum_{Z(x_1)=1}^{2} \tilde{\mathbf{q}}_E(Z(x_1), \ldots, Z(x_m)) \propto \prod_{(i,j) \in E'} \mathbf{q}_{ij}(Z(x_i), Z(x_j)),$$

so, by induction, it has uniform marginal at each single node corresponding to $x_2, \ldots, x_m$. Apply the same step to another subtree to obtain that the marginal for $Z(x_1)$ is also uniform.
(b) The proof for the first result is straightforward by induction based on the following fact: $A(E) = A(E') + \log 2$. The second result is a known fact for tree-structured graphical models.

(c) To understand the behavior of A, it is useful to interpret it as a function of parameter θ, $A(\theta)$ from now on, via

$$\theta_{ij} = \log \frac{\mathbf{q}(Z(x_i) \neq Z(x_j))}{\mathbf{q}(Z(x_i) = Z(x_j))} \text{ for } (i,j) \in E \; ; \; 0 \text{ otherwise,}$$

$$\theta_E = \{(\theta_{ij}) \mid 1 \leq i, j \leq m\},$$

$$\tilde{\mathbf{q}}_E() = \exp\left\{ \sum_{(i,j)\in E} \theta_{ij} \mathbb{I}(Z(x_i) \neq Z(x_j)) - B_E(\theta) \right\},$$

$$B(\theta_E) = \log \sum_Z \exp\left\{ \sum_{(i,j)\in E} \theta_{ij} \mathbb{I}(Z(x_i) \neq Z(x_j)) \right\},$$

$$A(\theta_E) = B(\theta_E) + \sum_{(i,j)\in E} \log \frac{1}{2}\mathbf{q}(Z(x_i) = Z(x_j)).$$

As a standard fact of exponential families, $B : \mathbb{R}^{m(m-1)/2} \to \mathbb{R}$ is a convex function with respect to θ_E. In addition, $\nabla_{\theta_E} B(\theta_E) = \tilde{\mathbf{q}}_E(Z(x_i) \neq Z(x_j))$. Due to the convexity, we havd:

$$B(\theta_E) \geq B(\theta_{E_0}) + (\theta_E - \theta_{E_0})\nabla_{\theta_E} B(\theta_{E_0}) \text{ and}$$

$$B(\theta_{E_0}) \geq B(\theta_E) + (\theta_{E_0} - \theta_E)\nabla_{\theta_E} B(\theta_E).$$

These inequalities lead to the desired result.

References

Abraham, C., Cornillon, P., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Statist.* **30**, 581–595.

Barron, A., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist* **27**, 536-561.

Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory* **54**, 781-790.

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993-1022.

Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.* **93**, 961-980.

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. Roy. Statist. Soc. Ser. B* **69**, 679-699.

Dabo-Niang, S., Ferraty, F. and Vieu, P. (2006). Mode estimation for functional random variable and its application for curves classication. *Far East J. Theor. Stat.* **18**, 93-119.

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

Duan, J., Guidani, M. and Gelfand, A. (2007). Generalized spatial Dirichlet processes. *Biometrika* **94**, 809-825.

Dunson, D. (2008a). Kernel local partition processes for functional data. Tech. Rep. 26, Department of Statistical Science, Duke University.

Dunson, D. (2008b). Nonparametric Bayes local partition models for random effects. *Biometrika*, to appear.

Dunson, D. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307-323.

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

Fernandez, C. and Green, P. (2002). Modelling spatially correlated data via mixtures: A Bayesian approach. *J. Roy. Statist. Soc. Ser. B* **64**, 805-826.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, New York.

Fraiman, R., Justel, A. and Svarc, M. (2008). Selection of variables for cluster analysis and classication rules. *J. Amer. Stat. Assoc.* **103**, 1294-1303.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test* **10**, 419-440.

Gelfand, A., Kottas, A. and MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100**, 1021-1035.

Ghosal, S. (2007). Dirichlet process, related priors and posterior asymptotics. Manuscript.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.

Griffin, J. and Steel, M. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 179-194.

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

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12**, 941-963.

James, G. M. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *J. Amer. Stat. Assoc.* **98**, 397-408.

Ma, P. and Zhong, W. (2008). Penalized clustering of large-scale functional data with multiple covariates. *J. Amer. Statist. Assoc.* 103, 625-636.

MacEachern, S. (2000). Dependent Dirichlet processes. Tech. rep., Ohio State University.

MacLehose, R. F. and Dunson, D. (2008). Nonparametric Bayes kernel-based priors for functional data analysis. *Statist. Sinica*, to appear.

Petrone, S., Guidani, M. and Gelfand, A. (2009). Hybrid Dirichlet processes for functional data. *J. Royal Stat. Soc. Ser. B*, to appear.

Pillai, N., Liang, F., Mukerjee, S., Wolpert, R. and Wu, Q. (2006). Characterizingthe function space for Bayesian kernel models. Tech. rep., Departmet of Statistical Science, Duke University.

Pritchard, J., Stephens, M. and Donnelly, P. (2000). Inference of populaton structure using multilocus genotype data. *Genetics* **155**, 945-959.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies.* Springer-Verlag, New York.

Ramsay, J. O. and Silverman, B. (2006). *Functional Data Analysis.* 2nd edition. Springer, New York.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahr. Verw. Gebiete* **4**, 10-26.

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

Stein, M. (1999). *Interpolation of Spatial Data*. Springer-Verlag, New York.

Sudderth, E., Torralba, A., Freeman, W. and Willsky, A. (2008), Describing visual scenes using transformed objects and parts. *Internat. J. Computer Vision* **77**.

Teh, Y., Jordan, M., Beal, M. and Blei, D. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566-1581.

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **32**, 1265-1269.

Tokushige, S., Yadohisa, H. and Inada, K. (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Comput. Statist.* **22**, 1-16.

Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential families, and variational inference. Tech. Rep. 649, Dept of Statistics, UC Berkeley.

Winn, J., Criminisi, A. and Minka, T. (2005). Object categorization by learned universal visual dictionary. in *Proc. IEEE Intl. Conf. on Computer Vision* (ICCV).

Wu, Y. and Ghosal, S. (2009). L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation, submitted.

Zhang, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum complexity density estimation. *Ann. Statist.* **34**, 2180-2210.

Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA.

E-mail: xuanlong@umich.edu

Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA.

E-mail: alan@stat.duke.edu