

The Dirichlet labeling process for functional data analysis

XuanLong Nguyen & Alan E. Gelfand *

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 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: 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 involving 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 et al. (2005); Teh et al. (2006)).

In this paper we are interested in mixture modeling for functional data. From the viewpoint of functional data analysis we are given a sample of n functions (curves) Y_1, \dots, Y_n over \mathbb{R}^d , each viewed as a realization

*XuanLong Nguyen is a postdoctoral researcher and Alan Gelfand is a Professor in the Department of Statistical Science at Duke University, Durham, NC 27708-0251. The work of the first author was supported in part through a SAMSI postdoctoral fellowship and the work of the second author was supported in part under NSF award DMS-0504953. The authors thank Sonia Petrone for valuable conversations, and David Dunson for helpful comments related to this work.

of a stochastic process Y . The curves are observed at a common set of locations $x_1, \dots, x_m \in D$, where D is a subset of \mathbb{R}^d .¹ 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 of this paper is to examine clustering of the set of curves. Formalizing the notion of clustering of curves raises several interesting challenges. First, we envision the Y_i 's as *noisy* versions of the curves θ_i . The θ_i 's are assumed to be smooth (i.e., 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 will require a distinct θ_i - no clustering. Second, we can envision local clustering, global clustering, and, possibly, attempt to formalize notions in between, notions of “partial” clustering. In particular, one can envision clustering curve realizations $\theta_1(x), \dots, \theta_n(x)$ at each location $x \in D$ into groups of observations using a DP mixture. 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 base measure (Gelfand et al., 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 et al., 2003;

¹The case where some locations are not common to all curves can be handled as well through inclusion of additional latent variables but is not detailed here.

Pritchard et al., 2000), as well as in the context of spatial and functional data (Duan et al., 2007; Petrone et al., 2008). In particular, our approach is based on the hybrid Dirichlet process mixture model first introduced by Petrone et al. (2008). Implicit in their modeling is a latent group allocation process, which we shall call the *Dirichlet labeling process*. This labeling process, which we denote by \mathbf{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 \mathbf{p} is a random probability measure on $\{1, \dots, k\}^D$, which 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 $\{1, \dots, k\}^D$. Explicitly, we mean that for any finite set of locations, $\{x_j, j = 1, 2, \dots, m\}$, \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 $k \rightarrow \infty$, 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 et al., 2008).

The novel contributions offered here is the detailed investigation of the Dirichlet labeling process model. This model 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. We also analyze the local and global clustering behavior in the overall mixture model. Statistical inference with the latent labeling process is expensive with a large number of local sites and clusters. We offer 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 pa-

pers introduce various constructions based on the Sethuraman’s stick-breaking representation, with varying weights assigned for different locations (e.g., Griffin and Steel (2006); Dunson and Park (2008); Duan et al. (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 (e.g., Pillai et al. (2006); MacLehose and Dunson (2008); Dunson (2008b,a)), i.e., $f(x) = \int K(x, u)\gamma(u)du$, 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 et al. (2007). This work 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 which 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.

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. Section 4 briefly discusses identifiability issues associated with the full hierarchical model we work with. Section 5 focuses on approximate estimation methods. Section 6 offers experiment results. We conclude with Section 7. All proofs and additional computational details can be found in the Appendix.

2 Formalizing the model

As described in the Introduction, we now define a mixture model for curve realizations Y_1, \dots, Y_n over \mathbb{R}^D , which are noisy versions of, respectively, $\theta_1, \dots, \theta_n$. The observations are obtained at local sites $x_1, \dots, x_m \in D$. Each curve θ_i is described by the label function $\{L_i : D \rightarrow \{1, \dots, k\}\}$, where k denotes the number of available labels (species). For any finite set of locations $x_1, \dots, x_m \in D$, we specify the random distribution $\mathbf{p}_{x_1, \dots, x_m}$ which is such that $\mathbf{p}_{x_1, \dots, x_m}(l_1, \dots, l_m) = P(L(x_1) = l_1, \dots, L(x_m) = l_m)$. In particular, the (random) label function L yields a marginal distribution which is a multinomial with probabilities \mathbf{p}_x at each site x , i.e., $P(L(x) = j) = \mathbf{p}_x(j)$ for $j = 1, \dots, k$. Below, the collection of $\mathbf{p}_{x_1, \dots, x_m}$ is specified to consistently determine a probability measure \mathbf{p} on $\{1, \dots, k\}^D$, which is randomly generated by the Dirichlet labeling process that we shall formally define. In addition, we need a collection of k ‘‘canonical’’ species $\theta_1^*, \dots, \theta_k^*$, which are drawn from G_0 independently of L . Conditioning on labels L_i and canonical species θ^* , the curve θ_i is realized at each location x by associating the value at x of the species indexed by the label $L_i(x)$, i.e., $\theta_i(x) = \theta_{j_i}^*(x)$ if $L_i(x) = j_i$. Formally:

$$\begin{aligned} \theta_j^* &\stackrel{iid}{\sim} G_0, \quad j = 1, \dots, k, \\ L_i | \mathbf{p} &\stackrel{iid}{\sim} \mathbf{p}, \quad i = 1, \dots, n, \\ \theta_i(x_t) | L, \theta^* &= \theta_{L_i(x_t)}^*, \quad i = 1, \dots, n; \quad t = 1, \dots, m \\ Y_i(x_t) | \theta_i(x_t) &\sim N(\theta_i(x_t), \tau^2). \end{aligned}$$

In addition, depending on application, there may be prior distributions for G_0 and τ . Also, there may be covariate information, say $\mathbf{Z}_i(x_t)$, which can be included in the mean for $Y_i(x_t)$. Alternatively, we can write $\theta_i \stackrel{iid}{\sim} G$ for $i = 1, \dots, n$, where G is a random measure on \mathbb{R}^D such that

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

We define the Dirichlet labeling process which generates the random multinomial distribution \mathbf{p} for the label function L . \mathbf{p} is a random probability measure on $\{1, \dots, k\}^D$. For locations x_1, \dots, x_m , $\mathbf{p}_{x_1, \dots, x_m}$ has a k^m -dimensional Dirichlet distribution:

$$(\mathbf{p}_{x_1, \dots, x_m}(j_1, \dots, j_m), j_i = 1, \dots, k) \sim Dir(\alpha \mathbf{q}_{x_1, \dots, x_m}(j_1, \dots, j_m), j_i = 1, \dots, k), \quad (2)$$

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

The base measure \mathbf{q} is constructed such that \mathbf{q} has a uniform marginal distribution at every location $x \in D$, i.e., $\mathbf{q}_x(1) = \dots = \mathbf{q}_x(k) = 1/k$. 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, \dots, x_m} the finite-dimensional distributions of F . Let $(\eta(x_1), \dots, \eta(x_m)) \sim F_{x_1, \dots, x_m}$, and consider the random vector $(F_{x_1}(\eta(x_1)), \dots, F_{x_m}(\eta(x_m))) \in [0, 1]^m$, 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, \dots, x_m} . The collection of finite-dimensional d.f. H_{F, x_1, \dots, x_m} characterizes a probability measure H_F on $[0, 1]^D$. Now, let us discretize $[0, 1]^m$ into hyper-cubes

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

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

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

Alternatively, for each $x \in D$, let $c_1(x), \dots, c_k(x)$ be an increasing sequence of threshold values in \mathbb{R} such that $F_x(c_j(x)) = j/k$, for $j = 1, \dots, k - 1$. Complement the sequence with $c_0(x) = -\infty$ and $c_k(x) = \infty$. Conditioning on the realization $\eta = (\eta(x_1), \dots, \eta(x_m))$, define function $Z : D \rightarrow \{1, \dots, k\}$ such that for each $j = 1, \dots, k$,

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

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

It is useful to note that just as \mathbf{q} is defined in terms of auxiliary variables $\eta \sim F$, \mathbf{p} can be defined directly in terms of auxiliary variables ξ without going through the labeling function $Z \sim \mathbf{q}$. Define a random function ξ on \mathbb{R}^D such that $\xi \sim H$, where $H \sim DP(\alpha F)$. For any $x \in D$, define:

$$\tilde{L}(x) = j \Leftrightarrow \xi(x) \in (c_{j-1}(x), c_j(x)] \Leftrightarrow F_x(\xi(x)) \in ((j-1)/k, j/k]. \quad (3)$$

Marginalizing over ξ and H we obtain a random probability distribution $\tilde{\mathbf{p}}$ generating \tilde{L} . It can be shown that $\mathbf{p} \stackrel{d}{=} \tilde{\mathbf{p}}$ and $L \stackrel{d}{=} \tilde{L}$. Indeed, for any $x_1, \dots, x_m \in D$, the random vector $\tilde{L} = (\tilde{L}(x_1), \dots, \tilde{L}(x_m))$ has

to satisfy, due to the definition of the Dirichlet process,

$$\begin{aligned}
& \left(\tilde{\mathbf{p}}(\tilde{L} = (j_1, \dots, j_m)), j_i = 1, \dots, k \right) \\
& \stackrel{d}{=} \left(\tilde{\mathbf{p}}(\xi(x_1) \in (c_{j_1-1}(x_1), c_{j_1}(x_1)], \dots, \xi(x_m) \in (c_{j_m-1}(x_m), c_{j_m}(x_m)]), j_i = 1, \dots, k) \right) \\
& \sim \text{Dir}(\alpha F(\eta(x_1) \in (c_{j_1-1}(x_1), c_{j_1}(x_1)], \dots, \eta(x_m) \in (c_{j_m-1}(x_m), c_{j_m}(x_m)]), j_i = 1, \dots, k) \\
& = \text{Dir}(\alpha \mathbf{q}_{x_1, \dots, x_m}(j_1, \dots, j_m), j_i = 1, \dots, k).
\end{aligned}$$

This implies that $\mathbf{p} \stackrel{d}{=} \tilde{\mathbf{p}}$ and $L \stackrel{d}{=} \tilde{L}$.

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} . We have given equivalent characterizations of \mathbf{p} in terms of latent process ξ , or in terms of latent process η and labeling vector Z . We shall see that the latter characterization is much more convenient to work with.

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, i.e., of 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 \mathbf{p} . From (2), \mathbf{p} and \mathbf{q} are related by relation: $\mathbf{p}|\mathbf{q} \sim DP(\alpha\mathbf{q})$. As a result, properties obtained for the labeling process Z can be easily incorporated into that for L . We start with the following elementary properties for \mathbf{p} :

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 $\mathbf{p}_x \sim \text{Dir}((\alpha/k)\mathbf{1})$.

(b) Let $L_1, L_2 | \mathbf{p} \stackrel{iid}{\sim} \mathbf{p}$. Then, unconditionally, $P(L_1(x) = L_2(x)) = \frac{1}{k} + (1 - 1/k) \frac{1}{\alpha+1}$.

(c) Let $L_1, L_2 | \mathbf{p} \stackrel{iid}{\sim} \mathbf{p}$, and $x_1, \dots, x_m \in D$. Then, unconditionally,

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

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

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

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 $\alpha \rightarrow \infty$, \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 $\mathbf{q}_{x_1, \dots, x_m}(\cdot)$ is of order $O(1/k^m)$ (cf. Prop. 3 and the Appendix), part (c) implies the global clustering probability $P(L_1 = L_2) \sim \frac{1}{\alpha+1} + \frac{\alpha}{\alpha+1} \cdot \frac{1}{k^m} \rightarrow \frac{1}{\alpha+1}$ as $m \rightarrow \infty$. On the other hand, at each local site x , the probability of clustering is substantially higher:

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

However, since the probability of global clustering is still $\frac{1}{\alpha+1}$, there are evident implications regarding either the specification of α or a prior for it.

When $k \rightarrow \infty$, 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 \mathbf{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, $\rho_{12}(x_1, x_2) = \text{cov}(\eta(x_1), \eta(x_2)) = \exp(-\phi_L \|x_1 - x_2\|)$ 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, \dots, 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 can write $\mathbf{q}(\phi_L, k)$. Although it is easy to generate a random sample of $(Z(x_1), \dots, Z(x_m)) \sim \mathbf{q}$, 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 $\rho_{12} = \text{Cov}(\eta(x_1), \eta(x_2))$. 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 $\phi_L \rightarrow \infty$ or $\|x_1 - x_2\| \rightarrow \infty$, $\rho_{12} \rightarrow 0$ 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 $\phi_L \rightarrow 0$ or $\|x_1 - x_2\| \rightarrow 0$, $Z(x_1)$ and $Z(x_2)$ share equal values with increasing probability.

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

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 $k \rightarrow \infty$, then*

$$P(Z(x_1) = i, Z(x_2) = j) = q_{x_1, x_2}(i, j) = \frac{1}{k^2}(R_{ij}(c_i, c_j) + o(1)), \quad (4)$$

$$P(Z(x_1) = i, Z(x_2) > j) = \frac{1}{k} \left(1 - \Phi \left(\frac{c_j - c_i \rho_{12}}{1 - \rho_{12}^2} \right) + o(1) \right), \quad (5)$$

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

$$R_{ij}(c_i, c_j) = \frac{1}{\sqrt{1 - \rho_{12}^2}} \exp - \frac{(c_i^2 + c_j^2)\rho_{12}^2 - 2\rho_{12}c_i c_j}{2(1 - \rho_{12}^2)}. \quad (6)$$

Prop. 3 can also be extended to arbitrary number of locations x_1, \dots, 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 Prop. 3. As $\rho_{12} \rightarrow 0$, we have $R_{ij}(c_i, c_j) \rightarrow 1$, so that $P(Z(x_1) = i, Z(x_2) = j) \rightarrow 1/k^2$, i.e., $Z(x_1)$ and $Z(x_2)$ become less dependent. On the other hand, as $\rho_{12} \rightarrow 1$, for any pair $i \neq j$, $R_{ij}(c_i, c_j) \rightarrow 0$, 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 $P(Z(x_2)|Z(x_1) = i) \approx \frac{1}{k} R_{ij}(c_i, c_j)$. $R_{ij}(c_i, c_j)$ 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 which is drawn from the random probability measure G (see Eqn. (1)), we have

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

where the randomness of G is due to the randomness of \mathbf{p} and θ^* . We posit that the canonical θ^* be smooth curves by placing a zero-mean Gaussian process prior G_0 on θ^* with variance σ_θ^2 and covariance function

$\rho_\theta(x_1, x_2) = \sigma_\theta^2 \exp -\phi_\theta \|x_1 - x_2\|^2$.² It is simple to obtain the following:

$$\begin{aligned}\mathbb{E}[\theta(x)|\mathbf{q}, G_0] &= \mathbb{E}[\theta^*(x)|G_0] = 0, \\ \mathbb{E}[\theta(x_1)\theta(x_2)|\mathbf{q}, G_0] &= \sum_{j=1}^k \mathbf{q}_{x_1, x_2}(j, j) \text{Cov}(\theta^*(x_1), \theta^*(x_2)).\end{aligned}$$

As $\|x_1 - x_2\| \rightarrow \infty$, $\text{Cov}(\theta^*(x_1), \theta^*(x_2)) \rightarrow 0$, so $\text{Cov}(\theta(x_1), \theta(x_2)|\mathbf{q}, G_0) \rightarrow 0$. As $\|x_1 - x_2\| \rightarrow 0$, $\sum_{j=1}^k \mathbf{q}_{x_1, x_2}(j, j) \rightarrow 1$, so $\text{Cov}(\theta(x_1), \theta(x_2)|\mathbf{q}, G_0) \rightarrow \sigma_\theta^2$. 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 . That is, let $\theta \sim G$. For any $x_1 \in D$,*

$$\lim_{x_2 \rightarrow x_1} \mathbb{E}(\theta(x_1) - \theta(x_2))^2 \rightarrow 0.$$

Although θ is mean square continuous, each realization is almost surely discontinuous as it is composed of multiple smooth segments of the canonical curves. The discontinuity is smoothed in the Y_i 's by the mixing with the noise or pure error process (also called the nugget process, i.e., i.i.d. random variables at locations $\epsilon(x) \sim N(0, \tau^2)$) to obtain $Y(x) = \theta(x) + \epsilon(x)$ for any $x \in D$.³ The joint density for $\mathbf{Y} = (Y(x_1), \dots, Y(x_m))$ given G and τ^2 is

$$f(\mathbf{Y}|G, \tau) = \int N_m(\mathbf{Y}|\theta, \tau^2 \mathbf{I}_m) G(d\theta). \quad (7)$$

It follows that $\mathbb{E}(\mathbf{Y}|\mathbf{q}, G_0) = \mathbb{E}(\theta^*|G_0)$ and the covariance matrix $\Sigma_{\mathbf{Y}}|\mathbf{q}, G_0 = \tau^2 \mathbf{I}_m + \Sigma_\theta$, where $(\Sigma_\theta)_{ij} = \text{Cov}(\theta(x_i), \theta(x_j)|\mathbf{q}, G_0)$. Finally, depending on applications it is simple to include covariate information into the mean structure of \mathbf{Y} .

4 Identifiability

The described Dirichlet 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

²Depending on applications, less smooth canonical curves can be desired, e.g., using exponential covariance function $\rho_\theta(x_1, x_2) = \sigma_\theta^2 \exp -\phi_\theta \|x_1 - x_2\|$.

³Alternatively, one could envision θ as the coefficient process of a kernel based representation, i.e. by convolving θ with a kernel basis function to obtain a rich class of smooth functions. We do not pursue this direction further in this paper.

is an important issue requiring careful prior specification. The previous section has 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 identifiability of the labeling L and canonical curves θ^* .

Suppose that we are interested in a representation which achieves dimensionality reduction with the goal to infer about both canonical curves θ^* and labeling L_1, \dots, L_n for observed replicates Y_1, \dots, Y_n . In this scenario the canonical curves can be viewed as basis functions and the label vectors L_1, \dots, L_n provides 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 6.

ϕ_L also plays an important role in the identifiability of the canonical species curves θ_j . When ϕ_L is close to 0, as shown by Prop 2 and Prop 3, label switching 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 , the canonical curves become more weakly identified. As we illustrate in Section 6 our model is able to recover segments of locations that admit relatively few switchings. 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, \dots, L_n as a means for characterizing and clustering the observed replicates Y_1, \dots, Y_n . In this scenario, strong constraints can be imposed upon θ^* to improve the model identifiability. 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 $\{1, \dots, k\}$. The previous discussion on properties of \mathbf{p} and \mathbf{q} suggests that for large k there is a

natural ordering of label values $\{1, 2, \dots, k\}$. 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 $\mathbb{E}(\theta_1^*) < \mathbb{E}(\theta_2^*) < \dots < \mathbb{E}(\theta_k^*)$. Note that such ordering constraints are not necessary to ensure model identifiability, but they would be expected to improve the mixing for simulated posterior inference.

5 Model fitting and inference

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

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

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 $[L_1, \dots, L_n, \theta_1^*, \dots, \theta_k^*, \phi_L, \alpha, \tau, \phi_\theta, \sigma_\theta | \mathbf{Y}]$. The updates of parameters $\alpha, \tau, \phi_\theta, \sigma_\theta$ are standard, see e.g., Duan et al. (2007). For canonical curves, under a Gaussian process, the prior for vector $\theta_j^* = (\theta_j^*(x_1), \dots, \theta_j^*(x_m))$ 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 equal to $\mathbb{I}(L_i(t) = j)$. The full conditional for θ_j^* has the form:

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

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

We now turn our attention to updating label vectors $L_i, i = 1, \dots, n$ and decay parameter ϕ_L . Due to the alternative characterization of latent labels L captured by (3), one simple method is to sample directly 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 $\tilde{\xi}_i$, which is a perturbed version of ξ by a small independent noise: $\tilde{\xi}_i = \xi + \epsilon$, 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 will 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 the Polya urn's 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 $Z_i \sim \mathbf{q}$ and the latent vector η_i for $i = 1, \dots, n$. Furthermore, by the virtue of Prop. (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. The Gibbs sampling procedure is now straightforward by applying the Polya urn sampling scheme. Here, we have, for say curve 1 at say, x_1 , that the conditional label distribution is

$$P(L_1(x_1)|L_1(x_2), \dots, L_1(x_m), \text{the rest}) \propto \sum_{i=2}^n \frac{\mathbb{I}(L_1 = L_i)}{\alpha + n - 1} N(Y_1|\theta_{L_i}) + \frac{\alpha}{\alpha + n - 1} N(Y_1|\theta_{L_1}) \mathbf{q}_{x_1, \dots, x_m}(L(x_1), \dots, L_1(x_m)|\phi_L, k).$$

The likelihood function under \mathbf{q} is obtained via Prop. 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, \dots, Z_n drawn from \mathbf{q} , observed values at locations x_1, \dots, x_m . We first consider the point estimation problem for ϕ_L . Suppose we have multiple curves, indexed by $i = 1, 2, \dots, n$, 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 closed form given in Proposition 2. The more typical scenario, however, is when m much larger than 2, while 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, \dots, x_m : $z = (z(x_1), \dots, z(x_m))$? An intuitive approach is to maximize a pseudo-likelihood

for z , which 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). We can show formally:

Proposition 5. *Suppose that $Z = (Z(x_1), \dots, Z(x_m))$ 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*

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

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 $k > 2$.

	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

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 $\phi_L = 0.5, k = 2$.

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 $Z = (z(x_1), \dots, z(x_m))$. We propose to use the aforementioned pseudo-likelihood to obtain what we term a ‘‘Gibbs posterior’’ distribution (Zhang, 2006) for ϕ_L as follows:

$$P_\lambda(\phi_L | Z) \propto \prod_{1 \leq i < j \leq m} \mathbf{q}(Z(x_i) = z(x_i), Z(x_j) = z(x_j)) | \phi_L)^\lambda \pi(\phi_L). \quad (8)$$

$\lambda > 0$ 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 6. *Suppose that $Z = (Z(x_1), \dots, Z(x_m))$ is drawn from \mathbf{q} equivalently $F_{\phi_L^*}$ for some $\phi_L^* > 0$; and that $\pi(\cdot)$ puts sufficient mass around ϕ_L^* (i.e., 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 :

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

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 $\{(i, j) \mid 1 \leq i < j \leq m\}$. E could be viewed as a collection of edges connecting the vertices $x_1, \dots, x_m \in D$ to form a graphical structure. Our strategy is to approximate the multivariate distribution $\mathbf{q}(Z(x_1), \dots, Z(x_m))$ by a graphical model distribution $\tilde{\mathbf{q}}$ defined as:

$$\tilde{\mathbf{q}}_E(Z(x_1), \dots, Z(x_m)) \propto \prod_{(i,j) \in E} \mathbf{q}(Z(x_i), Z(x_j)). \quad (9)$$

Consequentially, the conditional probability distribution for the labels is approximated by

$$\tilde{\mathbf{q}}_E(Z(x_1)|Z(x_2), \dots, Z(x_m)) \propto \prod_{j \neq 1} \mathbf{q}(Z(x_1), Z(x_j)).$$

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 7. Consider a class of probability distributions on $(Z(x_1), \dots, Z(x_m)) \in \{1, 2\}^m$:

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

where q_{ij} 's are any function on $\{1, 2\}^2$. Then the distribution $\tilde{\mathbf{q}}_E$ defined in (9) satisfies:

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

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 via the following result:

Proposition 8. (a) The marginal distribution under $\tilde{\mathbf{q}}_E$ is uniform (i.e., $1/2$).

(b) If E forms a spanning tree then $A(\theta_E) = -(m-2) \log 2$. Furthermore, $\tilde{\mathbf{q}}_E(Z(x_i), Z(x_j)) = \mathbf{q}(Z(x_i), Z(x_j))$

for any $(i, j) \in E$.

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

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

$$\begin{aligned} U &= \sum_{(i,j) \in E - E_0} \left(\tilde{\mathbf{q}}_{E_0}(Z(x_i) \neq Z(x_j)) \log \mathbf{q}(Z(x_i) \neq Z(x_j)) + \tilde{\mathbf{q}}_{E_0}(Z(x_i) = Z(x_j)) \log \mathbf{q}(Z(x_i) = Z(x_j)) \right) \\ V &= \sum_{(i,j) \in E - E_0} \left(\tilde{\mathbf{q}}_E(Z(x_i) \neq Z(x_j)) \log \mathbf{q}(Z(x_i) \neq Z(x_j)) + \tilde{\mathbf{q}}_E(Z(x_i) = Z(x_j)) \log \mathbf{q}(Z(x_i) = Z(x_j)) \right). \end{aligned}$$

For 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 $t = 1, \dots, m - 1$, assuming that $x_1 < x_2 < \dots < x_m$. 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)).

6 Applications

We demonstrate the behavior of the Dirichlet label process prior using simulated data in section 6.1. Sections 6.2 and 6.3 look at a collection of progesterone curves and a collection of images, respectively.

6.1 Synthetic data

First we illustrate the fitting of our 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 $[x_1, \dots, x_m] = [1, 3, \dots, 39]$ while leaving out 20 other locations $2, 4, \dots, 40$ for validation purposes. θ_j^* for $j = 1, \dots, k$ are independently drawn from a Gaussian process $GP(\mu_j, \phi_\theta, \sigma_\theta)$ at locations x_1, \dots, x_m , where $\mu_j = -1 + 2(j - 1)/(k - 1)$. The label vectors L_1, \dots, L_n are drawn from label process \mathbf{q} , which is drawn by known ϕ_L . Species $\theta_1, \dots, \theta_n$ are constructed by letting $\theta_i(x_t) = \theta_{L_i(t)}^*(x_t)$. Finally, the data collection Y_1, \dots, Y_n is obtained by mixing θ_i with an independent error process drawn from $N(0, \tau^2 I_m)$. We generate $n = 100$ sample curves using $k = 4$ canonical species curves. Parameter values for data generation are $\phi_\theta = 0.01$, $\sigma_\theta = 1$, $\phi_L = 0.05$, $\tau = 0.1$. For inference, we place a 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 canonical species curves are

obtained by running the MCMC algorithm for 4000 iterations after a burn-in period of 1000 iterations. An examination of running traces suggests the sampling algorithm mixes well.

Fig. 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 ($j = 1, \dots, k$) 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 identifiability of the canonical 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. Fig. 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 is obtained from our sampling algorithm. In both cases, a uniform distribution prior $Uni[0.0001, 1]$ is placed on parameter ϕ_L , while $\phi_\theta = 0.005, \sigma_\theta = 1, \tau = 0.1$ are fixed. For smaller value of true ϕ_L (top figures), the posterior is well-concentrated around the true value. For larger ϕ_L (bottom figures), the posterior mass is shifted to the right, because the canonical species curve estimates (due to weak identifiability) tend to over-switch between the modes.

6.2 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

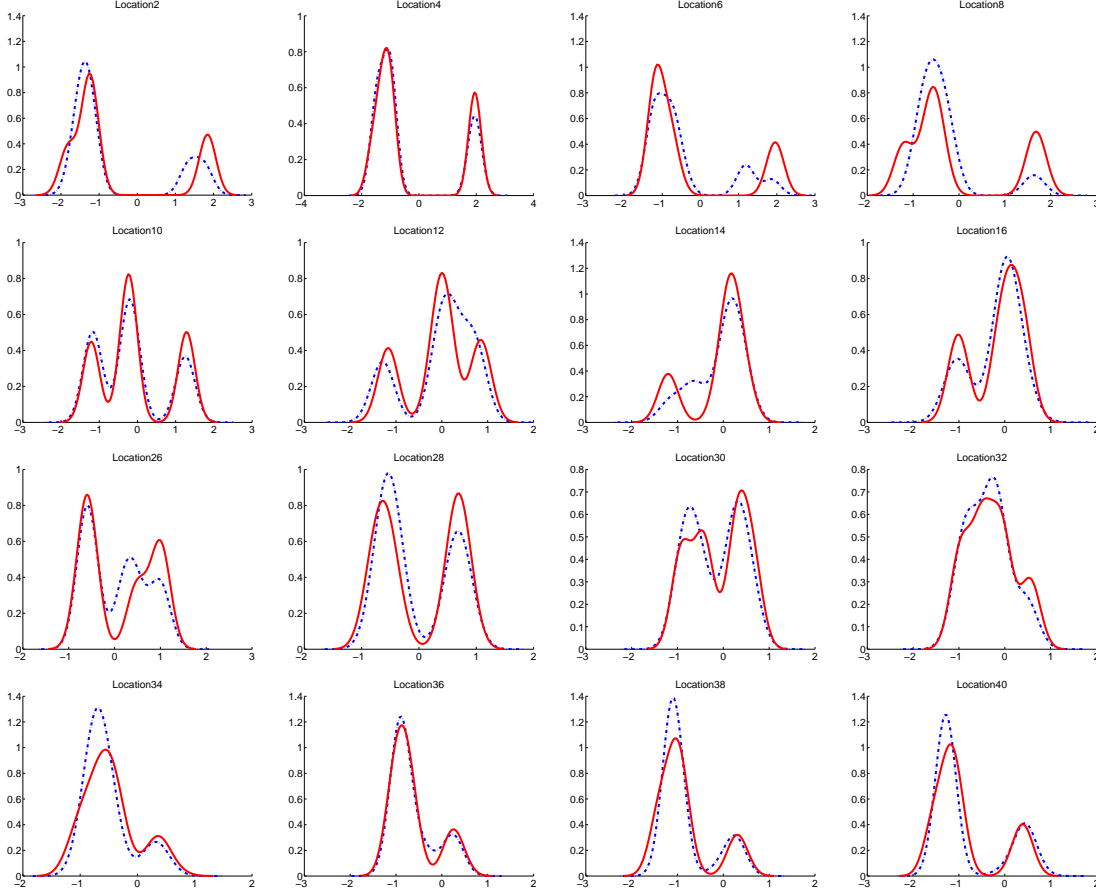


Figure 1. Evolution of posterior distributions at held-out locations $x = 2, 4, \dots, 40$. Solid plots are true distributions. Dashed plots are predictive distributions learned from the model.

non-contraceptive group, the remaining 22 cycles belong to the contraceptive group. This grouping is of course *unknown* to our analysis. See Fig. 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 Fig. 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 place a prior on the switching parameter $\phi_L \sim \text{Gam}(5, 2)$ so as to allow possible duplication of canonical curves in certain local segments. Canonical curves are drawn from mean-0 Gaussian process with a covariance matrix using decay parameter $\phi_\theta = 0.005$ and $\sigma_\theta = 1$. We fix the precision parameters $\tau = 1$, $\alpha = 1$. A discussion of the sensitivity of these parameters is included in the sequel. Samples from posterior distribution are

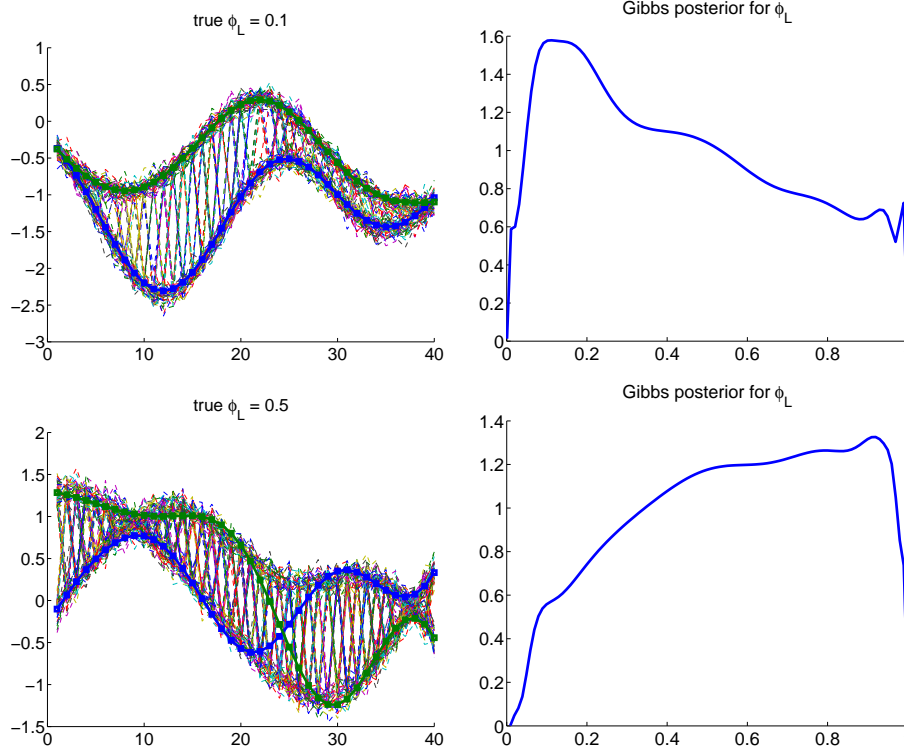


Figure 2. Illustration of canonical curve samples generated from the posterior in solid lines with squares. Figures to the right describe the corresponding Gibbs posterior for ϕ_L .

collected from 5000 MCMC iterations (discarding the first 1000). An examination of running traces suggests very fast mixing. Fig. 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. Fig. 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 Fig. 5, which illustrate the proportion of equal labels for pairs of curve replicates. Although global clustering is apparently not possible, one can observe local clustering effect by zooming in to the a curve segment corresponding to the last 5 days of the menstrual cycle. We also apply 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 are down-shifted by 2. Global

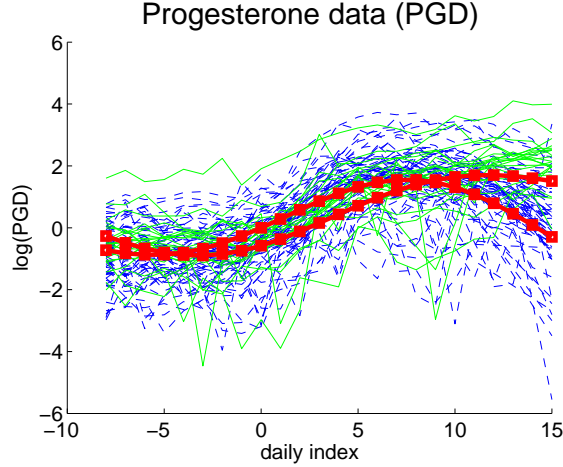


Figure 3. Monthly PGD cycle for contraceptive group (solid lines) and non-contraceptive group (dashed lines). Solid lines with squares are the mean estimate of canonical curves.

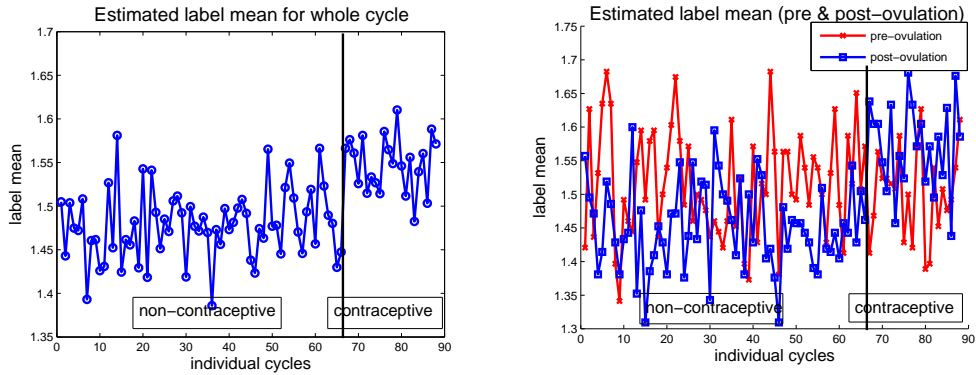


Figure 4. Left: Mean of estimated labels during the whole monthly cycle. Right: label means for pre and post-ovulation periods for 88 individuals (plots with x's and squares, resp.).

clustering is now easily achievable (see Fig. 6).

We now turn to a discussion of the effects of several parameters of interest on the identifiability of the canonical curves (see Fig. 7 for illustrative results). We observe that, as ϕ_L gets smaller, the model insists on increasingly global clusters (and less label switching for each replicate) resulting 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

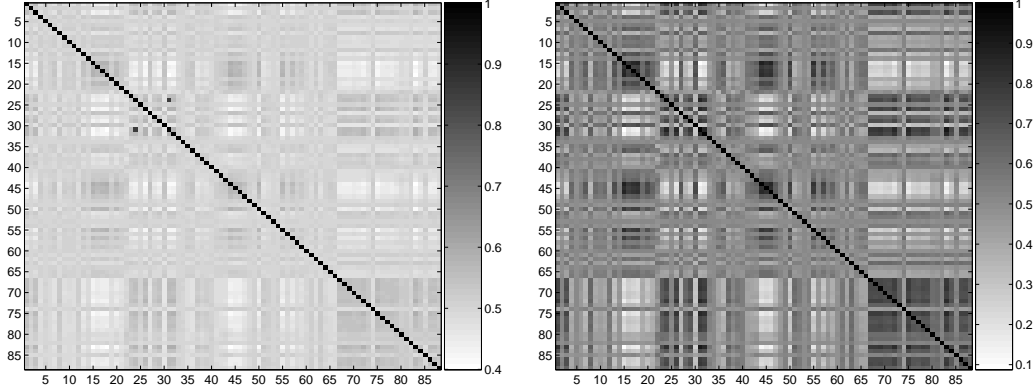


Figure 5. Heatmap illustrating proportion of equal labels for pairs of replicates for the whole curve (left), and a curve segment [20, 24] (i.e., last 5 days of the monitored cycle) (right).

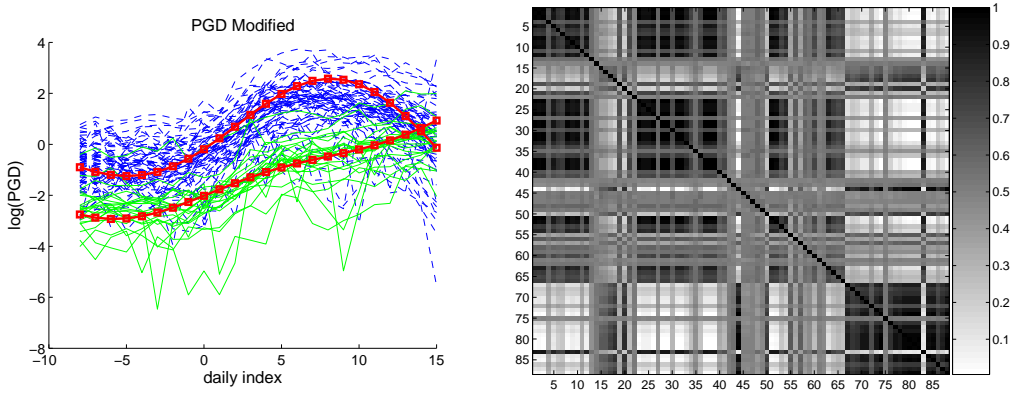


Figure 6. Analysis applied to the modified PG data set. Left: Mean estimate for the canonical curves. Right: Heatmap illustrating proportion of equal labels for pairs of replicates for the whole curve.

distinguishable canonical curves. The role of ϕ_θ is to dictate the smoothness of 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.

6.3 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 et al., 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 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 $i = 1, \dots, 80$, where $Y_i(x)$ is the color intensity of the location $x \in D = \{1, \dots, 26\} \times \{1, \dots, 40\}$

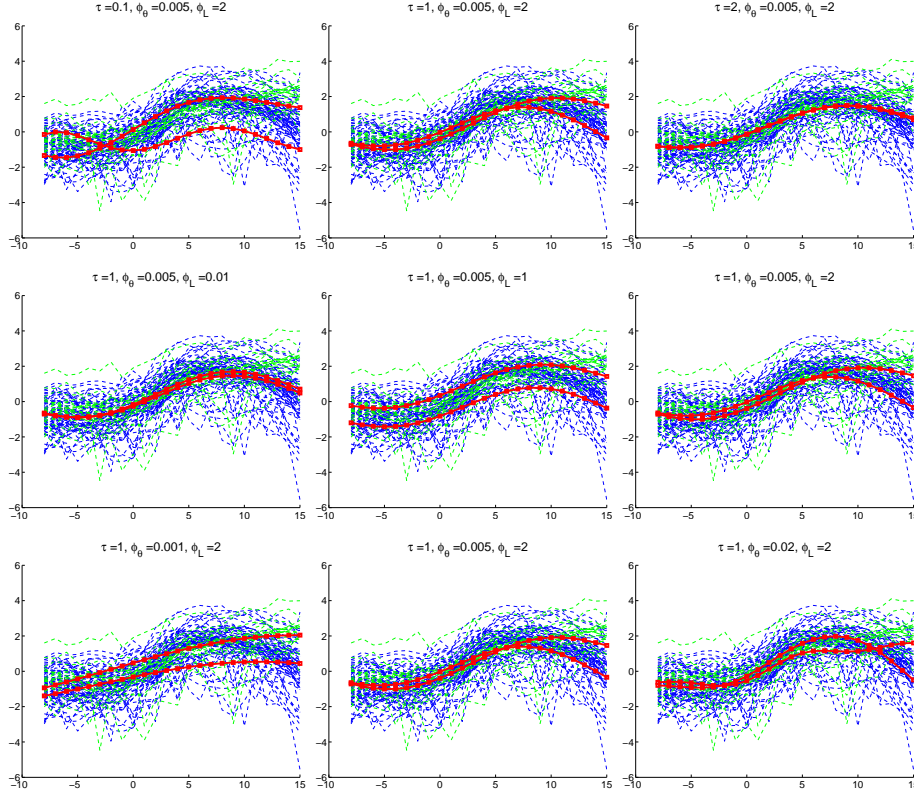


Figure 7. Top row: effects of $\tau = .1, 1, 2$. Second row: effects of $\phi_L = .01, 1, 2$. Third row: effects of $\phi_\theta = .001, .005, .02$.

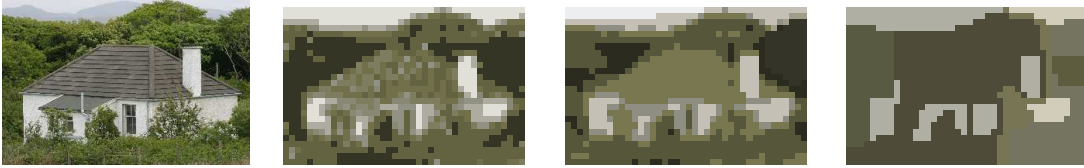


Figure 8: Effects of $\phi_L = 0.5, 0.05, 0.01$ on segmentation for the leftmost image.

in the i -th image. The color intensity consists of three numbers within interval $[0, 255]$ (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 yields 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 $i = 1, \dots, n; r = 1, 2, 3; x \in D$.

As described in Section 4 we introduce additional constraints into the prior structure for the canonical curves θ^* . In particular, we place a (truncated) normal prior with mean $[10 \ 10 \ 10]$ and variance $10^2 I_3$ on



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

θ_1^* , and a normal prior with mean $[240 \ 240 \ 240]$ and the same variance on θ_k^* . That is, we anchor the two extreme labels with the two extremes of the color scale (black and white colors). All remaining canonical species are given a relatively non-informative prior; for $j = 2, \dots, k - 1$, $\theta_j^{*r} \stackrel{iid}{\sim} N(\mu_j, \sigma_j) \cdot \mathbb{I}_{[0,255]}$, where $\mu_j = 128$. For all $j = 1, \dots, k$, we fix $\sigma_j = \sigma_\theta$, where $\sigma_\theta^{-2} \sim \text{Gam}(a_\sigma, b_\sigma)$. We set $a_\sigma = 0.4$ and $b_\sigma = 0.001$. For precision parameter τ we let $\tau^{-2} \sim \text{Gam}(a_\tau, b_\tau)$, 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 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 decompose 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 have experimented with different choices $r = 4, 6, 8$ and received comparable results. Finally, the latent labeling process $\mathbf{q}(\phi_L, k = 8)$ is specified for the whole domain using different choices of $\phi_L = 0.5, 0.1, 0.05, 0.01$.

The MCMC algorithm is run for 200 iterations. Sample obtained from the last 150 iterations are used for image segmentation. The segmented images are obtained by assigning to each image location the light intensity of the MAP estimate of the canonical curve at the same location. Fig. 9 provides examples of representative segmentation results. Fig. 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.

7 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 incorporating prior knowledge of canonical curves.

References

- 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.

- 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., Guindani, 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.
- (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, Series B*, 64, 805–826.
- Gelfand, A., Kottas, A., and MacEachern, S. (2005), “Bayesian nonparametric spatial modeling with Dirichlet process mixing,” *J. Amer. Statist. Assoc.*, 100, 1021–1035.
- 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.
- 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,” *Statistica Sinica*, to appear.
- Petrone, S., Guidani, M., and Gelfand, A. (2008), “Hybrid Dirichlet processes for functional data,” *Under journal revision*.
- Pillai, N., Liang, F., Mukerjee, S., Wolpert, R., and Wu, Q. (2006), “Characterizing the function space for Bayesian kernel models,” Tech. rep., Department 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.
- Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statistica Sinica*, 4, 639–650.
- Stein, M. (1999), *Interpolation of Spatial Data*, New York: Springer-Verlag.
- Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2008), “Describing Visual Scenes Using Transformed Objects and Parts,” *International Journal of Computer Vision*, 77.
- Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), “Hierarchical Dirichlet processes,” *J. Amer. Statist. Assoc.*, 101, 1566–1581.
- 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)*.
- Zhang, T. (2006), “From ϵ -entropy to KL-entropy: Analysis of Minimum Complexity Density Estimation,” *Ann. Statist.*, 34, 2180–2210.

A Proofs

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

Proof of Prop. 2. We derive the result for 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, i.e., $\tilde{\eta} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} [\eta(x_1) \ \eta(x_2)]^T$, 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 $\lambda_2 = \sigma_L^2 - \rho_{12}$. We obtain:

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

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 Prop. 3 We derive the result for 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:

$$c_j = c_{j-1} + \frac{1}{k}(\sqrt{2\pi}\sigma_L e^{c_j^2/2\sigma_L^2} + o(1)) \quad (10)$$

$$c_j = -c_{k-j} \quad (11)$$

$$c_{k/2} = 0 \text{ if } k \text{ is even.} \quad (12)$$

Approximating integrals by a Riemann sum, we have:

$$\begin{aligned} P(Z_1 = i, Z_2 = j) &= P(\eta_1 \in (c_{i-1}, c_i), \eta_2 \in (c_{j-1}, c_j)) \\ &= \frac{1}{2\pi\sqrt{\sigma_L^4 - \rho_{12}^2}} \int_{c_{i-1}}^{c_i} \int_{c_{j-1}}^{c_j} \exp - \frac{\sigma_L^2(\eta_1^2 + \eta_2^2) - 2\rho_{12}\eta_1\eta_2}{2(\sigma_L^4 - \rho_{12}^2)} d\eta_1 d\eta_2 \\ &= \frac{1}{2\pi\sqrt{\sigma_L^4 - \rho_{12}^2}} (c_i - c_{i-1})(c_j - c_{j-1}) \exp - \frac{\sigma_L^2(c_i^2 + c_j^2) - 2\rho_{12}c_i c_j}{2(\sigma_L^4 - \rho_{12}^2)} \\ &\stackrel{(10)}{=} \frac{1}{k^2} \frac{\sigma_L^2}{\sqrt{\sigma_L^4 - \rho_{12}^2}} \exp - \frac{(c_i^2 + c_j^2)\rho_{12}^2 - 2\rho_{12}\sigma_L^2 c_i c_j}{2(\sigma_L^4 - \rho_{12}^2)\sigma_L^2} (1 + o(1)). \end{aligned}$$

Next, by standard properties of multivariate Gaussian distributions, $\eta_2 | \eta_1 = u_1 \sim N(u_1 \rho_{12} / \sigma_L^2, \sigma_L^2 - \rho_{12}^2 / \sigma_L^2)$. As a result,

$$\begin{aligned} P(\eta_1 = i, \eta_2 \geq j) &= \int_{c_{i-1}}^{c_i} \frac{e^{-u_1^2/2\sigma_L^2}}{\sigma_L\sqrt{2\pi}} \left(1 - \Phi\left(\frac{c_j - u_1\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2}\right) \right) du_1 \\ &= (c_i - c_{i-1}) \left(\frac{e^{-c_i^2/2\sigma_L^2}}{\sigma_L\sqrt{2\pi}} \left(1 - \Phi\left(\frac{c_j - c_i\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2}\right) \right) \right) \\ &\stackrel{(10)}{=} \frac{1}{k} \left(1 - \Phi\left(\frac{c_j - c_i\rho_{12}/\sigma_L^2}{\sigma_L^2 - \rho_{12}^2/\sigma_L^2}\right) + o(1) \right). \end{aligned}$$

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

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

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

Extension of Prop. 3 to $m > 2$ locations. Letting $k \rightarrow \infty$, we obtain stochastic process $\mathbf{q}(\sigma_L, \phi_L, \infty)$ generating integer-valued random function $Z : D \rightarrow \mathbb{Z}$. This process converges to the continuous copula process in the following sense: For any collections of locations x_1, \dots, x_m and letting $Z_i = Z(x_i)$:

$$(\sigma_L \Phi^{-1}(Z(x_1)/k), \dots, \sigma_L \Phi^{-1}(Z(x_m)/k)) \Rightarrow N(0, \sigma_L, \phi_L).$$

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

$$P(Z(x_1) = j_1, \dots, Z(x_m) = j_m) = \frac{1}{k^m} (R_{j_1, \dots, j_m}(c_{j_1}, \dots, c_{j_m}) + o(1)), \quad (13)$$

where $o(1) \rightarrow 0$ uniformly for all such tuple (j_1, \dots, j_m) , and

$$R_{j_1, \dots, j_m}(c_{j_1}, \dots, c_{j_m}) = \sigma_L^m (\det A)^{1/2} \exp \sum_{i=1}^m c_{j_i}^2 \left(\frac{1}{2\sigma_L^2} - \frac{A_{ii}}{2} \right) - \sum_{s < t} c_{j_s} c_{j_t} A_{st}. \quad (14)$$

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 $Z(x_2), \dots, Z(x_m)$. Letting \tilde{A} denote the inverse of the covariance matrix for $Z(x_2), \dots, Z(x_m)$, we have:

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

$$P(Z(x_1) > j_1 | Z(x_2) = j_2, \dots, Z(x_m) = j_m) = 1 - \Phi \left(\frac{c_{j_1} - [\rho_{12} \dots \rho_{1m}] \tilde{A} [c_{j_2} \dots c_{j_m}]^T}{\sigma_L^2 - [\rho_{12} \dots \rho_{1m}]^T \tilde{A} [\rho_{12} \dots \rho_{1m}]} \right) + o(1).$$

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

$$\begin{aligned} \mathbb{E}(\theta(x_1) - \theta(x_2))^2 &= \mathbb{E} \sum_{j_1, j_2=1, \dots, k} \mathbf{p}_{x_1, x_2}(j_1, j_2) (\theta_{j_1}^*(x_1) - \theta_{j_2}^*(x_2))^2 \\ &= \sum_{j_1, j_2=1, \dots, k} \mathbb{E} \mathbf{p}_{x_1, x_2}(j_1, j_2) \mathbb{E} (\theta_{j_1}^*(x_1) - \theta_{j_2}^*(x_2))^2 \\ &= \sum_{j_1, j_2=1, \dots, k} \mathbf{q}_{x_1, x_2}(j_1, j_2) \mathbb{E} (\theta_{j_1}^*(x_1) - \theta_{j_2}^*(x_2))^2 \end{aligned}$$

Combining with properties of \mathbf{p} ,

$$\begin{aligned}\mathbb{E}(\theta(x_1) - \theta(x_2))^2 &= \sum_{j_1 \neq j_2} \mathbf{q}_{x_1, x_2}(j_1, j_2) (\mathbb{E}\theta^*(x_1)^2 + \mathbb{E}\theta^*(x_2)^2 - \mathbb{E}\theta^*(x_1)\mathbb{E}\theta^*(x_2)) \\ &\quad + \sum_{j=1}^k \mathbf{q}_{x_1, x_2}(j, j) \mathbb{E}(\theta^*(x_1) - \theta^*(x_2))^2.\end{aligned}$$

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 $j_1 \neq j_2$,

$$\mathbf{q}_{x_1, x_2}(j_1, j_2) \rightarrow 0 \text{ as } x_2 \rightarrow x_1.$$

Note that if F is a Gaussian process $GP(0, \sigma_L, \phi_L)$ and $k = 2$, this probability is available in closed form given by Prop. 2: For $j_1 \neq j_2$,

$$\mathbf{q}_{x_1, x_2}(j_1, j_2) = \frac{1}{\pi} \arccos \sqrt{\frac{1}{2} + \frac{\rho_{12}}{2\sigma_L^2}} \rightarrow 0 \text{ as } x_2 - x_1 \rightarrow 0.$$

More generally, suppose that $j_1 < j_2$. Recall the construction of Z via auxiliary random function $\eta \sim F$. Fix arbitrary $\epsilon > 0$. $P(Z(x_1) = j_1, Z(x_2) = j_2 | \eta(x_1) < c_{j_1} - \epsilon) \leq P(\eta(x_2) - \eta(x_1) > \epsilon) \leq \mathbb{E}(\eta(x_1) - \eta(x_2))^2 / \epsilon^2 \rightarrow 0$ as $x_2 \rightarrow x_1$ due to the mean square continuity of η . Now letting $\epsilon \rightarrow 0$, we obtain $P(\eta(x_1) \in [c_{j_1} - \epsilon, c_{j_1}]) \rightarrow 0$. This yields

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

as $x_2 \rightarrow x_1$.

Proof of Prop. 5. Let $\mathcal{D} = \{\|x_i - x_j\| | 1 \leq i, j \leq m\}$ and

$$p_\phi(d) = \frac{1}{\pi} \arccos \sqrt{1/2 + e^{-\phi d}/2}.$$

Let $n^+(d)$ ($n^-(d)$, resp.) be the number of (x_i, x_j) pairs such that $\|x_i - x_j\| = d$ and $\eta(x_i)\eta(x_j) \geq 0$ ($\eta(x_i)\eta(x_j) < 0$, 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

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

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

$$\sum_{d \in \mathcal{D}} n^+(d) \log p_{\hat{\phi}_L}(d) + n^-(d) \log(1/2 - p_{\hat{\phi}_L}(d)) \geq \sum_{d \in \mathcal{D}} n^+(d) \log p_{\phi_L^*}(d) + n^-(d) \log(1/2 - p_{\phi_L^*}(d)).$$

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

$$\sum_{d \in \mathcal{D}} n^+(d) \log \frac{p_{\hat{\phi}_L}(d) + p_{\phi_L^*}(d)}{2p_{\phi_L^*}(d)} + n^-(d) \log \frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)} \geq 0. \quad (15)$$

It is simple to see that both $\log \frac{p_{\phi}(d) + p_{\phi_L^*}(d)}{2p_{\phi_L^*}(d)}$ and $\log \frac{1 - p_{\phi}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)}$ are absolutely bounded by some constant $M > 0$ for any $\phi \geq 0$. From Prop. 2,

$$\begin{aligned} \mathbb{E}n^+(d) &= 2n(d)p_{\phi_L^*}(d) \\ \mathbb{E}n^-(d) &= 2n(d)(1/2 - p_{\phi_L^*}(d)). \end{aligned}$$

Applying McDiarmid's inequality, for any $\epsilon > 0$ we obtain:

$$\begin{aligned} P \left(\sup_{\phi \geq 0} \left| \sum_{d \in \mathcal{D}} \left(n^+(d) - 2n(d)p_{\phi_L^*}(d) \right) \log \frac{p_{\phi}(d) + p_{\phi_L^*}(d)}{2p_{\phi_L^*}(d)} \right. \right. \\ \left. \left. + \left(n^-(d) - 2n(d)(1/2 - p_{\phi_L^*}(d)) \right) \log \frac{1 - p_{\phi}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)} \right| \geq \epsilon \right) \leq 2 \exp \frac{-4\epsilon^2}{m(m-1)M^2}. \quad (16) \end{aligned}$$

Combining (16) and (15), we obtain:

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

In other words,

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

in \mathbf{q} -probability.

Note that for $\log x \leq 2(\sqrt{x} - 1)$, so

$$\begin{aligned} & p_{\phi_L^*}(d) \log \frac{2p_{\phi_L^*}(d)}{p_{\hat{\phi}_L}(d) + p_{\phi_L^*}(d)} + (1/2 - p_{\phi_L^*}(d)) \log \frac{1 - 2p_{\phi_L^*}(d)}{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)} \\ & \geq -2p_{\phi_L^*}(d) \left(\sqrt{\frac{p_{\phi_L^*}(d) + p_{\hat{\phi}_L}(d)}{2p_{\phi_L^*}(d)}} - 1 \right) - 2(1/2 - p_{\phi_L^*}(d)) \left(\sqrt{\frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{1 - 2p_{\phi_L^*}(d)}} - 1 \right) \\ & = \left(\sqrt{\frac{p_{\phi_L^*}(d) + p_{\hat{\phi}_L}(d)}{2}} - \sqrt{p_{\phi_L^*}(d)} \right)^2 + \left(\sqrt{\frac{1 - p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d)}{2}} - \sqrt{1/2 - p_{\phi_L^*}(d)} \right)^2 \\ & \geq \frac{1}{8} (p_{\hat{\phi}_L}(d) - p_{\phi_L^*}(d))^2. \end{aligned}$$

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 Prop. 6. Let $Z = (z(x_1), \dots, z(x_m))$. 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 $\epsilon_m > 0$:

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

where $C_1 = \int (\pi(\phi_L^*)/\pi(\phi))^\lambda d\phi$ is a constant. Define the following set:

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

By McDiarmid’s inequality, $P_Z(A_m(\epsilon)) \leq 2 \exp \frac{-4\epsilon^2}{m(m-1)M^2}$ for some constant $M > 0$. Applying union bounds, under joint distribution P^* we have, for any $\delta_m > 0$:

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

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 prior π ,

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

Thus we obtain,

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

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

Proof of Lem. 7. (sketch) Using standard calculations for exponential families, for each pair of values $(u, v) \in \{1, \dots, k\}^2$, 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 Prop. 8. (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:

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

so (by induction) it has uniform marginal at each single node corresponding to x_2, \dots, x_m . Applying 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 (and denote it by $A(\theta)$ from now on) via the following relations:

$$\begin{aligned}\theta_{ij} &= \log \frac{\mathbf{q}(Z(x_i) \neq Z(x_j))}{\mathbf{q}(Z(x_i) = Z(x_j))} \text{ for } (i, j) \in E ; 0 \text{ otherwise} \\ \theta_E &= \{(\theta_{ij}) \mid 1 \leq i, j \leq m\} \\ \tilde{\mathbf{q}}_E(Z) &= \exp \left\{ \sum_{(i,j) \in E} \theta_{ij} \mathbb{I}(Z(x_i) \neq Z(x_j)) - B_E(\theta) \right\} \\ B(\theta_E) &= \log \sum_Z \exp \left\{ \sum_{(i,j) \in E} \theta_{ij} \mathbb{I}(Z(x_i) \neq Z(x_j)) \right\} \\ A(\theta_E) &= B(\theta_E) + \sum_{(i,j) \in E} \log \frac{1}{2} \mathbf{q}(Z(x_i) = Z(x_j)).\end{aligned}$$

As a standard fact of exponential families, $B : \mathbb{R}^{m(m-1)/2} \rightarrow \mathbb{R}$ 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 have:

$$\begin{aligned}B(\theta_E) &\geq B(\theta_{E_0}) + (\theta_E - \theta_{E_0}) \nabla_{\theta_E} B(\theta_{E_0}) \\ B(\theta_{E_0}) &\geq B(\theta_E) + (\theta_{E_0} - \theta_E) \nabla_{\theta_E} B(\theta_E),\end{aligned}$$

and, using the above relationships, yields the desired result.