Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 4, *pp.* 755–782

Hybrid Dirichlet mixture models for functional data

Sonia Petrone,

Bocconi University, Milan, Italy

Michele Guindani

University of New Mexico, Albuquerque, USA

and Alan E. Gelfand

Duke University, Durham, USA

[Received July 2007. Final revision January 2009]

Summary. In functional data analysis, curves or surfaces are observed, up to measurement error, at a finite set of locations, for, say, a sample of n individuals. Often, the curves are homogeneous, except perhaps for individual-specific regions that provide heterogeneous behaviour (e.g. 'damaged' areas of irregular shape on an otherwise smooth surface). Motivated by applications with functional data of this nature, we propose a Bayesian mixture model, with the aim of dimension reduction, by representing the sample of n curves through a smaller set of canonical curves. We propose a novel prior on the space of probability measures for a random curve which extends the popular Dirichlet priors by allowing local clustering: non-homogeneous portions of a curve can be allocated to different clusters and the n individual curves can be represented as recombinations (hybrids) of a few canonical curves. More precisely, the prior proposed envisions a conceptual hidden factor with k-levels that acts locally on each curve. We discuss several models incorporating this prior and illustrate its performance with simulated and real data sets. We examine theoretical properties of the proposed finite hybrid Dirichlet mixtures, specifically, their behaviour as the number of the mixture components goes to ∞ and their connection with Dirichlet process mixtures.

Keywords: Bayesian non-parametrics; Dependent random partitions; Dirichlet process; Finite mixture models; Gaussian process; Labelling measures; Species sampling priors

1. Introduction

Functional data analysis is receiving increased interest in the scientific community, documented by a rapidly growing literature; Ramsay and Silverman (2005) and Ferraty and Vieu (2006) offer recent comprehensive references. In such analysis, curves or surfaces are observed, up to measurement error, for a sample of n individuals, i.e. $y_i(x) = \theta_i(x) + \varepsilon_i(x)$, $i = 1, 2, \ldots, n$, with $\varepsilon_i(x) \sim^{\mathrm{IID}} N(0, \sigma^2)$, $x \in D \subset \mathbb{R}^p$. Here, $\theta(\cdot)$ denotes the curve or surface and we focus on $p = 1$, e.g. x is time, or $p = 2$, e.g. x is a geographic co-ordinate. We assume that D spans a continuum in R^p, though everything that we discuss works also when D is finite or countable, e.g. for a lattice or a collection of areal units.

In this paper, we develop Bayesian non-parametric (NP) inference for estimating $\boldsymbol{\theta}_i = (\theta_i(x), x \in D)$ by borrowing strength from the other curves. Furthermore, we achieve dimension reduction

Address for correspondence: Sonia Petrone, Istituto di Metodi Quantitativi, Università Bocconi, Viale Isonzo 25, 20135 Milan, Italy.
E-mail: sonia.petrone@uni-bocconi.it

 1369–7412/09/71755

by representing the n observed curves by means of a smaller set of *canonical curves*, which we shall also refer to as 'curve species'.

A popular approach for dimension reduction in functional data analysis is functional principal components analysis (see for example Ramsay and Silverman (2005)). Clustering techniques based on mixture models are also receiving growing interest for the analysis of high dimensional data. In particular, the availability of efficient computational algorithms has led to a substantial development of Bayesian parametric and NP dimension reduction techniques through finite Dirichlet mixtures (i.e. finite mixture models with Dirichlet-distributed weights) or infinite components Dirichlet process (DP) mixtures. Mixtures of Gaussian kernels are widely used for modelling the distribution of multivariate data. In our context, the multivariate data that we model are the values of a curve at an arbitrary finite set of co-ordinates, say $Y_i = (Y_i(x_1), \ldots, Y_i(x_m))$, $i = 1, \ldots, n$. The kernel centroids are intepretable as values of 'canonical curves' at $x_1, \ldots, x_m$. In general, the goal is to use a number of canonical curves which is far fewer than n to describe the observed sample.

Bayesian hierarchical models and DP mixtures have been successfully exploited to represent the individual curves by means of an orthonormal basis expansion and to cluster the expansion coefficients (see Bigelow and Dunson (2009) and Ray and Mallick (2006)). In those approaches, the kernel centroids are defined by 'canonical vectors' of expansion coefficients. In computer modelling (e.g. Oakley and O'Hagan (2002)) and machine learning (Neal, 1997; Rasmussen and Williams, 2006) Gaussian process realizations are often used as a basis to model random functions. Finite mixtures and DP mixtures of Gaussian processes have been proposed also to model a sample of curves directly as in Shi and Wang (2008) or Gelfand *et al.* (2005) for spatial data. Here, the kernel centroids are modelled as independent and identically distributed (IID) realizations of a stationary Gaussian process $\mathcal{GP}$. In the simplest case, we may consider constant canonical curves; though restrictive, such a choice facilitates species identifiability. More generally, the shape of the canonical curves is regulated by the parameters in the mean and correlation function of the Gaussian process.

Bayesian finite and DP mixture models for functional data analysis usually propose to fit smooth curves, assuming *global* heterogeneity across individuals. However, in many applications, the individual curves may be quite smooth and similar except for some *local* heterogeneity. For example, for D countable or finite in $\mathbb{R}$, individual sequences $\boldsymbol{\theta}_i$ (e.g. DNA sequences in genetic studies) may show local mutations across individuals, at a few locations (genes). A two-dimensional example which we pursue further later involves a sample of magnetic resonance imaging brain images where grey matter level intensity is measured at a set of locations. Here, an otherwise smooth (healthy) image may show a few diseased regions of irregular shape. A finite Dirichlet mixture model represents the individual curves as globally selected from a population of canonical curves, or species, on D. Therefore, the model identifies a new species even when a curve is substantially different from the others only on a few portions of D. In other words, the model tends either to fit an 'average curve' or, if the number of mixture components is large (or infinite), to increase the number of canonical curves that are required for reconstructing the sample, thus missing the desired dimensional reduction goal.

In this work, we offer a more parsimonious mixture model to account for global and local heterogeneity, where the individual curves are represented as *recombinations* of the set of canonical curves. To be more specific, we propose a mixture model where the prior on the mixing distribution is an extension of the popular finite Dirichlet and DP priors. In fact, we suggest a general class of priors for NP Bayesian inference that extend the *global* allocation rules of the latter and allow for *dependent local* allocation. As an effect of the local allocation scheme, the curves will be described as *hybrid* species, obtained by recombining portions

of canonical curves; therefore we refer to the proposed class of priors as hybrid Dirichlet priors.

Our construction has an interpretation in terms of hidden labels. For any θ_i, we can assign a label $\gamma_i(x)$ to indicate the species that is chosen at x, i.e. we model a hidden label process with k-levels, acting *locally* on each individual curve, so that local species allocation is naturally induced. Moreover, the dependence structure in the labelling process controls the degree of species recombination in the sample. For example, mixtures of widely used hidden Markov models can be easily framed into our scheme. However, a Markov dependence might be considered too restrictive for general functional data. Therefore, we propose to model functional dependence in the label process through an auxiliary Gaussian copula. In addition to being more suitable for functional data analysis, such a choice facilitates prior elicitation and proves to be computationally attractive in an NP Bayes framework.

Local clustering has been addressed in the NP Bayes literature; however, most proposals are limited to modelling partially exchangeable data (see for example Teh *et al.* (2006) and references therein). In our context, although the observations can be separated into different groups corresponding to each x_j, the assumption of partial exchangeability is clearly too restrictive, since it implies conditional independence along x. In the more general framework of dependent DPs (MacEachern, 1999, 2001), most applications have made use of the so-called single-p dependent DP (e.g. Gelfand *et al.* (2005)). However, these models reduce to a DP on the joint distribution and therefore only allow global clustering. Recent proposals of multiple-p stick breaking priors (see for example Griffin and Steel (2006), Dunson and Park (2008) and Dunson *et al.* (2008)) allow for a more general dependence structure across the groups' random marginal distributions. However, the resulting functional dependence in the data, and consequently the shape of the curve realizations, have not been fully explored. Closer to our approach, Duan *et al.* (2007) defined a multivariate stick breaking construction for point-referenced spatial data. However, although they also referred to a hidden label process for local surface selection, their construction involves an infinite number of hidden variables at each location x, requiring a latent Gaussian process for each of the countable number of stick breaks. Hence, the label process is obscured as well as the chance for local and global clustering. Moreover, computations are cumbersome. MacEachern (2007) offered a somewhat simpler version that has been recently extended in Rodriguez *et al.* (2008).

Our proposal differs from those above in several aspects. Rather than specifying (stick breaking) priors for the marginal distributions of the $\theta(x_j)$, we focus on the prior on the *joint* distribution of the vector $(\theta(x_1), \ldots, \theta(x_m))$, and, ultimately, of the process $\boldsymbol{\theta} = (\theta(x), x \in D)$. Such a shift of interest proves to be crucial for controlling the functional dependence. In fact, we offer a unifying framework, based on species sampling models (Pitman, 1996), which includes several recent proposals as a special case. In this framework, we obtain hybrid DP mixtures as limits of finite mixtures as the number of components goes to ∞. This approach sheds light also on the relationship between finite and DP mixtures in the context of hidden variables models, extending results that were developed by Teh *et al.* (2006) for hidden Markov models. Although still challenging, computations in our model are simpler than in previous proposals. We suggest a fairly straightforward but effective Markov chain Monte Carlo (MCMC) algorithm, where the monitoring of labelling by site and by individual for each curve is facilitated by the introduction of a small amount of pure Gaussian error. Finally, we illustrate the performance of our model with a simulated data example as well as with the forementioned brain imaging data. The expected hybridization emerges, revealing the benefit of the modelling that we have introduced.

The format of the paper is as follows. In Section 2 we introduce the basic mixture of Gaussian

processes model, with a finite functional Dirichlet prior or a functional DP. Section 3 formalizes the notion of hybridization. In particular, Section 3.3 provides a careful examination of the weak limits of finite hybrid Dirichlet priors. Section 4 brings these processes to the functional data application of interest, including a discussion of the suggested Gaussian copula labelling prior and computational issues. Section 5 presents the results of both a simulated data analysis as well as analysis of the motivating brain images data. We discuss some final remarks and possible extensions in Section 6. Full details of our technical results as well as a detailed description of the MCMC algorithm used in Section 5 are provided in an appendix which is available online at `http://mypage.unibocconi.it/soniapetrone`.

The data and the program that was used to analyse them can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Mixture models for functional data

Let $\mathbf{Y}_i = (Y_i(x), x \in D)$, $i = 1, 2, \ldots, n$, be random curves defined on a (regular) domain $D \subseteq \mathbb{R}^p$. We have the formal model

$$\mathbf{Y}_i = \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n,$$

where the $\boldsymbol{\varepsilon}_i$ are independent realizations of a Gaussian white noise process with variance σ^2, denoted as $\mathcal{GP}(\mathbf{0}, \sigma^2)$. Equivalently, given the $\boldsymbol{\theta}_i$,

$$\mathbf{Y}_i | \boldsymbol{\theta}_i \overset{\text{ind}}{\sim} \mathcal{GP}(\boldsymbol{\theta}_i, \sigma^2). \tag{1}$$

All of the effort is in modelling the mean functions $\boldsymbol{\theta}_i$ which are usually specified as independent realizations of a Gaussian process. Instead, we allow borrowing of strength in the estimation by introducing probabilistic dependence across the $\boldsymbol{\theta}_i$s. We assume that they are sampled from a common (unknown) probability measure $\mathbf{G}$ on $\mathbb{R}^D$, i.e. $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n | \mathbf{G} \sim^{\text{IID}} \mathbf{G}$. Together with distribution (1), this gives a mixture of Gaussian processes model, that we write heuristically as

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n | \mathbf{G} \overset{\text{IID}}{\sim} \int \mathcal{GP}(\cdot | \boldsymbol{\theta}, \sigma^2) \, \mathrm{d}\mathbf{G}(\boldsymbol{\theta}). \tag{2}$$

For a random curve $\boldsymbol{\theta} = (\theta(x), x \in D)$ in $\mathbb{R}^D$, with probability measure $\mathbf{G}$, we shall denote by $G_{x_1,\ldots,x_m}$ the finite dimensional distribution of $(\theta(x_1), \ldots, \theta(x_m))$ at co-ordinates $(x_1, \ldots, x_m)$ (we use the same symbol for a probability measure on $\mathbb{R}^m$ and the corresponding distribution function (DF)). For convenience, the random curves are assumed to be observed at a common finite set of co-ordinates, say $(x_1, \ldots, x_m)$, so that the available data are $Y_i = (Y_i(x_1), \ldots, Y_i(x_m))$, $i = 1, \ldots, n$, with $Y_i(x_j) = \theta_i(x_j) + \varepsilon_i(x_j)$. (The case where the Y_is are observed at different sets of co-ordinates can be handled by augmenting the dimension of the θ_is to a common joint set of co-ordinates.) Let $\theta_i = (\theta_i(x_1), \ldots, \theta_i(x_m))$. Then, the finite dimensional characterization of model (2) is

$$\begin{aligned} Y_i | \theta_i &\overset{\text{ind}}{\sim} \mathcal{N}_m(\theta_i, \sigma^2 I_m), \\ \theta_i | G_{x_1,\ldots,x_m} &\overset{\text{IID}}{\sim} G_{x_1,\ldots,x_m}, \end{aligned} \tag{3}$$

where $\mathcal{N}_m(\cdot, \cdot)$ denotes the m-variate Gaussian distribution and I_m is the m-dimensional identity matrix. Integrating the θ_is out, we have

$$Y_i | G_{x_1,\ldots,x_m} \overset{\text{IID}}{\sim} \int \mathcal{N}_m(\theta, \sigma^2 I_m) \, \mathrm{d}G_{x_1,\ldots,x_m}(\theta), \tag{4}$$

i.e. a location mixture of Gaussian densities. In a Bayesian NP approach, a prior probability law is assigned to the random mixing probability measure $\mathbf{G}$. This requires assigning a prior, consistently, to its finite dimensional distributions $G_{x_1,\ldots,x_m}$, for any choice of m and $(x_1,\ldots,x_m)$. Let the prior on $\mathbf{G}$ almost surely select discrete probability measures of the form

$$\mathbf{G} = \sum_{j=1}^{k} p_j \delta_{\boldsymbol{\theta}_j^*}, \tag{5}$$

where k is a finite integer and $\delta_{\boldsymbol{\theta}}$ denotes a probability measure that is degenerate on $\boldsymbol{\theta}$, i.e. $\mathbf{G}$ concentrates probability masses p_j on curve atoms $\boldsymbol{\theta}_j^*$ in $\mathbb{R}^D$. Moreover, let $(p_1,\ldots,p_k) \sim \mathcal{D}(\alpha_1,\ldots,\alpha_k)$, a Dirichlet distribution with parameters $(\alpha_1,\ldots,\alpha_k)$, and $\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_k^*$ be IID according to a non-atomic probability measure $\mathbf{G}_0$ on $\mathbb{R}^D$, independently of the p_js. Correspondingly,

$$G_{x_1,\ldots,x_m} = \sum_{j=1}^{k} p_j \delta_{\theta_j^*}, \tag{6}$$

where the $\theta_j^* = (\theta_j^*(x_1),\ldots,\theta_j^*(x_m))$ are IID according to $G_{0,x_1,\ldots,x_m}$. Then, model (2) reduces to a finite mixture of Gaussian processes, with finite dimensionals

$$Y_i | G_{x_1,\ldots,x_m} \overset{\text{IID}}{\sim} \sum_{j=1}^{k} p_j \mathcal{N}_m(\theta_j^*, \sigma^2 I_m). \tag{7}$$

In the Bayesian literature, finite mixture models such as model (7) are usually regarded as indexed by the parameters $\{(p_1,\ldots,p_k),(\theta_1^*,\ldots,\theta_k^*)\}$. Here, instead, the model is parameterized in terms of the mixing distribution. As a prior on the space of DFs on $\mathbb{R}^m$, the probability measure that almost surely selects distributions of the form (6) is usually called a (finite dimensional) Dirichlet prior (Ishwaran and Zarepour, 2002). We write $G_{x_1,\ldots,x_m} \sim \mathrm{DP}_k\{(\alpha_1,\ldots,\alpha_k), G_{0,x_1,\ldots,x_m}\}$. Usually, a 'non-informative' symmetric Dirichlet distribution is used for the mixing weights, $(p_1,\ldots,p_k) \sim \mathcal{D}(\alpha/k,\ldots,\alpha/k)$, $0 < \alpha < \infty$, and we say that $G_{x_1,\ldots,x_m}$ has a symmetric Dirichlet prior, which is denoted by $G_{x_1,\ldots,x_m} \sim \mathrm{DP}_k(\alpha, G_{0,x_1,\ldots,x_m})$. Extending to the functional case, we say that the random probability measure (RPM) on $\mathbb{R}^D$ defined by equation (5) has a *functional* DP_k prior, $\mathbf{G} \sim \mathrm{fDP}_k\{(\alpha_1,\ldots,\alpha_k), \mathbf{G}_0\}$, or $\mathbf{G} \sim \mathrm{fDP}_k(\alpha, \mathbf{G}_0)$ for the symmetric case. Clearly, the finite dimensional distributions of $\mathbf{G}$ are DP_k, with base measures $G_{0,x_1,\ldots,x_m}$ driven by $\mathbf{G}_0$, for any choice of m and $x_1,\ldots,x_m$.

When the number of components k is uncertain, rather than treating it as random, it has become common practice to let $k = \infty$ and to use DP mixtures. Suppose that $\mathbf{G}$ is as in equation (5), but let now $k = \infty$ and $(p_1, p_2, \ldots)$ have a stick breaking prior with parameter α, i.e. $p_1 = V_1$,

$$p_j = V_j \prod_{i=1}^{j-1} (1 - V_i), \qquad V_i \overset{\text{IID}}{\sim} \mathrm{beta}(1, \alpha).$$

Then, $\mathbf{G}$ has a DP prior, with parameters α and $\mathbf{G}_0$ (Sethuraman, 1994). In fact, we say that $\mathbf{G}$ has a *functional* DP prior, $\mathbf{G} \sim \mathrm{fDP}(\alpha, \mathbf{G}_0)$. It is easy to see that the finite dimensional distributions $G_{x_1,\ldots,x_m}$ are DPs, $G_{x_1,\ldots,x_m} \sim \mathrm{DP}(\alpha, G_{0,x_1,\ldots,x_m})$, and they are dependent since their base measures are all driven by $\mathbf{G}_0$.

The parameterization of the model in terms of the mixing distribution is useful to clarify the relationship between finite and DP mixtures. It can be shown that, if $(G_k, k \geqslant 1)$ is a sequence of random DFs with $G_k \sim \mathrm{DP}_k(\alpha, G_0)$, then G_k converges in distribution to $G \sim \mathrm{DP}(\alpha G_0)$ as $k \to \infty$ (Muliere and Secchi (1995) and Ishwaran and Zarepour (2002), theorem 3).

Both the DP_k and the DP are special cases of the more general family of (proper) *species sampling priors* (Pitman, 1996). To be more specific, a proper species sampling prior for an RPM

G on a measurable space $(\Theta, \sigma(\Theta))$ almost surely selects distributions of the form (5), where, more generally, $k \leqslant \infty$, $p_j \geqslant 0$, $\Sigma_{j=1}^{k} p_j = 1$ and the θ_j^*s are IID from a non-atomic probability measure G_0 on Θ, independently of the p_js. The DP_k and the DP correspond to specific choices of the distribution on the random weights. In this context, the atoms θ_j^* are referred to as 'species' and **G** describes a population comprised of $k \leqslant \infty$ species, populated in proportions that are determined by $p_1, \ldots, p_k$.

In statistical applications to mixture modelling, species sampling priors are of particular interest since ties can be obtained with positive probability in sampling from the mixing distribution. Therefore, the number of species that are 'discovered' in the sample is usually less than $\min(k, n)$, thus revealing that the data $Y_1, \ldots, Y_n$ can in fact be described by means of a mixture with a few components. This type of dimension reduction is often referred to as *clustering* in the Bayesian NP literature. For example, if $\theta_1 = \ldots = \theta_{n_1} \neq \theta_{n_1+1} = \ldots = \theta_n$, then $\theta_1, \ldots, \theta_{n_1}$ are said to be in the same cluster, whereas $\theta_{n_1+1}, \ldots, \theta_n$ belong to a second cluster, and only two mixture components are needed to describe the distribution of the data $Y_1, \ldots, Y_n$. However, clustering the θ_is does not correspond, in general, to a classification procedure. In a location mixture of Gaussians like model (7), the components play the role of kernels, usually with no physical interpretation. So, clustering the θ_is amounts to describing the data through a smaller dimensional set of 'kernel centres' (canonical curves in the functional version (2)). A classification mixture model would require instead more flexible group-specific distributional assumptions, e.g. a scale–location mixture of Gaussians, or an NP model for each component (Ishwaran and James, 2003).

Species sampling priors are characterized by the predictive rules that allocate the θ_is into different species (Pitman, 1996). In particular, for the symmetric $\mathrm{DP}_k(\alpha, G_0)$, we have $\theta_1 \sim G_0$ and

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{\alpha(k - d_n)/k}{\alpha + n} G_0 + \sum_{j=1}^{d_n} \frac{n_j + \alpha/k}{\alpha + n} \delta_{\theta_{(j)}^*}, \tag{8}$$

for $k \leqslant n$, where $\theta_{(1)}^*, \ldots, \theta_{(d_n)}^*$ are the distinct values among $\theta_1, \ldots, \theta_n$, in the order as they appear. For $k \to \infty$, the predictive distribution (8) converges to

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \sum_{j=1}^{d_n} \frac{n_j}{\alpha + n} \delta_{\theta_{(j)}^*}, \tag{9}$$

i.e. the predictive rule characterizing the $\mathrm{DP}(\alpha, G_0)$.

3. Hybrid Dirichlet priors

A limitation of DP_k or DP mixtures for multivariate data is that they can only model *global* clustering, or global mutations. As is evident from distributions (8) and (9), θ_{n+1} is either one of the previously observed species or a completely new species. In applications to multivariate or functional data (as in Section 1), it may be desirable, instead, to allow also for *local* mutations or local clustering. In Section 5, we show that DP_k or DP mixtures often generate as many species as the sample size. The model succeeds in fitting the data well but misses the aim of dimension reduction to a smaller number of canonical curves. We enable the possibility of *hybrid species*, where a curve may be characterized by different species at different co-ordinates.

3.1. Hybrid Dirichlet priors for random probability measures on $\mathbb{R}^m$

Consider the finite dimensional distributions $G_{x_1,\ldots,x_m}$ which, in this subsection, we denote

simply by G. We extend the idea of species sampling priors and imagine that G is the result of a process of global and local mutations. To be more specific, we start from a base population of $k \leqslant \infty$ species $\theta^*_j = (\theta^*_{j,1}, \ldots, \theta^*_{j,m})$, $j = 1, \ldots, k$, IID according to a non-atomic distribution G_0 on $\mathbb{R}^m$. By the effect of the local mutations, hybrid species $(\theta^*_{j_1,1}, \ldots, \theta^*_{j_m,m})$ emerge, where the first component is from species j_1, the second from species $j_2, \ldots$, and so on. We say that a random DF G on $\mathbb{R}^m$ has a (proper) *hybrid species sampling prior* if, almost surely,

$$G = \sum_{j_1=1}^{k} \ldots \sum_{j_m=1}^{k} p(j_1, \ldots, j_m) \delta_{\theta^*_{j_1,1}, \ldots, \theta^*_{j_m,m}}, \tag{10}$$

where $p(j_1, \ldots, j_m)$ represents the proportion of (hybrid) species $(\theta^*_{j_1,1}, \ldots, \theta^*_{j_m,m})$ in the population, $p(j_1, \ldots, j_m) \geqslant 0$, $\Sigma_{j_1=1}^{k} \ldots \Sigma_{j_m=1}^{k} p(j_1, \ldots, j_m) = 1$, and $\theta^*_j \sim^{\text{IID}} G_0$, independently of the $p(j_1, \ldots, j_m)$s.

If $p(j_1, \ldots, j_m) > 0$ only when $j_1 = \ldots = j_m$, we are back to the usual definition of species sampling priors. Our extension can be interpreted as a model that allows for *local effects* of a hidden factor with k levels. Indeed, the weights $(p_1, \ldots, p_k)$ in equation (5) define a random probability mass function, say p, on $\{1, \ldots, k\}$. We can interpret p as the distribution inducing labels drawn from $\{1, \ldots, k\}$, i.e., if $\theta | G \sim G$, and G has a proper species sampling prior, then $\theta = \theta^*_j$ if the label $\gamma = j$. The label's distribution is modelled as $\Pr(\gamma = j | p, \theta^*_1, \ldots, \theta^*_k) = p_j$, $j = 1, \ldots, k$, with a prior on p. The latter could be a Dirichlet distribution in the case of a DP_k prior or a stick breaking prior for the DP (with $k = \infty$).

More generally, the weights $p(j_1, \ldots, j_m)$ in equation (10) define a random probability mass function on $\{1, \ldots, k\}^m$, which we still denote by p. Now, p can be interpreted as the distribution of a random *vector* of labels $\gamma = (\gamma_1, \ldots, \gamma_m)$, with $\gamma_l \in \{1, \ldots, k\}$, $l = 1, \ldots, m$. If $\theta = (\theta_1, \ldots, \theta_m) | G \sim G$, G as in equation (10), then $\theta(x_l) = \theta^*_j(x_l)$ if $\gamma_l = j$. Conditionally on p and the θ^*_js, $\gamma \sim p$, i.e. $\Pr(\gamma_1 = j_1, \ldots, \gamma_m = j_m | p, \theta^*_1, \ldots, \theta^*_k) = p(j_1, \ldots, j_m)$.

In this general framework, the DP_k or the DP prior are extended by appropriately specifying the labelling prior on p. Some recent proposals for NP priors can be regarded as special cases of equation (10). For example, the generalized spatial DP by Duan *et al.* (2007) takes $k = \infty$ and defines a multivariate stick breaking labelling prior. The prior that we propose here is, instead, a natural extension of the finite DP_k prior.

3.1.1. Hybrid finite Dirichlet priors

Let G be as in equation (10), with $k < \infty$. Note that the ith marginal of G is

$$G_i = \sum_{j=1}^{k} p_i(j) \delta_{\theta^*_{j,i}}$$

where $p_i(\cdot)$ is the ith marginal of the probability measure p, and $\theta^*_{1,i}, \ldots, \theta^*_{k,i} \sim^{\text{IID}} G_{0,i}$, where $G_{0,i}$ is the ith marginal of G_0. It is natural to require that an extension of the DP_k for a random DF on $\mathbb{R}^m$ still has DP_k marginals, which is true if $(p_i(1), \ldots, p_i(k)) \sim \mathcal{D}(\alpha_{i,1}, \ldots, \alpha_{i,k})$ for all $i = 1, \ldots, m$. In particular, G_i has a symmetric DP_k prior if $\alpha_{i,j} = \alpha/k$ for every j. It seems natural also to assume that the random weights in equation (10) have a joint Dirichlet distribution, centred on a probability measure q on $\{1, \ldots, k\}^m$:

$$\Pr(\gamma_1 = j_1, \ldots, \gamma_m = j_m) = E\{p(j_1, \ldots, j_m)\} = q(j_1, \ldots, j_m).$$

We use the notation $p \sim \mathcal{D}(\alpha q)$, with q denoting both the probability measure and the family of weights $\{q(j_1, \ldots, j_m), j_i = 1, \ldots, k; i = 1, \ldots, m\}$ and $\alpha > 0$. Under these assumptions, we say that G defined by equation (10) has a *hybrid* finite Dirichlet prior with parameters α, q and G_0,

$G \sim \text{hDP}_k(\alpha q, G_0)$. The marginals G_i have a symmetric DP_k prior if q has uniform marginals, i.e. $q_i(j) = 1/k$, $j = 1, \ldots, k$. The choice of the distribution q is discussed in Section 4.1.

3.1.2. *Hybrid Dirichlet process*

Let G be as in equation (10), but now let $k = \infty$. Then, the weights $p(j_1, \ldots, j_m)$ define an RPM p on $\{1, 2, \ldots\}^m$. Let us assume that p has a DP prior with base probability measure q, i.e. $p \sim \text{DP}(\alpha q)$, so that $E\{p(A)\} = q(A)$, for any $A \subset \{1, 2, \ldots\}^m$. Then, we say that G has a *hybrid* DP prior with parameters α, q and G_0, and write $G \sim \text{hDP}(\alpha q, G_0)$. We show in Section 3.3 that the hybrid DP arises as the weak limit of an hDP_k under appropriate conditions on the labelling prior.

3.2. *Functional hybrid Dirichlet priors*

Now, we extend the discussion of hybrid species sampling priors to the functional case, i.e. we consider an RPM $\mathbf{G}$ on $\mathbb{R}^D$, where D is a countable or continuous subset of $\mathbb{R}^p$. Roughly speaking, we say that $\mathbf{G}$ has a *functional* hybrid species sampling prior if its random finite dimensional distributions $G_{x_1,\ldots,x_m}$ have a hybrid species sampling prior as defined by equation (10), with consistent parameters. More precisely, Kolmogorov consistency must be ensured almost surely for the *family* of *random* DFs $\{G_{x_1,\ldots,x_m}, m \geqslant 1, x_1, \ldots, x_m\}$. A less developed discussion of these aspects is provided in Duan *et al.* (2007). Let $\mathcal{M}(\mathbb{R}^D)$ be the space of probability measures on $\mathbb{R}^D$. The law of an RPM on $\mathbb{R}^D$ is often defined by specifying how it selects an element in $\mathcal{M}(\mathbb{R}^D)$; that is how the functional DP_k and DP were defined in Section 2.

Alternatively, one may specify how the prior selects a family $\mathcal{G}_{\mathbf{G}}$ of consistent finite dimensional distributions, since they characterize an element in $\mathcal{M}(\mathbb{R}^D)$. Let $\mathbf{G}_0$ be a non-atomic probability measure on $\mathbb{R}^D$ (e.g. a Gaussian process) and $\mathbf{p}$ an RPM on $\{1, 2, \ldots, k\}^D$, with $k \leqslant \infty$. As usual, let $G_{0,x_1,\ldots,x_m}$ and $p_{x_1,\ldots,x_m}$ be the finite dimensional distributions of $\mathbf{G}_0$ and $\mathbf{p}$ respectively. A *functional hybrid species sampling prior* for an RPM on $\mathbb{R}^D$ selects a family $\mathcal{G}_{\mathbf{G}}$ by first choosing a sample of curves $\boldsymbol{\theta}^*_j = \{\theta^*_j(x), x \in D\}$, $j = 1, \ldots, k$, IID from $\mathbf{G}_0$ and, independently, a realization p of the RPM $\mathbf{p}$. Then, the family of finite dimensional distributions is defined as

$$G_{x_1,\ldots,x_m} = \sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} p_{x_1,\ldots,x_m}(j_1, \ldots, j_m)\, \delta_{\theta^*_{j_1}(x_1),\ldots,\theta^*_{j_m}(x_m)}, \tag{11}$$

for all $m \geqslant 1, x_1, \ldots, x_m \in D$. The difference from the previous definition (10) is that here the prior selects a *family* $\mathcal{G}_{\mathbf{G}} = \{G_{x_1,\ldots,x_m}, m \geqslant 1, x_1, \ldots, x_m \in D\}$. Kolmogorov consistency is required by construction for both the families $\{p_{x_1,\ldots,x_m}, m \geqslant 1, x_1, \ldots, x_m \in D\}$ and $\{G_{0,x_1,\ldots,x_m}, m \geqslant 1, x_1, \ldots, x_m \in D\}$ that are used in equation (11). It is easy to prove that this ensures that the family $\mathcal{G}_{\mathbf{G}}$ is consistent, defining a probability measure $\mathbf{G}$ on $\mathbb{R}^D$. Extending the discussion below equation (10), we may regard the probability measure $\mathbf{p}$ driving the weights $p_{x_1,\ldots,x_m}$ in equation (11) as the probability law of a stochastic *process* of labels $\boldsymbol{\gamma} = (\gamma(x), x \in D)$ with each $\gamma(x) \in \{1, \ldots, k\}$. If $\boldsymbol{\theta}$ is a random curve in $\mathbb{R}^D$, with $\boldsymbol{\theta}|\mathbf{G} \sim \mathbf{G}$, then $\theta(x) = \theta^*_j(x)$ if $\gamma(x) = j$, for $x \in D$, and $\boldsymbol{\gamma}|\mathbf{p}, \theta^*_1, \ldots, \theta^*_k \sim \mathbf{p}$. Different specifications for the labelling prior $\mathbf{p}$ lead to different classes of priors on $\mathbf{G}$. Below, we give functional versions of the hybrid DP_k and DP.

3.2.1. *Functional hybrid Dirichlet priors*

For $k < \infty$, we can naturally extend the hDP_k to the functional case assuming that the weights $p_{x_1,\ldots,x_m}$ in equation (11) have a joint Dirichlet distribution. However, the parameters of the

Dirichlet prior must be chosen consistently across varying choices of $(x_1, \ldots, x_m)$. Hence, let $p_{x_1,\ldots,x_m} \sim \mathcal{D}(\alpha q_{x_1,\ldots,x_m})$, where the family $\{q_{x_1,\ldots,x_m}, m \geqslant 1, x_1, \ldots, x_m \in D\}$ defines a probability measure $\mathbf{q}$ on $\{1, \ldots, k\}^D$. From the properties of a DP, this is equivalent to $\mathbf{p} \sim \mathrm{DP}(\alpha\mathbf{q})$. Then we say that the RPM $\mathbf{G}$ in equation (11) has a *functional* hybrid Dirichlet prior with parameters α, $\mathbf{q}$ and $\mathbf{G}_0$, $\mathbf{G} \sim \mathrm{fhDP}_k(\alpha\mathbf{q}, \mathbf{G}_0)$.

The labelling prior is centred on the probability measure $\mathbf{q}$, $\Pr(\gamma \in A) = E\{\mathbf{p}(A)\} = \mathbf{q}(A)$, for $A \subset \{1, \ldots, k\}^D$, with finite dimensional distributions

$$\Pr\{\gamma(x_1) = j_1, \ldots, \gamma(x_m) = j_m\} = E\{p_{x_1,\ldots,x_m}(j_1, \ldots, j_m)\} = q_{x_1,\ldots,x_m}(j_1, \ldots, j_m).$$

In most cases, a continuity property is desirable for the label process; e.g. $\gamma(x') \to^{\mathrm{d}} \gamma(x)$ if $x' \to x$, which is true, for example, if $q_{x,x'}(i, i)$ converges to $q_x(i)$ for all $i \in \{1, \ldots, k\}$ and $q_{x,x'}(i, j)$ converges to 0 for $i \neq j$, as $x' \to x$.

3.2.2. Functional hybrid Dirichlet process

A functional hybrid DP is introduced similarly, for $k = \infty$, $\mathbf{q}$ a probability measure on $\{1, 2, \ldots\}^D$ and $\mathbf{p} \sim \mathrm{fDP}(\alpha\mathbf{q})$. We write $\mathbf{G} \sim \mathrm{fhDP}(\alpha\mathbf{q}, \mathbf{G}_0)$. The finite dimensional distributions of $\mathbf{G}$ are as in equation (11), where $(\theta_j^*(x_1), \ldots, \theta_j^*(x_m)) \sim^{\mathrm{IID}} G_{0,x_1,\ldots,x_m}$ and $p_{x_1,\ldots,x_m}$ is the random finite dimensional distribution of $\mathbf{p}$ on $\{1, 2, \ldots\}^m$. Note that $p_{x_1,\ldots,x_m} \sim \mathrm{DP}(\alpha q_{x_1,\ldots,x_m})$; therefore, $G_{x_1,\ldots,x_m} \sim \mathrm{hDP}(\alpha q_{x_1,\ldots,x_m}, G_{0,x_1,\ldots,x_m})$. In particular, $G_x \sim \mathrm{hDP}(\alpha q_x, G_{0,x})$.

3.3. Weak limits of hybrid Dirichlet priors

As discussed in Section 2, the DP_k may be viewed as a finite approximation of the DP. In mixture models, that property is most relevant for studying the sensitivity of finite mixtures and their relationship with DP mixtures as k increases. It has been successfully exploited for MCMC computations (Ishwaran and James, 2001) and Teh *et al.* (2006) have extended those results for hidden Markov models. In this section, we give further results for more general hidden (labelling) measures, by examining the limit behaviour of a hybrid DP_k prior. We confine our study to the case of hybrid Dirichlet priors on random distributions on $\mathbb{R}^m$, although extensions to the functional case can be envisioned. For notational simplicity, here we drop the dependence on $(x_1, \ldots, x_m)$ but we make explicit the dependence on k, denoting a random DF on $\mathbb{R}^m$ simply by G_k. Since hybrid Dirichlet priors do not have IID support points because of the recombination of the θ_j^*s, it is not immediate to generalize the well-known results for the limit of a DP_k to this setting. We show that the limiting behaviour of the hybrid Dirichlet priors depends crucially on the sequence of the labelling measures q_k, $k \geqslant 1$. The proofs are provided in section A.1 of the appendix.

First, we obtain some results that are based on a representation of the hybrid DP_k and DP priors as mixtures of DPs (Antoniak, 1974), with transition measure given by a 'hybrid' version of the empirical distribution of $\theta_1^*, \ldots, \theta_k^*$. Recall that an RPM G on a space Θ has a mixture of DPs probability law, with transition measure ν, if $G|H \sim \mathrm{DP}(\alpha H)$ and H is a random measure with distribution ν. We write $G \sim \int \mathrm{DP}(H)\,\mathrm{d}\nu(H)$. If H is non-random, the mixture of DPs reduces to $G \sim \mathrm{DP}(H)$.

Let $G_k \sim \mathrm{hDP}_k(\alpha_k q_k, G_0)$, where the labelling measure q_k is not restricted to having uniform marginals. Here, we regard q_k as a probability measure on $\{1, 2, \ldots\}^m$, with support $\{1, \ldots, k\}^m$. Let $(\theta_j^* = (\theta_{1,j}^*, \ldots, \theta_{m,j}^*), j = 1, 2, \ldots)$ be a random sample from G_0, and define the RPMs

$$Q_k = Q_k(\theta_1^*, \theta_2^*, \ldots) = \sum_{j_1=1}^{\infty} \cdots \sum_{j_m=1}^{\infty} q_k(j_1, \ldots, j_m)\delta_{\theta_{1,j_1}^*, \ldots, \theta_{m,j_m}^*}, \tag{12}$$

for $k = 1, 2, \ldots$. Given a probability measure q on $\{1, 2, \ldots,\}^m$, define a random measure Q similarly, replacing q_k with q. Finally, denote by μ_k and μ respectively the probability laws of Q_k and Q. Then, the hybrid Dirichlet priors admit the following representation.

Proposition 1. Let G be an RPM on $\mathbb{R}^m$ and Q_k and Q be defined as above. Then,

(a) an $\mathrm{hDP}(\alpha q, G_0)$ prior for G is a mixture of DPs $\int \mathrm{DP}(\alpha Q)\, \mathrm{d}\mu(Q)$ and
(b) an $\mathrm{hDP}_k(\alpha_k q_k, G_0)$ prior for G is a mixture of DPs $\int \mathrm{DP}(\alpha_k Q_k)\, \mathrm{d}\mu_k(Q_k)$.

This result is extended to the functional hDP_k prior in section A.1 of the appendix. Proposition 1 suggests that the weak limit of hybrid DP_k priors depends on the limiting behaviour of the sequence $(Q_k, k \geqslant 1)$.

Proposition 2. Let $(G_k, k \geqslant 1)$ be a sequence of RPMs on $\mathbb{R}^m$ with $G_k \sim \mathrm{hDP}_k(\alpha_k q_k, G_0)$. Let Q_k be defined by equation (12), $k \geqslant 1$. Assume that $\alpha_k \to \alpha$, $0 < \alpha < \infty$, for $k \to \infty$.

(a) If Q_k converges in distribution to a non-random probability measure Q_∞ on $\mathbb{R}^m$, the $\mathrm{hDP}_k(\alpha_k q_k, G_0)$ converges weakly to a $\mathrm{DP}(\alpha Q_\infty)$.
(b) If Q_k converges in distribution to an RPM Q_∞ on $\mathbb{R}^m$, with probability law μ, then the $\mathrm{hDP}_k(\alpha_k q_k, G_0)$ converges weakly to $\int \mathrm{DP}(\alpha Q_\infty)\, \mathrm{d}\mu(Q_\infty)$.

The proof is based on proposition 1 and the fact that, under the assumptions, the integral $\int \mathrm{DP}(\alpha Q_k)\, \mathrm{d}\mu_k(Q_k)$ converges weakly to $\int \mathrm{DP}(\alpha Q_\infty)\, \mathrm{d}\mu(Q_\infty)$. Of course, the behaviour of the random measures Q_k depends crucially on that of the sequence of labelling measures q_k. The case where $q_k(j_1, \ldots, j_m) = q_{j,k} > 0$ if $j_1 = \ldots = j_m = j$, with $j = 1, \ldots, k$, and it is 0 otherwise, reduces the $\mathrm{hDP}_k(\alpha_k q_k, G_0)$ to a finite Dirichlet prior. In this case, Q_k is a weighted empirical distribution of the sample $\theta_1^*, \ldots, \theta_k^* \sim^{\mathrm{IID}} G_0$. In particular, it is their empirical DF for a symmetric DP_k prior, where $q_{j,k} = 1/k$. If $\alpha_k \to \alpha$, $0 < \alpha < \infty$ and $\max(q_{1,k}, \ldots, q_{k,k}) \to 0$ for $k \to \infty$, then it can be shown that Q_k converges in distribution to the (non-random) DF G_0. Thus, by proposition 2, G_k converges in distribution to a random DF $G \sim \mathrm{DP}(\alpha G_0)$; this is in accordance with theorem 3 (part 2) by Ishwaran and Zarepour (2002).

Finding general conditions on the q_k parameters such that Q_k converges to a probability measure is beyond the scope of this paper (see corollary 2.5 in Berti *et al.* (2006)). However, the following theorems relate more explicitly the behaviour of hybrid Dirichlet priors to that of the labelling measures q_k.

We start with the simple case $m = 1$, i.e. G_k is a random DF on $\mathbb{R}$. In this case, the hDP_k reduces to a DP_k.

Theorem 1. Let $(G_k, k \geqslant 1)$ be a sequence of DFs on $\mathbb{R}$, with $G_k \sim \mathrm{DP}_k(\alpha_k q_k, G_0)$. Suppose that $\alpha_k \to \alpha$, $0 < \alpha < \infty$ for $k \to \infty$.

(a) If $q_k(j) \to 0$ for all $j = 1, 2, \ldots$ for $k \to \infty$, then $G_k \to G$ in distribution, where $G \sim \mathrm{DP}(\alpha G_0)$.
(b) If $q_k(j) \to q(j)$ for any $j = 1, 2, \ldots$, with $q(1), q(2), \ldots$ defining a probability measure q on $\{1, 2, \ldots\}$, then $G_k \to G$ in distribution, where $G \sim \mathrm{hDP}(\alpha q, G_0)$.

Although case (a) is known, a comparison between the two parts clarifies the role of the sequence of q_ks. In part (a), the limit of $q_k(j)$ is 0 for all js, so q_k does not converge to a probability measure. In fact, the proof of part (a) considers the vector of the *ordered* weights in equation (6), which converges in distribution to a random sequence of weights $p^* = (p_1^*, p_2^*, \ldots)$ having a Poisson–Dirichlet distribution (Kingman, 1975). Hence, G_k converges in distribution to $G = \Sigma_{j=1}^{\infty} p_j^* \delta_{\theta_j^*}$, with the θ_j^*s IID from G_0; such a G is a $\mathrm{DP}(\alpha G_0)$. In part (b), the limit q of

q_k is a probability measure, and the proof uses the fact that $p_k \sim \mathcal{D}(\alpha q_k)$ converges in distribution to an RPM $p \sim \mathrm{DP}(\alpha q)$, so that the limit of G_k is a hybrid DP prior.

Now consider the general case $m \geqslant 1$. Let $G_k \sim \mathrm{hDP}_k(\alpha_k q_k, G_0)$, $k \geqslant 1$, and $G_{k,i}$ be its ith marginal. Then, $G_{k,i} \sim \mathrm{DP}_k(\alpha_k q_{k,i}, G_{0,i})$, where $q_{k,i}$ and $G_{0,i}$ are respectively the ith marginals of q_k and of G_0. Thus, theorem 1 applies to the random marginals of G_k. To extend the result to the joint DF G_k, we need to introduce further notation. Let $C_{d,n_1,\ldots,n_d,o}$ denote a *configuration* of the indices $(j_1,\ldots,j_m)$ in $\{1,2,\ldots\}^m$, characterized by d distinct values, repeated respectively $n_1,\ldots,n_d$ times, in the order given by o; for example, if $m=3$, $C_{1,3,(1,1,1)} = C_{1,3}$ denotes the set of triples (i,i,i) with $i \in \{1,2,\ldots\}$, $C_{2,1,2,(1,2,1)}$ is the set of triples (i,j,i) with $i, j \in \{1,2,\ldots\}$, $i \neq j$, etc. Denote by $\mathcal{C}$ the class of all possible configurations. Given a configuration C, and the sample $(\theta_j^* = (\theta_{1,j}^*,\ldots,\theta_{m,j}^*), j=1,\ldots,k)$ from G_0, let $G_{0,C}$ be the DF of the vector $(\theta_{1,j_1}^*,\ldots,\theta_{m,j_m})$ whose co-ordinates are chosen from the θ_j^*s according to the labels $(j_1,\ldots,j_m)$ in C. For example, if $m=3$ and $C = C_{1,3}$, $G_{0,C} = G_0$; if $C = C_{2,1,2,(1,2,1)}$, $G_{0,C} = G_{0,2}G_{0,1,3}$, where $G_{0,2}$ and $G_{0,2,3}$ are the marginal distributions of $\theta_{2,j}^*$ and $(\theta_{1,j}^*, \theta_{3,j}^*)$ respectively. Then for the joint DF G_k on $\mathbb{R}^m$ we have the following limiting result.

Theorem 2. Let $(G_k, k \geqslant 1)$ be a sequence of random DFs on $\mathbb{R}^m$, $G_k \sim \mathrm{hDP}_k(\alpha_k q_k, G_0)$. Suppose that $\alpha_k \to \alpha$, $0 < \alpha < \infty$, for $k \to \infty$.

(a) If $q_k(j_1,\ldots,j_m) \to 0$ for all $(j_1,\ldots,j_m) \in \{1,2,\ldots\}^m$, then $G_k \to G$ in distribution, with $G \sim \mathrm{DP}(\alpha G_q)$, where the base measure G_q is given by $G_q = \Sigma_{C\in\mathcal{C}}\, q(C) G_{0,C}$, with $q(C) = \lim_{k\to\infty}\{q_k(C)\}$, $C \in \mathcal{C}$.

(b) If $q_k(j_1,\ldots,j_m) \to q(j_1,\ldots,j_m)$ for all $(j_1,\ldots,j_m)$, where the $q(j_1,\ldots,j_m)$s define a probability measure q on $\{1,2,\ldots\}^m$, then $G_k \to G$ in distribution, where $G \sim \mathrm{hDP}(\alpha q, G_0)$.

In practice, the expression of the base measure G_q in part (a) is simplified, since many of the weights $q(C)$ are usually negligible. Note the difference between samples from the limit G in parts (a) and (b). If $\theta_i = (\theta_i(x_1),\ldots,\theta_i(x_m)) | G \overset{\mathrm{IID}}{\sim} G$ and $G \sim \mathrm{DP}(\alpha G_q)$, only global ties can be modelled, although G_q may model species that present inhomogeneous traits, or irregular areas of random shape. Instead, if $G \sim \mathrm{hDP}(\alpha q, G_0)$ (case (b)), it is possible to model global and local clusters alike, i.e. the individual vectors may share just some of their co-ordinates.

Condition (b) of theorem 2 is not satisfied if q_k has uniform marginals $q_{k,i}$, i.e. at each co-ordinate the weights are symmetric Dirichlet. In fact, in that case, $q_{k,i}(j) = 1/k \to q_i(j) = 0$ for all $j = 1,2,\ldots$, so q_i cannot be the marginal of a probability measure q on $\{1,2,\ldots\}^m$. Therefore, if a symmetric Dirichlet distribution is used, as is common in applications of mixture models, we should be aware that the properties of the model for large k may be quite different from those for small k. In fact, as we illustrate in Section 5, hDP_k mixture models usually succeed in giving a good reconstruction of the data for small values of k, by allowing local and global clustering; however, the previous results should be kept in mind when studying sensitivity to the choice of k.

4. Applications to mixture modelling

In this section, we apply hybrid Dirichlet priors to mixture models, as in Section 2. Again, we consider a sample of curves $\mathbf{Y}_i = (Y_i(x), x \in D)$, observed at a finite set of locations $(x_1,\ldots,x_m)$. The model for these data is described by expression (3), except here $G_{x_1,\ldots,x_m} \sim \mathrm{hDP}_k(\alpha q_{x_1,\ldots,x_m}, G_{0,x_1,\ldots,x_m})$. Integrating out the θ_is, from equation (10) we obtain

$$Y_i | G \overset{\mathrm{IID}}{\sim} \sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} p_{x_1,\ldots,x_m}(j_1,\ldots,j_m)\mathcal{N}_m\{(\theta_{j_1,1}^*,\ldots,\theta_{j_m,m}^*), \sigma^2 I_m\}, \tag{13}$$

that is a location mixture of Gaussian distributions with local random effects.

The role of the labelling component of the hDP_k prior may be further specified. As discussed in the previous sections, finite Dirichlet and DP mixtures imply a global random partition of the mean vectors θ_is; only a scalar $\gamma_i \in \{1, \ldots, k\}$ labels the species allocation. Under equation (6), $\Pr(\theta_i = \theta_{i'} = \theta_j^* | p) = \Pr(\gamma_i = \gamma_{i'} = j | p) = p_j^2$. Instead, hybrid Dirichlet priors provide dependent local random partitions. Thus, we imagine a vector of labels $\gamma_i = (\gamma_i(x_1), \ldots, \gamma_i(x_m))$. For each co-ordinate x, $Y_i(x)$ and $Y_{i'}(x)$ are described by the same mean vectors $\theta_i(x) = \theta_{i'}(x)$ if $\gamma_i(x) = \gamma_{i'}(x)$. Moreover, the partition that is induced at co-ordinate x_1 is related to that obtained at x_2 through the functional dependence that is expressed by the weights p; in particular, $\Pr\{\gamma_i(x_1) = \gamma_{i'}(x_1) = j_1, \gamma_i(x_2) = \gamma_{i'}(x_2) = j_2 | p\} = p^2_{x_1,x_2}(j_1, j_2)$. Thus, the model can be reformulated in terms of the hidden label process,

$$Y_i | \theta_i, \gamma_i, \theta_1^*, \ldots, \theta_k^* \overset{\text{ind}}{\sim} \mathcal{N}_m\{\theta^*(\gamma_i), \sigma^2 I_m\}, \tag{14}$$

where $\theta^*(\gamma) = (\theta_{j_1,1}, \ldots, \theta_{j_m,m})$ if $\gamma = (j_1, \ldots, j_m)$,

$$\gamma_i | p_{x_1,\ldots,x_m}, \theta_1^*, \ldots, \theta_k^* \overset{\text{IID}}{\sim} p_{x_1,\ldots,x_m},$$
$$p_{x_1,\ldots,x_m} | \xi \sim \mathcal{D}\{\alpha q_{x_1,\ldots,x_m}(\cdot | \xi)\},$$
$$\theta_j^* \overset{\text{IID}}{\sim} G_{0,x_1,\ldots,x_m},$$

and $p_{x_1,\ldots,x_m}$ and the θ_j^*s are independent. Above, the labelling prior is centred on a parametric model $q_{x_1,\ldots,x_m}(\cdot|\xi)$, such that

$$\Pr\{\gamma_i = (j_1, \ldots, j_m) | \xi\} = E\{p_{x_1,\ldots,x_m}(j_1, \ldots, j_m) | \xi\} = q_{x_1,\ldots,x_m}(j_1, \ldots, j_m | \xi),$$

with ξ being a vector of hyperparameters. For $\alpha \to \infty$, the Dirichlet labelling prior degenerates on the parametric model $q_{x_1,\ldots,x_m}(\cdot|\xi)$.

Consistency of the model for varying grids x is obtained by regarding the γ_is as the values at $x_1, \ldots, x_m$ of a hidden *process* of labels $\boldsymbol{\gamma}$, as in Section 3.2. Thus, hybrid Dirichlet mixtures offer a general framework for capturing the effect of a hidden labelling process in inference for functional data.

4.1. A copula model for the label process

If $k < \infty$, a common choice in mixture models is to assume a symmetric Dirichlet distribution on the mixing weights. The case of a symmetric DP_k prior corresponds to the choice of $\mathbf{q}$, where $q_{x_1,\ldots,x_m}$ is degenerate on the 'diagonal' of the hypercube $\{1, \ldots, k\}^m$, i.e., for any choice of $(x_1, \ldots, x_m)$, $q_{x_1,\ldots,x_m}(j_1, \ldots, j_m)$ equals $1/k$ if $j_1 = \ldots = j_m$ and 0 otherwise. Hence, $p_{x_1,\ldots,x_m}(j_1, \ldots, j_m) = 0$ almost surely unless $j_1 = \ldots = j_m$, so the label vector γ_i has almost surely identical co-ordinates, $\gamma_i(x_1) = \ldots = \gamma_i(x_m)$. In this section we provide a more general model for $\mathbf{q}$ with flexible dependence structure and uniform marginals.

The idea is to assign $q_{x_1,\ldots,x_m}$ by means of an auxiliary absolutely continuous DF, say $H_{0,x_1,\ldots,x_m}$. To be more specific, we take a DF $H_{0,x_1,\ldots,x_m}$ on $(0,1)^m$ with uniform marginals, i.e. $H_{0,x_1,\ldots,x_m}$ is an m-variate copula with suitable dependence structure. Let us partition $(0,1)^m$ in hypercubes $C_{j_1,\ldots,j_m}$ with sides $((j_i - 1)/k, j_i/k]$, the first closed also on the left, and let

$$q_{x_1,\ldots,x_m}(j_1, \ldots, j_m) = H_{0,x_1,\ldots,x_m}(C_{j_1,\ldots,j_m}) \qquad j_i = 1, \ldots, k, \quad i = 1, \ldots, m. \tag{15}$$

Then, $q_{x_1,\ldots,x_m}$ has uniform marginals on $\{1, \ldots, k\}$ and a dependence structure induced by $H_{0,x_1,\ldots,x_m}$. The appeal of this construction is that sampling from $q_{x_1,\ldots,x_m}$ becomes straightforward (it is enough to sample from the continuous $H_{0,x_1,\ldots,x_m}$, instead of computing k^m weights

as would be otherwise required). In what follows, we use a Gaussian copula. More specifically, we consider $L = (L_1, \ldots, L_m) \sim N_m(0, \Sigma)$, with marginal DFs $F_1, \ldots, F_m$, and let $U_i = F_i(L_i)$, $i = 1, \ldots, m$. Then, each U_i has a uniform distribution on $(0, 1)$ and their joint distribution $H_0 = H_0(\cdot; \Sigma)$ reflects the dependence structure of the underlying Gaussian distribution. The latter construction is extended to the functional case by considering a stationary Gaussian process $\mathbf{L} = (L(x), x \in D)$ with mean function identically 0 and covariance function $\sigma(x - x')$ say, $\mathbf{L} \sim \mathbf{F} = \mathcal{GP}\{0, \sigma(x - x')\}$. For any $x \in D$, let $U(x) = F_x\{L(x)\}$. Then, the random process $\mathbf{U} = (U(x), x \in D)$ has probability law $\mathbf{H}_0 = \mathbf{H}_0(\cdot|\xi)$; for any choice of $m \geqslant 1$ and $(x_1, \ldots, x_m)$, the vector $(U(x_1, \ldots, U(x_m)) \sim H_{0,x_1,\ldots,x_m}$ where $H_{0,x_1,\ldots,x_m}$ is an m-variate Gaussian copula. The family of probability measures $q_{x_1,\ldots,x_m}$ that is obtained by discretizing $H_{0,x_1,\ldots,x_m}$ defines a probability measure $\mathbf{q}$ on $\{1, \ldots, k\}^D$, with uniform marginals and dependence structure regulated by $\mathbf{H}_0$. We use this $\mathbf{q}$ in the modelling above and for the applications that are described in Section 5.

4.2. Study of the prior

The nature and the behaviour of realizations from an hDP_k prior with a copula model for the label process can be best appreciated by means of a simulation study. Here, we imagine that x is univariate, e.g. time. Our aim is to illustrate three basic features of the hybrid Dirichlet prior: the choice of the canonical curves, through the base measure G_0; the modelling of the species recombination (hybridization), through the labelling prior; the clustering, that results from the predictive rule. Sensitivity to k and in particular the behaviour of the prior for increasing k has been discussed in Section 3.3.

As base measure G_0, we use an m-variate Gaussian distribution with zero mean and covariance matrix $\sigma_0^2\, R(\phi_0)$, where $R(\phi_0)$ is the correlation matrix with (i, j)th entry $\exp\{-\phi_0(x_i - x_j)^2\}$, $\phi_0 \geqslant 0$, i.e. the finite dimensional distribution of a Gaussian process with exponential covariance function. If σ_0^2 is large, we expect high variability across the pure species. Thus, small differences between two vectors θ_i and $\theta_{i'}$ are explained by the random error, rather than being described by two different species, say θ_j^* and $\theta_{j'}^*$. The decay parameter ϕ_0 controls the 'smoothness' of the vector realizations. In general, we imagine smooth canonical curves and, thus, ϕ_0 will typically be small.

The labelling prior is centred on a distribution $q = q(\cdot; \phi_q)$ specified via the auxiliary Gaussian copula. Here, we use the copula that is obtained from an m-variate Gaussian distribution with mean vector zero, variance 1 and correlation matrix $R(\phi_q)$. Note that this choice satisfies the continuity property that was discussed in Section 3.2, since for $x \to x'$ the correlation between $\gamma(x)$ and $\gamma(x')$ goes to 1. The labelling prior controls the amount of recombination of the pure vectors θ_j^*s. To discourage too much local selection, q may be fairly concentrated on the diagonal $(i, \ldots, i)$ of $\{1, \ldots, k\}^m$; the limit case of q degenerate on the diagonal is obtained for $\phi_q = 0$ and corresponds to the DP_k prior on G. For the copula model, there is an interplay between k and the probability that is assigned on the diagonal of $\{1, \ldots, k\}^m$; given ϕ_q, this probability decreases as k becomes larger.

Fig. 1 shows samples from the hDP_k prior, for various values of the hyperparameters. We consider $\theta = (\theta(x_1), \ldots, \theta(x_m))|G \sim G$ and $G \sim \mathrm{hDP}_k(\alpha q, G_0)$, with q and G_0 as above, on a grid of $m = 200$ equally spaced points in $D = (0, 100)$. Here, $k = 5$, $\alpha = 5$ and $\sigma_0^2 = 9$, whereas ϕ_0 and ϕ_q vary as detailed in Fig. 1. In particular, $\phi_0 = 0.01$ illustrates the limit case of constant canonical curves θ_j^* ($\phi_0 = 0$). Instead, $\phi_q = 0.1$ exemplifies the effects of a strong correlation in the label process; as a consequence, we observe none or just a few change-points. If $\phi_q = 3$, we allow for an accentuated local selection; therefore, we may observe (slight or abrupt) changes

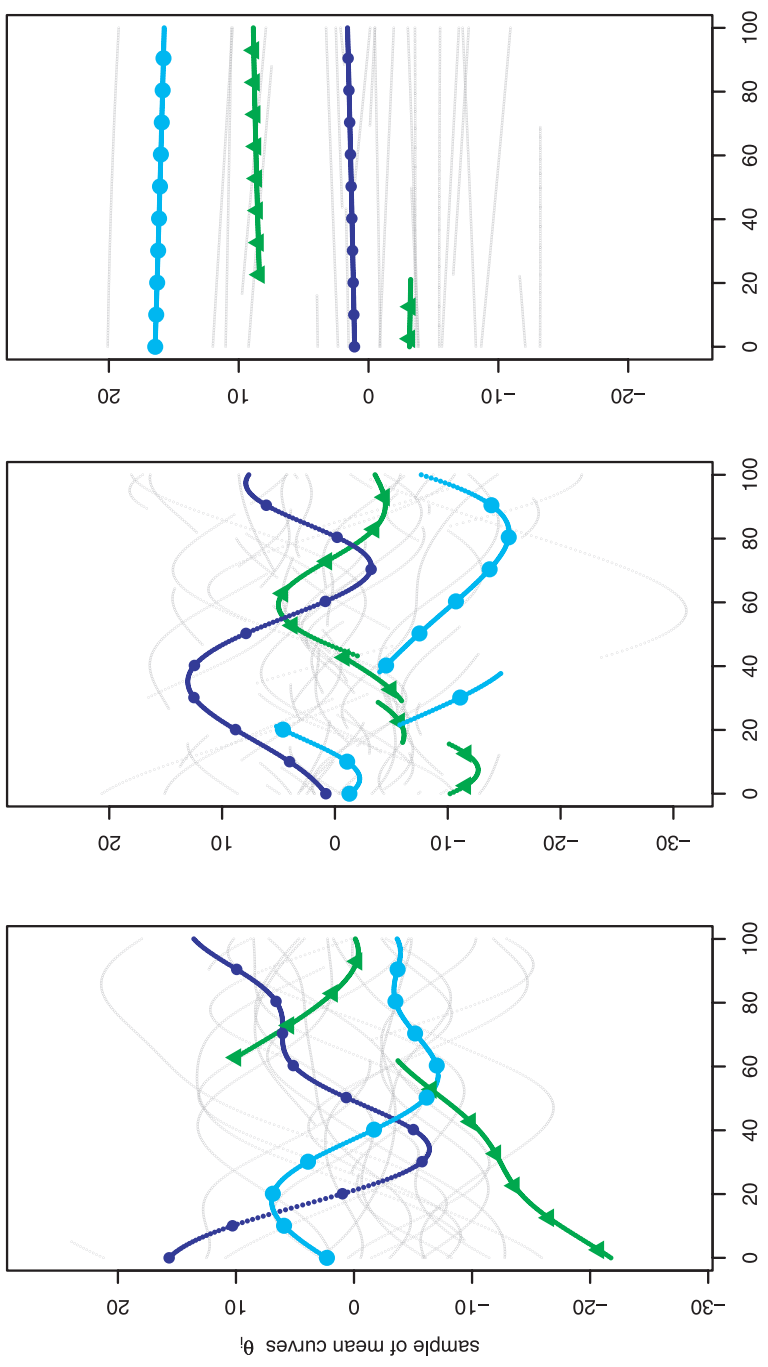

Fig. 1. Samples from the hDP$_k$ prior, for various values of the hyperparameters (some pure and hybrid curves are highlighted; the pure species have continuous trajectories whereas hybrid curves are clearly characterized by their discontinuities; details are given in Section 4.2): (a) $\phi_0 = 3$ and $\phi_q = 0.1$; (b) $\phi_0 = 3$ and $\phi_q = 3$; (c) $\phi_0 = 0.01$ and $\phi_q = 0.1$

in the trajectories. The curves are essentially IID samples from $E(G) = \Sigma_C\, q(C) G_{0,C}$, with $G_{0,C}$ as in theorem 2. For illustration, in Fig. 1 we highlight a few pure and hybrid curves.

Now consider the joint distribution of $(\theta_1, \ldots, \theta_n)$, with $n > 1$. Since, $\theta_i(x)|G_x \sim^{\text{IID}} G_x$, with $G_x \sim \text{DP}_k(\alpha, G_{0,x})$, marginally $\theta_1(x) \sim G_{0,x}$ and $\theta_{i+1}(x)|\theta_1(x), \ldots, \theta_i(x)$ is assigned to one of the species observed at x, or to a new species, according to distribution (8). The *joint* predictive rule of $\theta_{i+1}|\theta_1, \ldots, \theta_i$ takes into account the possibility of dependent local mutations. To describe the structure of the implied joint clustering, we exploit the hidden labels formulation of the model, i.e. we sample from the joint distribution of the θ_is and the labels,

$$\pi(\theta_1, \ldots, \theta_n, \gamma_1, \ldots, \gamma_n) = \pi(\theta_1, \ldots, \theta_n|\gamma_1, \ldots, \gamma_n)\, \pi(\gamma_1, \ldots, \gamma_n).$$

A sample from $\pi(\gamma_1, \ldots, \gamma_n)$ is obtained by drawing the components one at a time, from the Dirichlet updating rule; namely $\gamma_1 \sim q$, and, for $i > 1$,

$$\gamma_{i+1}|\gamma_1, \ldots, \gamma_i \sim \frac{\alpha}{\alpha + i} q + \frac{1}{\alpha + i} \sum_{j=1}^{d_i} n_j \delta_{\gamma_j^*}, \tag{16}$$

where $\gamma_1^*, \ldots, \gamma_{d_i}^*$ are the vectors among $\gamma_1, \ldots, \gamma_i$ that differ at least in one co-ordinate, and n_j is the frequency of γ_j^*, $j = 1, \ldots, d_i$. Sampling from distribution (16) is made simple by the copula labelling prior, since we can sample the underlying, continuous vector $U \sim H_0$ instead of the high dimensional q. Rule (16) suggests that the Dirichlet parameter α controls the probability of global ties; however, global and local ties can occur, with probabilities governed by q. If q models strong correlation between labels, such that $q(j, \ldots, j) > q(j_1, \ldots, j_m)$ for unequal j_is, the probability that γ_{n+1} is one of the vectors already in the sample, say γ_i^*, is greater if $\gamma_i^* = (j, \ldots, j)$, so pure or canonical species tend to be more frequent in the sample, whereas hybrid species are more rare. Given $(\gamma_1, \ldots, \gamma_n)$, the mean curves $\theta_1, \ldots, \theta_n$ are generated according to the labels' configurations. In practice, a sample from $\pi(\theta_1, \ldots, \theta_n|\gamma_1, \ldots, \gamma_n)$ is obtained by first generating $\theta_1^*, \ldots, \theta_k^* \sim^{\text{IID}} G_0$ and then letting $\theta_i = \theta^*(\gamma_i)$, where $\theta^*(\gamma)$ is defined as in distribution (14).

As a proof of purpose, we provide a simple illustration of the clustering that is implied by the hDP_k prior in Fig. 2, for $n = 10$, $k = 5$ and $m = 100$, and hyperparameters $\sigma_0^2 = 9$, $\phi_0 = 3$, and $\phi_q = 0.1$ and $\alpha = 5$ (Figs 2(a) and 2(b)) or $\phi_q = 1$ and $\alpha = 20$ (Figs 2(c) and 2(d)). Figs 2(a) and 2(c) show a sample $\gamma_1, \ldots, \gamma_{10}$ from the Dirichlet updating rule (16). The frequency of each configuration is reported on the outermost right-hand side of the panels. Figs 2(b) and 2(d) show the corresponding sample of $(\theta_1, \ldots, \theta_{10})$. For α small and strongly correlated labels, we tend to observe global ties, whereas higher values of α and ϕ_q favour local selection and local clustering. In the example of Figs 2(a) and 2(b), we distinguish two global clusters; more specifically, five θ_is share the constant label $\gamma(x) = 1$ and three have $\gamma(x) = 4$, for all x. In addition, we observe local clusters; two of the θ_is are hybrids since they belong to different species at different points. Some of these behaviours are explicitly highlighted in Fig. 2. Figs 2(c) and 2(d) exemplify the situation where α and ϕ_q are large, to favour local selection and local clustering; this is clearly indicated by the larger number of change-points on the trajectories of the hybrid curves.

4.3. Markov chain Monte Carlo computations

As discussed at the beginning of this section, the hybrid Dirichlet prior leads to a mixture model (13) with local random effects. Bayesian inference via MCMC sampling for mixture models usually takes advantage of the hidden labels to facilitate sampling; see, for example, Marin and Robert (2007). However, the local nature of the likelihood in our case makes standard MCMC algorithms not computationally feasible; sampling the k^m mixing weights in distribution (13) is not manageable. We integrate the p weights out and, using distribution (14), we focus on the posterior

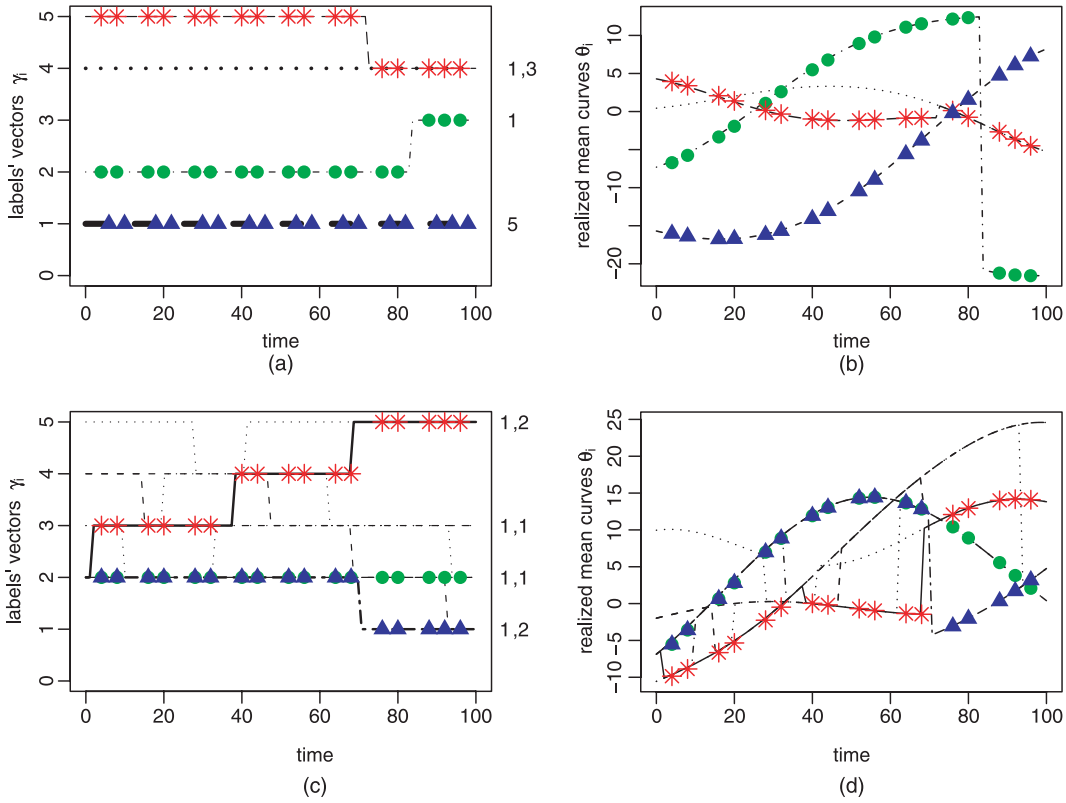

Fig. 2. Example of simulation of one sample from the joint distribution of $(\gamma_1,\ldots,\gamma_n)$ and $(\theta_1,\ldots,\theta_n)|(\gamma_1,\ldots,\gamma_n)$ (the $n = 10$ vectors are clustered both globally and locally, as detailed in Section 4.2): here, $m = 100$, $k = 5$, $\sigma_0^2 = 9$ and $\phi_0 = 3$, and (a), (b) $\phi_q = 0.1$ and $\alpha = 5$, and (c), (d) $\phi_q = 1$ and $\alpha = 20$

$$\pi(\gamma_1,\ldots,\gamma_n,\theta_1^*,\ldots,\theta_k^*|y_1,\ldots,y_n) \propto \prod_{i=1}^{n} N\{y_i|\theta^*(\gamma_i),\sigma^2\}\,\Pr(\gamma_1,\ldots,\gamma_n)\prod_{j=1}^{k} g_0(\theta_j^*),$$

where $\theta^*(\gamma_i)$ is defined as in distribution (14), g_0 is the density corresponding to G_0 and $\Pr(\gamma_1,\ldots,\gamma_n)$ is characterized by the predictive rule (16). The challenges for MCMC computations are evident in the discreteness of the γ_is, the presence of the allocation prior q in rule (16) and the way that the γ_is enter the likelihood. We suggest a computational strategy that can be described in two steps.

Recall that $\gamma_i|p \sim^{\mathrm{IID}} p$ and $p \sim \mathcal{D}(\alpha q)$. The first step requires rewriting the latter in a computationally useful way. In fact, if q is assigned through the copula construction, we can set $\gamma_i = (j_1,\ldots,j_m)$ if and only if $U_i \in C_{j_1,\ldots,j_m}$, where $U_1,\ldots,U_n$ are auxiliary random variables such that $U_i|H \sim^{\mathrm{IID}} H$, $H \sim \mathrm{DP}(\alpha H_0)$. Indeed,

$$p(j_1,\ldots,j_m) = \Pr\{\gamma_i = (j_1,\ldots,j_m)|p\} = \Pr(U \in C_{j_1,\ldots,j_m}|H) = H(C_{j_1,\ldots,j_m}),$$

and, by the well-known properties of the DP, the vector of probabilities $H(C_{j_1,\ldots,j_m})$ on the finite partition in hypercubes of $(0,1)^m$ has a Dirichlet distribution with parameters $H_0(C_{j_1,\ldots,j_m}) = q(j_1,\ldots,j_m)$, as required. The high dimensional discrete labels γ_i are replaced by the easier-to-sample, continuous, U_i. The posterior becomes

$$\pi(u_1,\ldots,u_n,\theta_1^*,\ldots,\theta_k^*|y_1,\ldots,y_n) \propto \prod_{i=1}^{n} N\{y_i|\theta^*(U_i),\sigma^2\}\,\pi(u_1,\ldots,u_n)\,\prod_{j=1}^{k} g_0(\theta_j^*),$$

where $\pi(\cdot)$ is the probability law of $U_1,\ldots,U_n$ characterized by the Polya urn scheme that is associated with the DP. However, the update of the U_is in a Gibbs sampler is still complicated, because of the local nature of the likelihood term (piecewise constant over the hypercubes $C_{j_1,\ldots,j_m}$). It is worth noting that sampling the vectors U_i one co-ordinate at a time would not be feasible, owing to the peculiar nature of the random conditional distribution of $U_i(x_1)|U_i(x_2),\ldots,U_i(x_m)$ for a DP (Ramamoorthi and Sangalli, 2006); here, that would lead to a Markov chain that is not reversible.

Thus, we perturb the U_is with a small noise or 'jitter'; details are provided in section A.2 of the appendix. Basically, the jitter is an artificial perturbation of the label U_i, say $\tilde{U}_i$, such that

$$P(\tilde{U}_i \in C_{j_1,\ldots,j_m}|U_i \in C_{j_1,\ldots,j_m}) \approx 1,$$

so that the resulting allocation is not significantly altered by the perturbation. Moreover, we assume that

$$\tilde{U}_i(x_j)|U_i(x_j) = u \overset{\text{ind}}{\sim} f_j\{\tilde{u}_i(x_j)|u\}, \qquad j=1,\ldots,m, \quad i=1,\ldots,n,$$

i.e. independence across the xs. Thus, we can augment the parameter space and approximate the posterior distribution with its 'jittered' version,

$$\begin{aligned}\pi(u_1,\ldots,u_n,\tilde{u}_1,\ldots,\tilde{u}_n,\theta_1^*,\ldots,\theta_k^*|y_1,\ldots,y_n) \propto{}& \prod_{i=1}^{n}\prod_{j=1}^{m} N[y_i(x_j)|\theta^*\{\tilde{u}_i(x_j)\},\sigma^2] \\ &\times \prod_{i=1}^{n}\prod_{j=1}^{m} f_j\{\tilde{u}_i(x_j)|u_i(x_j)\}\,\pi(u_1,\ldots,u_n)\prod_{j=1}^{k} g_0(\theta_j^*).\end{aligned} \tag{17}$$

The aim of the jitter is evident. The jittered posterior is easier to sample from, since we can now separate the update of $\tilde{U}_i$ from that of U_i. Moreover, given the independence of the U_is, the posterior allocation is determined by sampling only from m-dimensional distributions. An MCMC algorithm for the slightly perturbed posterior distribution (17) can be easily described in the following steps (details are given in section A.2 of the appendix).

(a) Update $\tilde{U}_1,\ldots,\tilde{U}_n$, one co-ordinate at a time from the full conditional of $\tilde{U}_i(x_j)$,

$$\pi\{\tilde{u}_i(x_j)|u_i(x_j),y_1,\ldots,y_n\} \propto N[y_i(x_j)|\theta^*\{\tilde{u}_i(x_j)\}]\,f_j\{\tilde{u}_i(x_j)|u_i(x_j)\}.$$

(b) Update $U_1,\ldots,U_n$ one at a time. The full conditional of U_i,

$$\pi(u_i|\tilde{u}_i,u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_n) \propto \prod_{j=1}^{m} f_j\{\tilde{u}_i(x_j)|u_i(x_j)\}\,\pi(u_1,\ldots,u_n),$$

is the standard full conditional for the point masses of an m-dimensional DP mixture model. We use the strategy that was described by Bush and MacEachern (1996) for improving possible slow mixing due to the update of the U_is one at a time.

(c) Update $\theta_1^*,\ldots,\theta_k^*$ one at a time, on the basis of the allocation structure that is defined by the label process $\gamma(x)$. The full conditional of θ_j^* is $\mathcal{N}(\mu_m,\Lambda_m)$, where $\mu_m = (1/\sigma^2)\Lambda_m \Sigma_{i,l:\gamma_i(x_l)=j} Y_i(x_l)$, and

$$\Lambda_m = \left[\frac{1}{\sigma_0^2}\,R^{-1}(\phi_0) + \frac{1}{\sigma^2}\sum_{i,l} I\{\gamma_i(x_l)=j\}\right]^{-1}.$$

The specific implementation and the full conditionals that are used for the simulation and application example in Section 5 are detailed in the appendix.

4.4. Alternative label process models

Many other choices for the random labelling measure could be considered. For example, we might assume a Markov dependence for the label process. Then, our framework would lead to Bayesian mixtures of hidden Markov models, a development which is in line with Teh *et al.* (2006). In a rather different spirit, Green and Richardson (2002) employed a hidden Potts model for spatial count data. Here, the set of xs is finite and the association is captured through a proximity matrix. Under the Potts model, $P\{\gamma(x_r) = j_r, r = 1, 2, \ldots, m|\delta\} \propto \exp\{-\delta \Sigma_{r \neq r'} I(j_r = j_{r'})\}$. As known, calculation of the normalizing constant (a function of δ) requires summation over $(j_1, j_2, \ldots, j_m) \in \{1, 2, .., k\}^m$. More similar to our approach is the work of Duan *et al.* (2007). They specified a generalized DP model through the formalization of multivariate stick breaking distributions to create partitions of an m-dimensional hypercube. In their application, they created labels by using a countable collection of IID Gaussian process replicates, say $\mathbf{Z}_u$, $u = 1, 2, \ldots$. Then, $P\{\gamma(x_r) = j_r, r = 1, 2, \ldots, m\} = P\{Z_u(x_r) < 0, u < j_r, Z_{j_r} > 0, r = 1, 2, \ldots, m\}$. For a fixed variance, these probabilities are determined by the mean and decay parameters of the Gaussian processes; if those parameters are random, then the stick breaking distribution is, as well. Model fitting is done by truncating the stick breaking to, say, k pieces but computation remains challenging since it requires k Gaussian processes instead of just one.

Finally, we observe that the copula construction of q is attractive for its simplicity and flexibility. However, it suggests an ordering of the labels' values that is not further exploited when we generate the associated species. Such behaviour is common in mixture modelling, given the standard assumption of IID support points for the mixing distribution. In other words, a change in the label's value reflects a change-point in the curve trajectory, but *a priori* the values of the curve before and after the change-point are identically distributed; of course, the prior assumption is updated with inferential evidence. Otherwise, we could introduce an order among the θ_j^*; however, that would require *ad hoc* assumptions and may not be easy when m is large, unless G_0 is degenerate on constant species (see Rodriguez *et al.* (2008)). A different approach may use a 'symmetric' labelling prior, i.e. a model where q is *uniform* on each subspace of label configurations (see Section 3.3). Then, the copula construction might be helpful to specify the values of q on each configuration set, also taking into account the continuity requirement of Section 3.2. Of course, we would need to consider a different DF H_0 or a different partition of $(0, 1)^m$. These extensions are an open research direction. A recent, interesting, proposal of a local labelling process is given by Dunson (2008).

5. Empirical study: hDP$_k$ mixtures for spatial data

We discuss the behaviour of our model where x is a spatial co-ordinate and the random curve is a surface. The data $Y_i = (Y_i(x_1), \ldots, Y_i(x_m))$, $i = 1, \ldots, n$, are observations of different surfaces at m sites. We introduce a common mean term $\mu = (\mu(x_1), \ldots, \mu(x_m))$ in model (3), so that

$$Y_i|\mu, \theta_i, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(\mu + \theta_i, \sigma^2 I_m),$$
$$\theta_i|G_{x_1,\ldots,x_m} \overset{\text{IID}}{\sim} G_{x_1,\ldots,x_m}.$$

More generally, the mean term would include the effects of covariates, e.g. $\mu_i = X_i\beta$. We compare two choices for the prior on G—the standard symmetric DP$_k(\alpha, G_{0,x_1,\ldots,x_m})$ and the hybrid

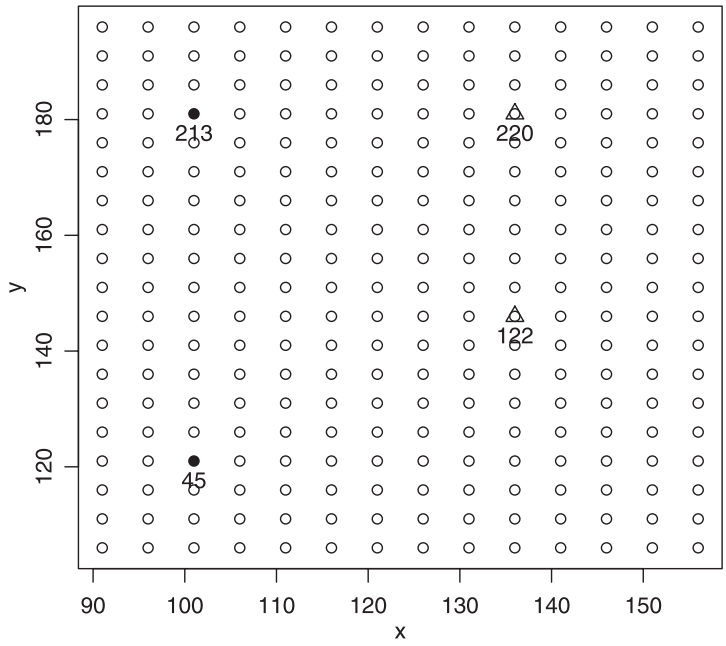

Fig. 3. Regular grid of 14×19 points where we collect measurements for the two applications that are discussed in Section 5 ($N = 266$): the locations that are denoted by '•' are explicitly discussed in Section 5.1; those denoted by 'Δ' are discussed in Section 5.2

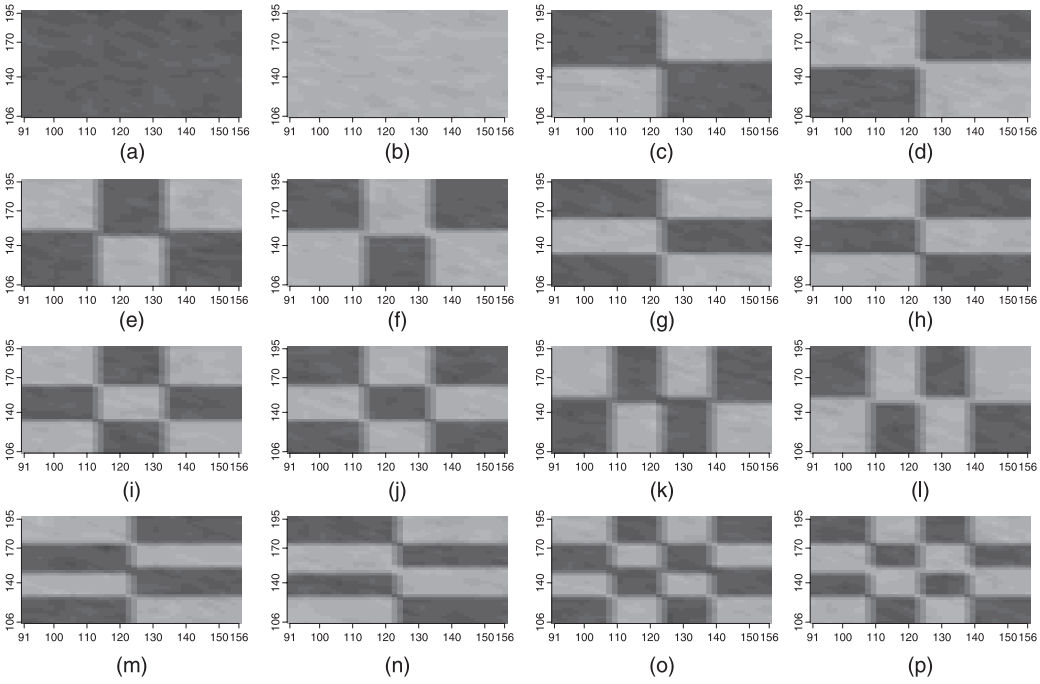

Fig. 4. Image plots of the various observational patterns that were created for the simulation in Section 5.1: (a) $i = 3$; (b) $i = 6$; (c) $i = 9$; (d) $i = 12$; (e) $i = 15$; (f) $i = 18$; (g) $i = 21$; (h) $i = 24$; (i) $i = 27$; (j) $i = 30$; (k) $i = 33$; (l) $i = 36$; (m) $i = 39$; (n) $i = 42$; (o) $i = 45$; (p) $i = 48$

$\mathrm{DP}_k(\alpha q, G_{0,x_1,\ldots,x_m})$. As base measure $G_{0,x_1,\ldots,x_m}$, in both cases we use the finite dimensional distribution of an isotropic Gaussian process with mean 0, variance σ_0^2 and exponential correlation function with decay parameter ϕ_0. The labelling prior is centred on a measure q defined from a Gaussian copula, as in Section 4.2. To facilitate comparison between the models, we fix the mass parameter $\alpha = 1$. The priors for μ, σ^2 and σ_0^2 are chosen following Gelfand *et al.* (2005). The priors for ϕ_0 and ϕ_q are given below.

We expect that the hDP_k mixture model provides a good reconstruction of the curve trajectories with a sensibly smaller number of canonical species than the DP_k mixture. It is worth stressing that both models are location mixtures of Gaussians; the hDP_k mixture model (13) allows also local smoothing, by local selection of the mean vector components, but in both models the measurement error variance is constant along the curves and across the mixture components. As usual with location mixtures, this implies a trade-off between the size of the measurement error and the number of species that are required to describe the data, as we discuss in Section 5.2. Also, an inadequate specification of the base measure G_0 may require more species to describe the sample of curves. However, choosing G_0 degenerate on constant species ($\phi_0 = 0$) may be attractive since such choice facilitates species identifiability. In fact, if

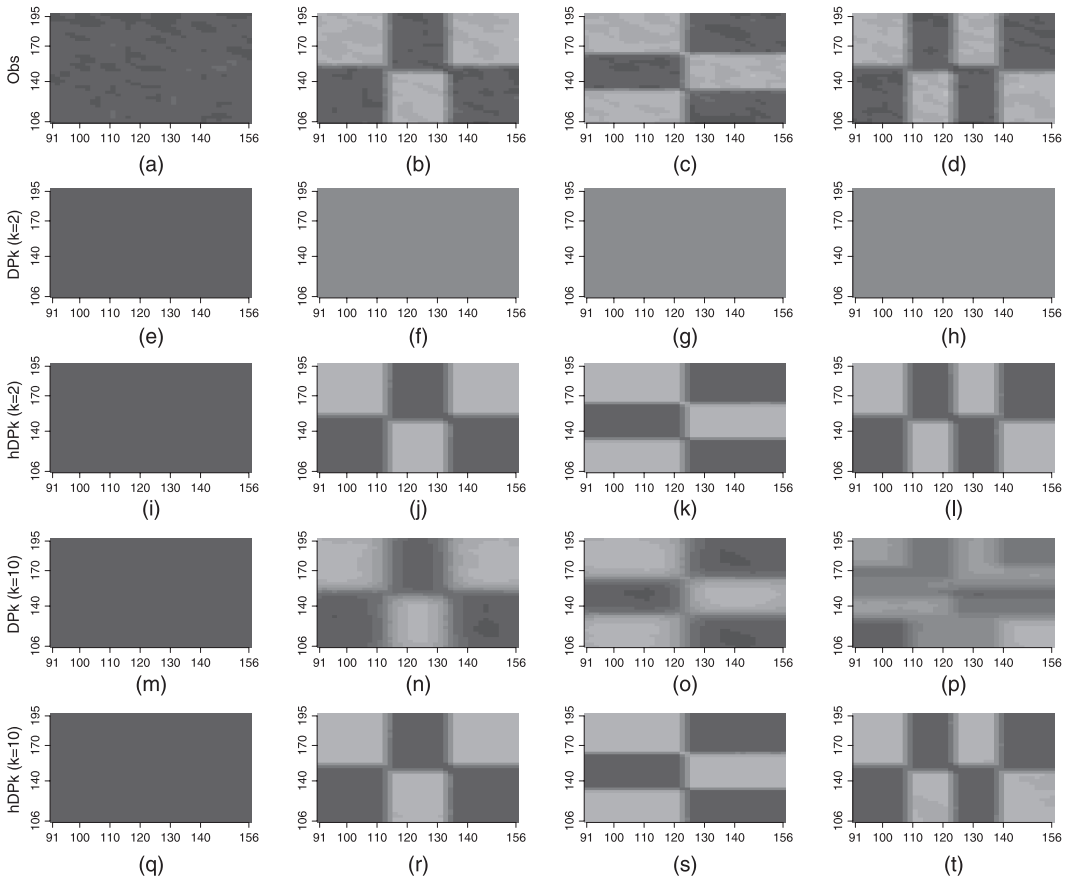

Fig. 5. By rows, image plots of (a)–(d) the simulated data and (e)–(t) posterior predictive means from a DP_k or an hDP_k prior, specifically, (e)–(h) DP_k, $k = 2$, (i)–(l) hDP_k, $k = 2$, (m)–(p) DP_k, $k = 10$, and (q)–(t) hDP_k, $k = 10$: by columns, (a), (e), (i), (m), (q) correspond to observation $i = 3$, (b), (f), (j), (n), (r) $i = 13$, (c), (g), (k), (o), (s) $i = 23$ and (d), (h), (l), (p), (t) $i = 33$

G_0 is too flexible, pure and hybrid species may be confounded, even if the model still succeeds in giving a good reconstruction of the data. Prior knowledge about the nature of the pure species θ_j^* clearly helps identifiability and can be easily incorporated in the prior (see Section 4.2). For example, in modelling possibly abrupt changes in an otherwise smooth surface, e.g. the degree of impairment of different regions in the brain, we may require that each cluster be characterized by realizations of a highly correlated base process. This requirement is easily expressed in the model by centring the prior on ϕ_0 and ϕ_q on small values (recall that a small value of ϕ_q discourages too many local changes). In the examples that follow, however, we use a weak inverse gamma IG(0.5, 1.0) distribution for ϕ_0 and ϕ_q so that the prior mean for both decay parameters is 0.5, which corresponds to an effective range of $\frac{1}{2}$ of the maximum intersites distance that is observed in our data set.

In our model, k is the number of species in the population. Again, we expect the hDP_k to succeed in reconstructing the data with small k, whereas the DP_k should require many more

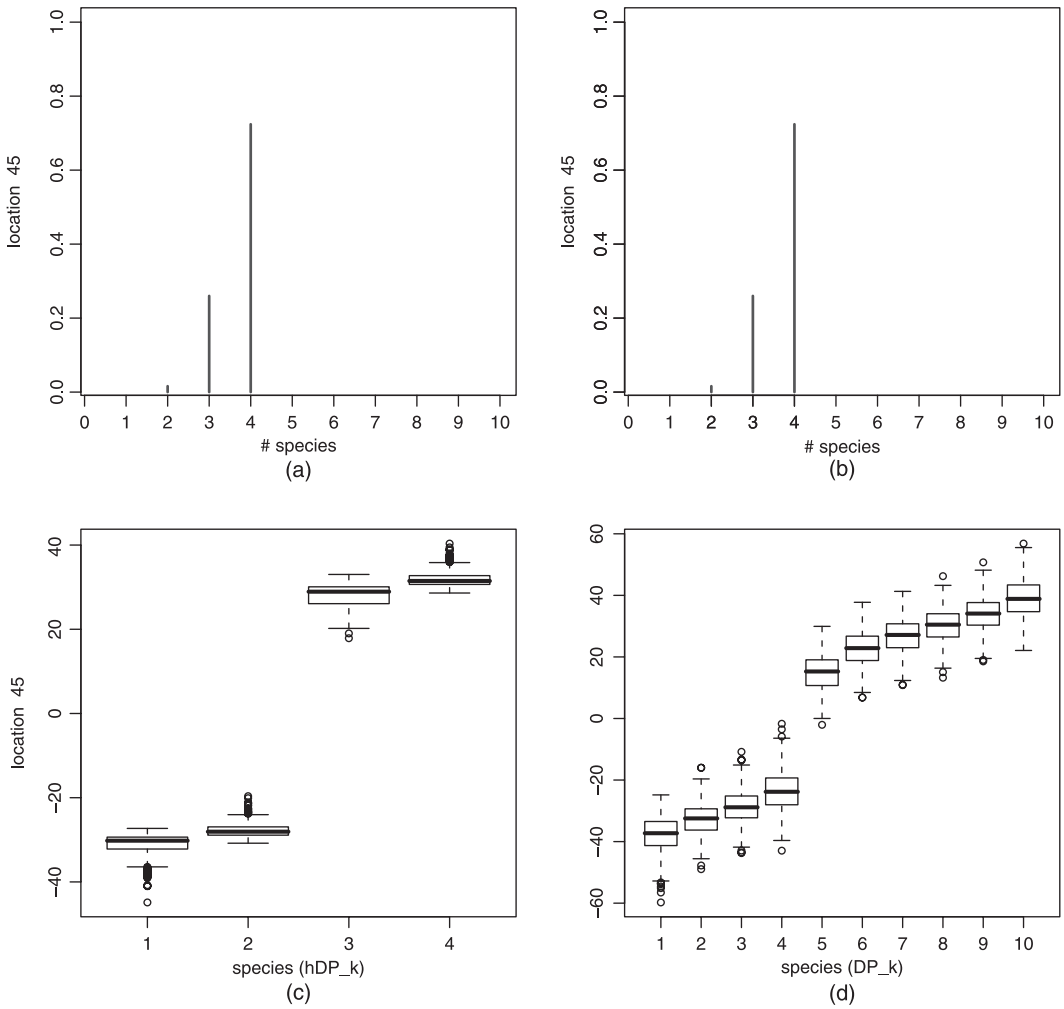

Fig. 6. Posterior number of clusters at (a) x_{213} and (b) x_{45} under an hDP_k ($k = 10$), for the application in Section 5.1, and boxplots of the ordered values of $\theta^*(x_{45})$ corresponding to $d(x_{45})$, for (c) the hDP_k and (d) DP_k ($k = 10$)

'available' species. We do not study the case where k is random but note that the number of species needed to describe the sample (or 'discovered' species) is random, with a prior which is implicitly given by the prior on G. Rather than a random k, we studied the behaviour of the model for increasing k (Section 3.3). With the copula labelling prior used in the next examples, for large k, the hDP_k tends to encourage more species. Note that part (b) of theorem 2 applies, since here we took q with uniform marginals, to have a direct comparison with the usual symmetric DP_k mixtures (one can easily modify the symmetric copula construction by changing the hypercube partition). In this case, when k increases the hDP_k tends to $\text{DP}(\alpha G_q)$, whereas the DP_k tends to $\text{DP}(\alpha G_0)$, so the behaviour that we obtain in the next examples is not unexpected. The form of the base measure in the limit DP for the hybrid Dirichlet mixture still enables better modelling of the presence of irregular areas on the surfaces than the $\text{DP}(\alpha G_0)$, and the resulting description of the data remains more parsimonious.

Computations are implemented by using the MCMC algorithm that was described in Section 4.3 and detailed in section A.2 of the appendix. Here, we set the jitter variance $\eta^2 = 0.01$.

5.1. Simulated data

To illustrate the behaviour of our model, we first investigate a simulated data set. The data set can be viewed as a toy example for the brain magnetic resonance imaging data that are analysed

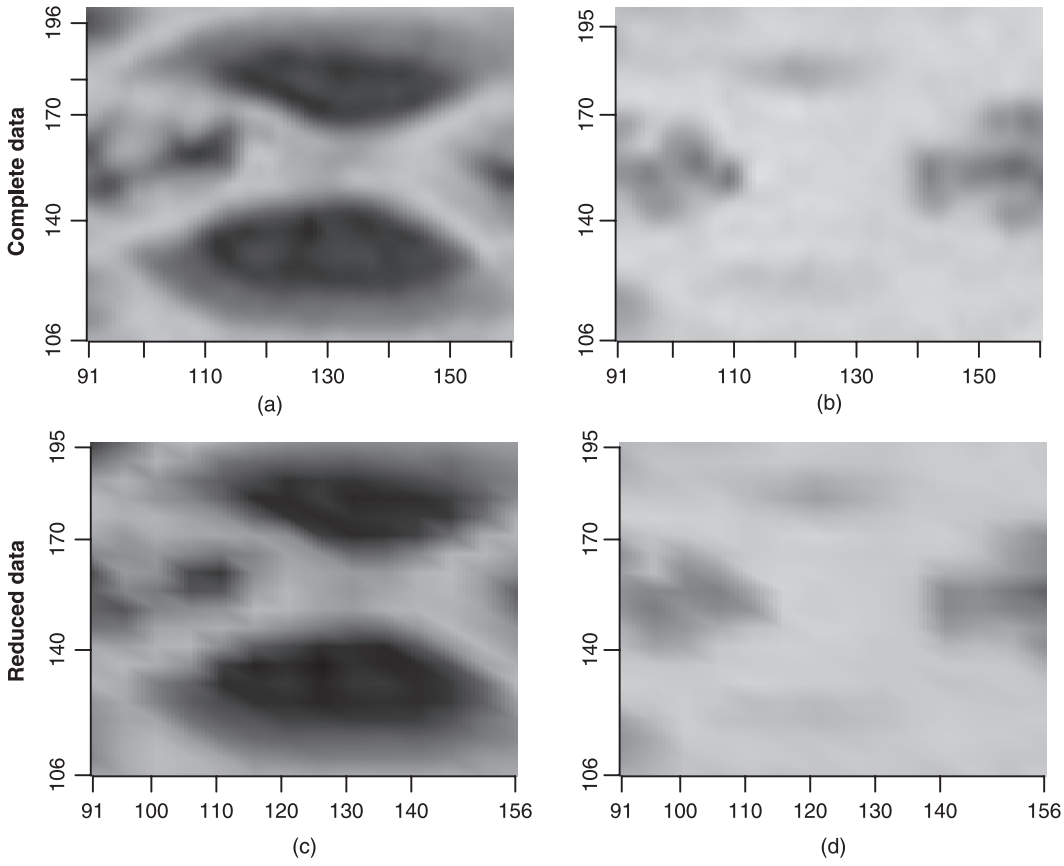

Fig. 7. Effect of the data reduction: image plots of the original data set for (a) $i = 6$ and (b) $i = 12$, and image plots of the measurements that were collected on the grid in Fig. 3 for (c) $i = 6$ and (d) $i = 12$

in Section 5.2. In fact, the observations are generated on the same regular grid of 14×19 points (Fig. 3) that is considered for that application. For each site x and each replicate i, we observe a noisy realization of either one of two base (or 'pure') processes. To be more specific, we take a common mean $\mu = 90$ and generate two independent species $\theta_j^* \sim \mathcal{N}\{m_j, \tau^2 R(\phi)\}$, $j = 1, 2$, with $m_1 = -30$ and $m_2 = 30$, $\tau = 3$ and $\phi = 0.5$. Then, we create a sample of noisy hybrid surfaces $Y_i(x) = \mu + \theta_j^*(x) + \varepsilon_i(x)$, $\varepsilon_i(x) \sim^{\text{IID}} \mathcal{N}(0, \sigma^2 = 3)$, with $j = 1$ or $j = 2$ according to the values of the co-ordinates of x and the replicate i. More specifically, we generate $n = 48$ replicates, three belonging to each of the 16 different patterns ('checkerboards') that are illustrated in Fig. 4. Note that these data are not simulated from the model since the species θ_1^* and θ_2^* are not IID, and the 'hybridization' and the clustering are arbitrarily chosen, rather than being induced by the prior as in Section 4.2. Recall that here we fix $\alpha = 1$.

We compare the behaviour of the DP_k and hDP_k for $k = 2$ and $k = 10$. As expected, with $k = 2$, the DP_k relies only on $k = 2$ 'global' species and cannot recover the full pattern of the data. In contrast, the hDP_k captures well the observed pattern by allowing local choices of the relevant process. When $k = 10$, the DP_k struggles to recover complex patterns fully. See Fig. 5, where we show the image plots of the original observations and the posterior means estimated by the two models, for $k = 2$ and $k = 10$.

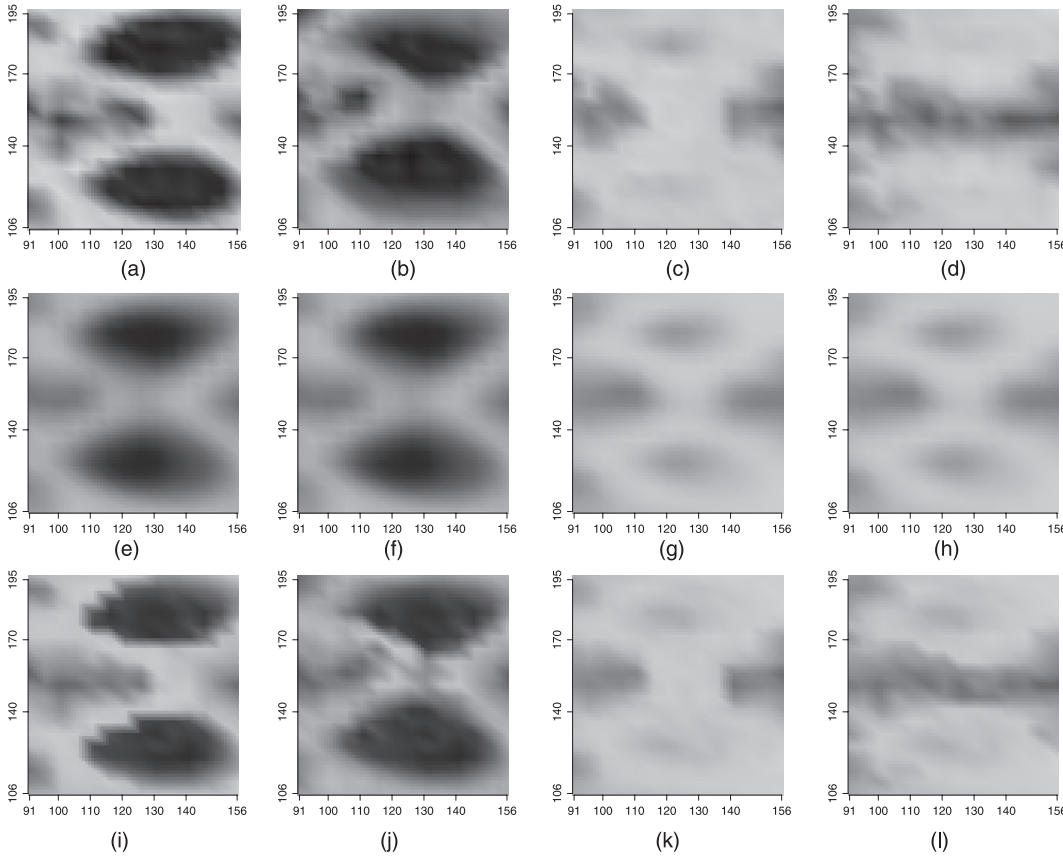

Fig. 8. By rows, (a)–(d) image plots of observations and posterior predictive means for (e)–(h) the DP_k prior and (i)–(l) the hDP_k prior, with $k = 2$, for some individuals considered in the application in Section 5.2: specifically, by columns, (a), (e), (i) $i = 2$, (b), (f), (j) $i = 6$, (c), (g), (k) $i = 12$ and (d), (h), (l) $i = 17$

If we fix k sufficiently large, the DP_k (and the DP) roughly identifies 16 clusters; hence, the fit is greatly improved. However, the hDP_k can anyway provide a simpler description of the data. Figs 6 (a) and 6(b) show the posterior distribution on the random partition of the replicates that is implied by the DP_k and hDP_k, for $k = 10$. Of course, the random partition at a single location x is determined by the posterior of $\theta_1(x), \ldots, \theta_n(x)$ and it is characterized by the number $d_n(x)$ of distinct values $\theta_1^*(x), \ldots, \theta_{d_n(x)}^*$ observed at x, together with the size and composition of the groups. We show the posterior of $d_n(x)$ at two locations, x_{213} and x_{45}, under the hDP_k. The corresponding figure for the DP_k (which is not shown) insists on $d_n(x) = k = 10$ species. For the hDP_k, the posterior modes of $d_n(x_{213})$ and $d_n(x_{45})$ are respectively 3 and 4, suggesting that the data might support more variety in the pure species. In fact, this reflects the variability of the data around the two pure species as well as the choice of the hyperparameters, in particular α and the standard prior that is used for ϕ_0 and ϕ_q. Figs 6(c) and 6(d) report the boxplot of the ordered values of the distinct $\theta^*(x)$s for those iterations where $d_n(x)$ is equal to the posterior mode (for both models). Here, $x = x_{45}$. It is evident that the hDP_k supports the existence of only two species.

5.2. Brain magnetic resonance imaging images

Alzheimer's disease is a neurodegenerative disease, that induces hippocampal atrophy in the brain (Ashburner *et al.*, 2003). Magnetic resonance imaging images of the brain of patients who are affected by the disease show impaired as well as normal regions. The statistical analysis of brain images is a notoriously difficult task, primarily because of the large amount of data and the evident non-stationarity of the spatial processes that are involved. In particular, regions that are far apart in the brain might show higher correlations than neighbouring regions. For a discussion of this issue and a solution from a Bayesian perspective, see, for example, DuBois Bowman (2005).

We analyse magnetic resonance imaging data from 17 patients. The data have been provided

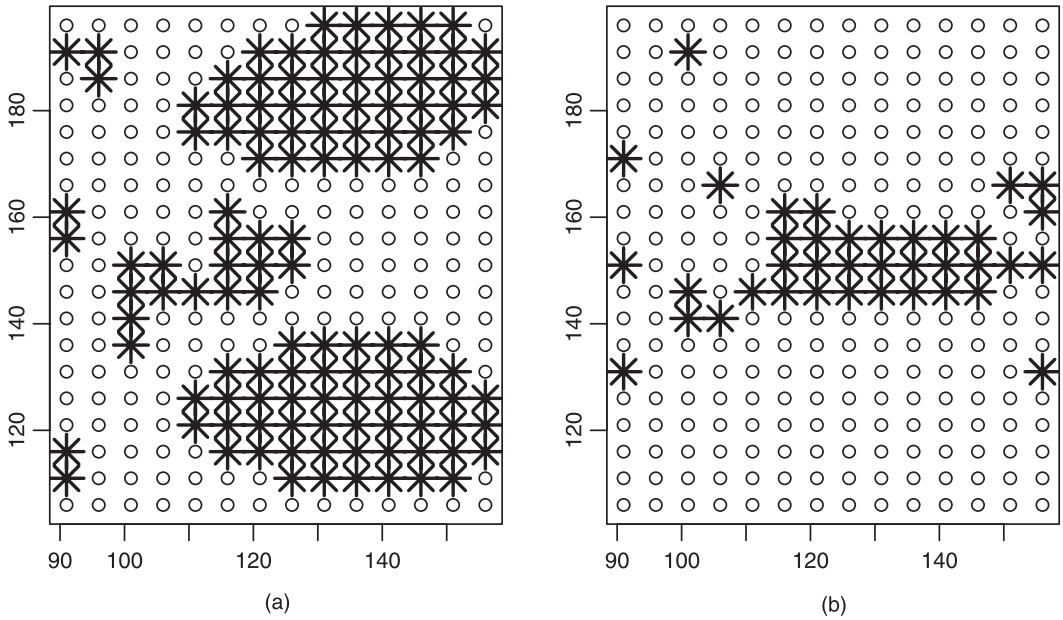

Fig. 9. Posterior probability maps of pixel impairment for individuals (a) $i = 2$ and (b) $i = 17$: see Section 5.2

by the Laboratory of Epidemiology and Neuroimaging, Centro San Giovanni di Dio-Fatebenefratelli, Brescia, Italy, and have been previously normalized by using the freely available SPM5 software (`http://www.fil.ion.ucl.ac.uk/spm/`; see Worsley and Friston (1995)). For simplicity, we reduce the original data set and record grey density matter intensity only on a regular two-dimensional grid of 14×19 pixels encompassing the hippocampus. The data are treated as continuous and point referenced. Image plots of the full and reduced data set for two patients are presented in Fig. 7; the reduction of the data to a smaller grid does not seem to alter significantly the perception of the main features of the biological processes affecting the brain.

Ideally, we would like to be able to capture the action of two processes: one describing the features of healthy individuals and another characterizing the impaired regions of the brain. However, since we consider a location mixture of Gaussian distributions, more than two species

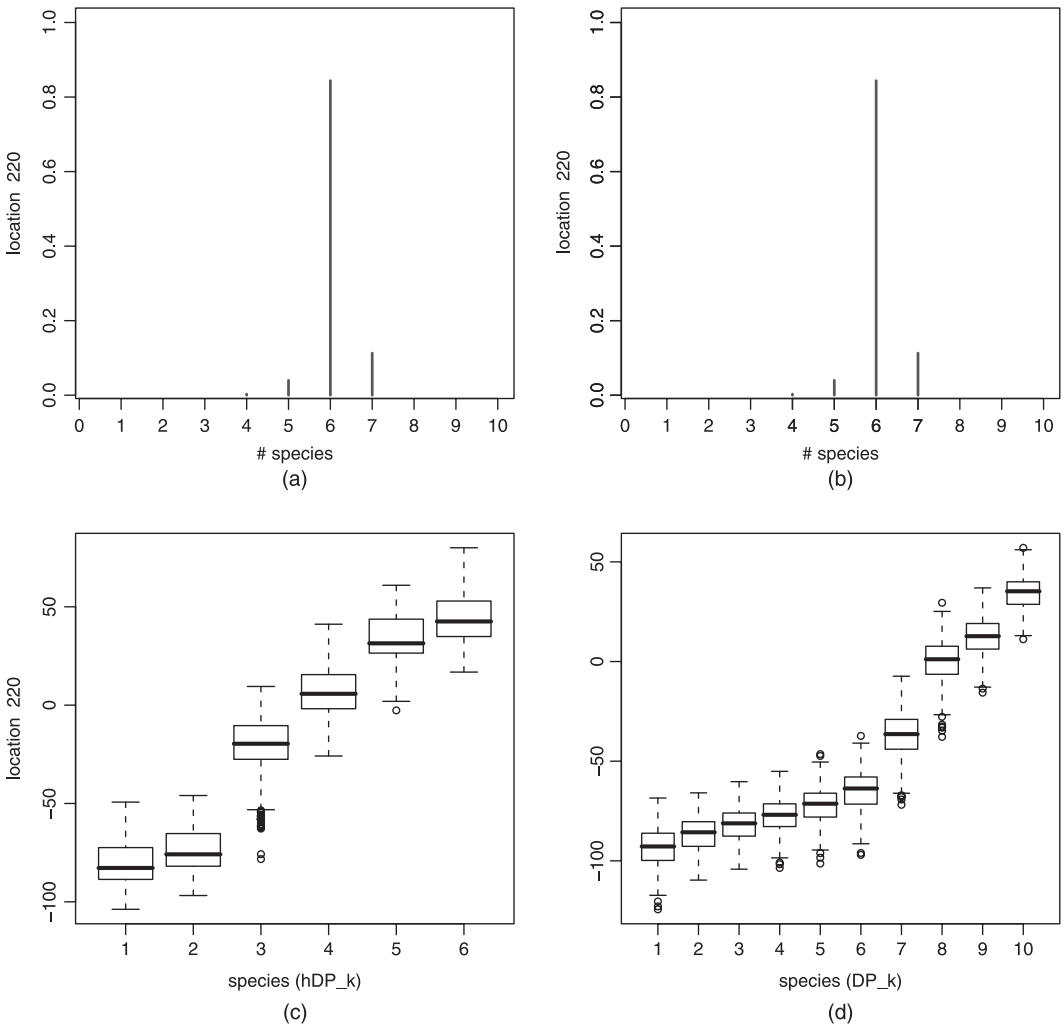

Fig. 10. Posterior number of clusters at (a) x_{122} and (b) x_{220} under an hDP$_k$ ($k = 10$), for the application in Section 5.2, and boxplots of the ordered values of $\theta^*(x_{220})$ corresponding to $d(x_{220}) = 6$, for (c) the hDP$_k$ and (d) DP$_k$ ($k = 10$)

(i.e. kernels) may be required for reconstructing the data if the variability inside the healthy and diseased groups is different and/or larger than the measurement error that is explained by σ^2.

In Fig. 8, for four patients, we show the image plots of the observations and the posterior predictive means for the DP_k and hDP_k models, $k=2$. It is evident that, with small k, the hDP_k captures the features of the data better; as a result, we have evidence to support a model of two types of regions: normal and impaired. In general, the hDP_k describes the existence of two clearly separate processes, whereas the DP_k tends to overshrink the estimates for observations that are quite dispersed. Fig. 9 shows a posterior probability map of pixel impairment for two individuals. For illustration, here we consider a pixel impaired if $p[\theta_i(x) = \min_j\{\theta_j^*(x)\}|\mathrm{data}] > 0.7$. Posterior probability maps of this kind, jointly with an appropriate loss function, may be used to define multicomparison procedures in a coherent Bayesian framework (Friston and Penny, 2003; Müller *et al.*, 2004).

More generally, it may be imagined that different levels of impairment arise for different regions, thus suggesting the existence of $k > 2$ groups. As in the simulated data set, the DP_k and the DP models tend to recognize as many species as the number of replicates (or number of available species, whichever is less), thus defeating any dimension reduction purposes. However, for the hDP_k, when $k = 10$, Fig. 10 reports the posterior distribution of the number of species selected at x_{122} and x_{220} and the boxplots of the ordered θ^* in x_{220}. Unlike the DP_k, the hDP_k does suggest the existence of species (surfaces) corresponding to different levels of impairment.

6. Final remarks

In the context of functional data, we have generalized the functional DP_k and DP priors to versions that enable more parsimonious representations in terms of species of curves, accounting for global and local heterogeneity in the data. We have illustrated our method with spatial data, but more general applications in the area of supervised learning can be envisaged. Examples may include the analysis of time course gene expression data in bioinformatics or clustering of time series in econometrics. We can also consider clustering objects other than curves, e.g. shapes or sets in two or three dimensions. We can imagine adding a dynamic aspect to our specifications, considering the curves to be temporally evolving rather than conditionally independent. Further promising interchanges are with fields such as population genetics, where our construction of dependent local partitions may allow modelling of the partial evolution of multivariate or functional processes. On the methodological side, further development of labelling measures would be useful. We gave a notion of hybrid species priors which extends species sampling models. As briefly discussed in Section 4.4, an open direction of research is to study appropriate specifications of the labelling measure to facilitate the characterization of the hybrid species sampling models in terms of the predictive rule.

Acknowledgements

The authors thank Patrizia Berti and Pietro Rigo for helpful discussions about Section 3.3, and Giovanni B. Frisoni, Lorena Bresciani and Michela Pievani at the Centro San Giovanni di Dio, Brescia, Italy, for providing the data in Section 5.2 and helpful support. The authors also acknowledge useful comments of two reviewers and the Associate Editor which led to a much improved presentation. The research of the third author was supported in part by National Science Foundation award DMS 0505085 and the first author's by Ministero dell' Università e della Ricerca Scientifica grant PRIN 2006131039.

References

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

Ashburner, J., Csernansky, J. G., Davatzikos, C., Fox, N. C., Frisoni, G. B. and Thompson, P. M. (2003) Computer-assisted imaging to assess brain structure in healthy and diseased brains. *Lancet Neurol.*, **2**, 79–88.

Berti, P., Pratelli, L. and Rigo, P. (2006) Almost sure convergence of random probability measures. *Stochastics*, **78**, 91–97.

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

Bush, C. A. and MacEachern, S. N. (1996) A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **83**, 275–285.

Duan, J. A., Guindani, M. and Gelfand, A. E. (2007) Generalized spatial Dirichlet process models. *Biometrika*, **94**, 809–825.

DuBois Bowman, F. (2005) Spatio-temporal modelling of localized brain activity. *Biostatistics*, **6**, 558–575.

Dunson, D. B. (2008) Nonparametric Bayes local partition models for random effects. *Biometrika*, to be published.

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

Dunson, D. B., Xue, Y. and Carin, L. (2008) The matrix stick-breaking process: flexible Bayes meta analysis. *J. Am. Statist. Ass.*, **103**, 317–327.

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

Friston, K. J. and Penny, W. (2003) Posterior probability maps and SPMs. *Neuroimage*, **19**, 1240–1249.

Gelfand, A., Kottas, A. and MacEachern, S. (2005) Bayesian nonparametric spatial modeling with Dirichlet processes mixing. *J. Am. Statist. Ass.*, **100**, 1021–1035.

Green, P. J. and Richardson, S. (2002) Hidden Markov models and disease mapping. *J. Am. Statist. Ass.*, **97**, 1055–1070.

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

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

Ishwaran, H. and James, L. F. (2003) Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhya* A, **65**, 577–592.

Ishwaran, H. and Zarepour, M. (2002) Dirichlet prior sieves in finite Normal mixtures. *Statist. Sin.*, **12**, 941–963.

Kingman, J. F. C. (1975) Random discrete distributions (with discussion). *J. R. Statist. Soc.* B, **37**, 1–22.

MacEachern, S. N. (1999) Dependent nonparametric process. *Proc. Bayesn Statist. Sci. Sect. Am. Statist. Ass.*, 50–55.

MacEachern, S. N. (2001) Decision theoretic aspects of dependent nonparametric processes. In *Bayesian Methods with Applications to Science, Policy, and Official Statistics* (ed. E. George), pp. 551–560. International Society for Bayesian Analysis.

MacEachern, S. N. (2007) Discussion of "Bayesian nonparametric modelling for spatial data using Dirichlet processes", by Gelfand, Guindani and Petrone. In *Bayesian Statistics 8* (eds J. M. Bernardo, J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.

Marin, J. M. and Robert, C. P. (2007) *Bayesian Approach: a Practical Approach to Computational Bayesian Statistics*. Berlin: Springer.

Muliere, P. and Secchi, P. (1995) A note on a proper Bayesian bootstrap. *Technical Report*. Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, Pavia.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Am. Statist. Ass.*, **99**, 990–1001.

Neal, R. M. (1997) Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *Technical Report 9702*. Department of Statistics, University of Toronto, Toronto. (Available from `http://www.kernel-machines.org/papers/upload_11071_mc-gp.ps`.)

Oakley, J. and O'Hagan, A. (2002) Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, **89**, 769–784.

Pitman, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory; Papers in Honor of David Blackwell* (eds T. S. Ferguson, L. S. Shapley and J. B. MacQueen), pp. 245–267. Hayward: Institute of Mathematical Statistics.

Ramamoorthi, R. V. and Sangalli, L. (2006) On a characterization of Dirichlet distribution. In *Bayesian Statistics and Its Applications* (eds S. K. Upadhyay, U. Singh and D. K. Dey), pp. 385–397. Varanasi: Anshan.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*, 2nd edn. New York: Springer.

Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Ray, S. and Mallick, B. (2006) Functional clustering by Bayesian wavelet methods. *J. R. Statist. Soc.* B, **68**, 305–332.

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008) Latent stick-breaking processes. To be published.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.
Shi, J. Q. and Wang, B. (2008) Curve prediction and clustering with mixtures of Gaussian process functional regression models. *Statist. Comput.*, **18**, 267–283.
Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical Dirichlet processes. *J. Am. Statist. Ass.*, **101**, 1566–1581.
Worsley, K. J. and Friston, K. J. (1995) Analysis of fMRI time-series revisited—again. *Neuroimage*, **2**, 173–181.