

NONPARAMETRIC FUNCTIONAL DATA ANALYSIS THROUGH BAYESIAN DENSITY ESTIMATION

ABEL RODRÍGUEZ, DAVID B. DUNSON, AND ALAN E. GELFAND

ABSTRACT. In many modern experimental settings, observations are obtained in the form of functions, and interest focuses on inferences on a collection of such functions. Some examples are conductivity-temperature-depth (CTD) data in oceanography, dose-response models in epidemiology and time-course microarray experiments in biology and medicine. In this paper we propose a hierarchical model that allows us to simultaneously estimate multiple curves nonparametrically by using dependent Dirichlet Process mixtures of Gaussians to characterize the joint distribution of predictors and outcomes. Function estimates are then induced through the conditional distribution of the outcome given the predictors. The resulting approach allows for flexible estimation and clustering, while borrowing information across curves. We also show that the function estimates we obtain are consistent on the space of integrable functions. As an illustration, we consider an application to the analysis of CTD data in the north Atlantic.

¹Abel Rodriguez is Ph.D. candidate, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708, abel@isds.duke.edu. David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Science, P.O. Box 12233, RTP, NC 27709, dunson1@niehs.nih.gov. Alan E. Gelfand is James B. Duke professor, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708, alan@isds.duke.edu. This work was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Key words and phrases. Nonparametric regressions; Consistency; Functional clustering; Dependent Dirichlet process; Nonparametric Bayes; Random probability measure.

This work was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Science. The authors would like to thank Susan Lozier and Robert Wolpert for helpful comments.

1. INTRODUCTION

Current scientific interest goes beyond estimating and comparing parameters among populations. In many cases, interest lies on the functional relationships between variables, and how these change under different experimental conditions. That is, given pairs $\{(\mathbf{y}_{ij}, \mathbf{x}_{ij})\}$ where $j = 1, \dots, J$ indexes an experimental condition and $i = 1, \dots, n_j$ indexes an observation within the experiment, $\mathbf{y}_{ij} \in \mathbb{R}^q$ and $\mathbf{x}_{ij} \in \mathbb{R}^p$, we are interested in 1) jointly estimating functions $f_1, \dots, f_J : \mathbb{R}^p \rightarrow \mathbb{R}^q$ that describe the relationship between predictors and outcomes; 2) testing hypotheses about the dependence between the functions; and 3) predicting the function under new experimental conditions. Depending on the application at hand, these functions might correspond to the conditional mean responses, quantiles of the conditional distributions, or even the conditional variances; while the inference goal might be multiple comparison of curves, functional clustering or spatial prediction of the functional relationship. In this paper, we develop a class of models that can tackle such joint inference problems from a Bayesian nonparametric perspective.

Popular approaches for nonparametric functional estimation can be broadly divided in three main groups. One simple yet powerful alternative is kernel regression methods. These methods represent the unknown function as a linear combination of the observed values of the outcome variable, using covariate-based weights (Altman, 1992; Chu and Marron, 1991; Fan et al., 1995). Another class of methods assumes that the functions of interest can be represented as a linear combination of basis functions. The problem of estimating the function reduces to estimation of the basis coefficients. Splines, wavelets and reproducing kernel methods fall in this broad category (Vidakovic, 1999; Truong et al., 2005). A third alternative is to assume that the functions in question are realizations of stochastic processes, with the Gaussian process (GP) being a common choice (Rasmussen and Williams, 2006).

Different approaches have been used to extend these methodologies to collections of functions. For example, when the function of interest is modeled as a linear combination of basis functions, hierarchical models on the basis coefficients can be used to accommodate different types of dependence. This approach has been successfully exploited by authors such as Rice and Silverman (1991); Wang (1998); Guo (2002); Wu and Zhang (2002) and Morris and Carroll (2006) to construct ANOVA and random effect models for curves. Along similar lines, Bigelow and Dunson (2005) and Ray and Mallick (2006) have used Dirichlet process priors as part of the hierarchical specification of the model in order to induce clustering across curves. Behseta et al. (2005) develops a hierarchical Gaussian process (GP) model, which treats individual curves as realizations of a GP centered on a GP mean function.

These methods are based on specifications for the conditional distributions $p_1(\mathbf{y}|\mathbf{x}), \dots, p_J(\mathbf{y}|\mathbf{x})$, where $p_j(\mathbf{y}|\mathbf{x})$ denotes the distribution of the outcome \mathbf{y} given the predictor \mathbf{x} under experimental condition j . In this paper, we consider a completely different approach. Instead of modeling the conditional distributions directly, we induce a prior on the space of functions indirectly through a model on the collection of joint distributions $p_1(\mathbf{y}, \mathbf{x}), \dots, p_J(\mathbf{y}, \mathbf{x})$ that uses mixtures of dependent Dirichlet processes (MacEachern, 2000; DeIorio et al., 2004; Gelfand et al., 2005; Rodriguez et al., 2006). This method is conceptually related to the double kernel method of Fan et al. (1996); Fan and Yim (2004), which induces a frequentist conditional density estimate through multivariate density estimation. However, we focus on a Bayesian approach, generalizing the method of Müller et al. (1996) to a setting where multiple dependent curves are of interest. The model induces a rich error structure for the conditional distributions, accommodating non-Gaussian and heteroscedastic errors. Function estimates reduce to kernel-weighted mixtures of linear models, where the location and variances of the kernels are automatically chosen. Our method provides domain adaptive smoothing for each curve

while avoiding an arbitrary choice of basis functions or the use of complicated and inefficient MCMC algorithms typically required for adaptive function estimation.

As we obtain a joint posterior distribution for the full conditional response distributions, we can conduct inferences on regression functions characterized in terms of the mean, a quantile or even the variance. In addition, multivariate responses and predictors can be accommodated without complications, while also allowing interactions in a flexible manner. Under fairly general conditions, the method produces consistent estimates on a dense subset of the space of integrable functions on compact subsets of \mathbb{R}^p . As an illustration, we focus on functional clustering applications using the nested Dirichlet process (Rodriguez et al., 2006) as a building block in our model. Functional clustering has become popular as a hypothesis generating mechanism. For example, in the analysis of time-course expression experiments (Ramoni et al., 2002; Luan and Li, 2003; Wakefield et al., 2003), functional clustering is used to identify coregulated genes, which are typically assumed to be members of a common transcription pathway.

Section 2 reviews the definition of the Dirichlet process and the single-curve model of Müller et al. (1996). In section 3 we introduce our method for multiple curves and discuss properties. In section 4 we give conditions for posterior consistency of these models, providing theoretical support for our approach. Section 5 illustrates the approach through application to temperature profile data in the North Atlantic, and section 6 contains a discussion.

2. SINGLE-FUNCTION NONPARAMETRIC REGRESSION

2.1. The Dirichlet process. Let $(\mathcal{X}, \mathcal{B})$ be a complete and separable metric space (typically $\mathcal{X} = \mathbb{R}^p$ and \mathcal{B} are the Borel sets on \mathcal{X}), and let $H \in \mathcal{H}$ be a probability measure on $(\mathcal{X}, \mathcal{B})$. A Dirichlet process (DP) (Ferguson, 1973, 1974) with baseline measure H_0 and precision α , denoted $\text{DP}(\alpha H_0)$, defines a probability measure on the space \mathcal{H} , such that $(H(B_1), \dots, H(B_L)) \sim \text{Dir}(\alpha H_0(B_1), \dots, \alpha H_0(B_L))$

for any partition B_1, \dots, B_L of \mathcal{X} . The Dirichlet Process can also be defined as a stick-breaking prior.

Let $\pi_i \sim \text{Beta}(1, \alpha)$ and $\boldsymbol{\eta}_i \sim H_0$, for all i be independent. If H is defined as

$$(1) \quad H(\cdot) = \sum_{i=1}^{\infty} w_i \delta_{\boldsymbol{\eta}_i}(\cdot),$$

where $w_i = \pi_i \prod_{j=1}^{i-1} (1 - \pi_j)$ and $\delta_x(\cdot)$ represents a degenerate distribution at x , then $H \sim \text{DP}(\alpha H_0)$ (Sethuraman, 1994). This characterization shows that H is almost surely discrete, making the Dirichlet process an unappealing model for continuous data.

An alternative is to consider Dirichlet process mixture (DPM) models (Escobar, 1994; Escobar and West, 1995), where the Dirichlet process is used as the prior on the random mixing distribution over the parameters of a continuous distribution,

$$(2) \quad \mathbf{z} \sim g(\cdot) \quad g(\cdot) = \int k(\cdot|\boldsymbol{\eta})H(d\boldsymbol{\eta}) \quad H \sim \text{DP}(\alpha H_0).$$

Therefore, the DPM induces a prior on g indirectly through a prior on the mixing distribution H . A popular choice is the DPM of Gaussian distributions, where $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $k(\cdot|\boldsymbol{\eta}) = \phi_p(\cdot|\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is the p -variate normal kernel with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$.

Given an iid sample $\mathbf{z}^n = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, the posterior of the mixing distribution, $H^n(\cdot|\mathbf{z}^n)$, is distributed as a mixture of Dirichlet processes (MDP) (Antoniak, 1974),

$$H^n(\cdot|\mathbf{z}^n) \sim \int \text{DP} \left(\alpha H_0 + \sum_{i=1}^n \delta_{\boldsymbol{\theta}_i} \right) p(d\boldsymbol{\theta}_1, \dots, d\boldsymbol{\theta}_n | \mathbf{z}^n)$$

and the optimal density estimator under squared error loss, $g^n(\mathbf{z})$, is the posterior predictive distribution

$$(3) \quad \begin{aligned} g^n(\mathbf{z}) &= \mathbb{E}_{H^n} \left[\int k(\mathbf{z}|\boldsymbol{\eta})H^n(d\boldsymbol{\eta}|\mathbf{z}^n) \right] \\ &= \int k(\mathbf{z}|\boldsymbol{\eta})H_0^n(d\boldsymbol{\eta}|\mathbf{z}^n), \end{aligned}$$

where

$$H_0^n(\cdot|\mathbf{z}^n) = \int \frac{\alpha H_0(\cdot) + \sum_{i=1}^n \delta_{\boldsymbol{\theta}_i}(\cdot)}{\alpha + n} p(d\boldsymbol{\theta}_1, \dots, d\boldsymbol{\theta}_n | \mathbf{z}^n)$$

is the posterior mean of the mixing distribution.

Computation for DPM models is typically carried out using one of three different approaches: Pólya urn schemes that marginalize out the unknown distribution H (MacEachern, 1994; Escobar and West, 1995; Bush and MacEachern, 1996), truncation methods that use finite mixture models to approximate the DP (Ishwaran and James, 2001; Green and Richardson, 2001), and Reversible Jump algorithms (Green and Richardson, 2001; Jain and Neal, 2000; Dahl, 2003).

2.2. Nonparametric regression through Dirichlet process mixtures. Consider the following application of the model for multivariate density estimation described in (2), which, in recognition of Müller et al. (1996), we will call the MEW model:

$$\begin{aligned}
 \mathbf{z}_i &= (\mathbf{y}_i, \mathbf{x}_i) \sim \mathbf{N}_{p+q}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) & i = 1, \dots, n \\
 (\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) &\sim H & H \sim \text{DP}(\alpha H_0) \\
 (4) \quad H_0 &= \text{NIW}_{p+q}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0) & \alpha \sim \text{G}(a_\alpha, b_\alpha) \\
 \boldsymbol{\theta}_0 &\sim \mathbf{N}_{p+q}(\boldsymbol{\theta}_{00}, \mathbf{D}_{00}) & \boldsymbol{\Sigma}_0 \sim \text{W}_{p+q}(\gamma, \boldsymbol{\Sigma}_{00}) \\
 \kappa_0 &\sim \text{G}(a_\kappa, b_\kappa),
 \end{aligned}$$

where NIW_p denotes the p -variate Normal-Inverse-Wishart distribution, G denotes the gamma distribution, W denotes the p -variate Wishart distribution (see appendix A for details on the parametrization of these densities), and the parameters at the top level are partitioned as

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_x) \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}.$$

In this model, hyperpriors on the parameters of the baseline measure H_0 and the precision parameter α have been incorporated to make the DPM more flexible and borrow information parametrically across components. Müller et al. (1996) proposed a slight variant of this model in order to indirectly induce a prior on a mean regression function, $f(\mathbf{x}) = \mathbb{E}(\mathbf{y}|\mathbf{x})$. From the density estimate for the joint distribution $g^n(\mathbf{z})$ described in (3), a posterior estimate for the conditional density can be obtained as

$$(5) \quad g^n(\mathbf{y}|\mathbf{x}) = \int \frac{\phi_q(\mathbf{y}|\boldsymbol{\theta}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\theta}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy})\phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})}{\int \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n)} H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n).$$

In turn, a nonparametric estimate $f^n(\mathbf{x})$ of the mean regression function of \mathbf{y} on \mathbf{x} , $f(\mathbf{x})$, can be obtained from (5) by calculating the conditional expectation,

$$(6) \quad f^n(\mathbf{x}) = \mathbb{E}(\mathbf{y}|\mathbf{x}, \mathbf{z}^n) = \int \frac{(\boldsymbol{\theta}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\theta}_x))\phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})}{\int \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n)} H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n).$$

For any fixed \mathbf{x} , the conditional distribution in (5) is a locally weighted mixture of normals, with the conditional expectation in (6) reducing to a local mixture of linear functions. This rich structure allows for heteroscedastic and non-Gaussian errors, as well as for very flexible mean functions. Indeed, we show in section 4 that any integrable function on a compact set can be arbitrarily well approximated by the functions arising from this model. The location and variance of the kernels are automatically chosen by the model according to the marginal distribution of the predictor variables. Therefore, the model provides local adaptive smoothing, while avoiding awkward choices typical in other methods based on basis expansions or Gaussian processes.

Note that, as $\alpha \rightarrow 0$, the prior on H becomes a single point mass with probability one, and the model in (4) reduces to a normal linear regression model. Hence, since the linear parametric model is nested within our specification, we can test the parametric model against a nonparametric alternative

by examining the posterior probability of a single component in the mixture. This avoids the need for specially tailored MCMC algorithms, such as the method of Basu and Chib (2003).

By concentrating on other summaries of the conditional posterior distribution, the model can also be used for quantile or variance regression. In addition, it can be readily extended to accommodate categorical outcomes and predictors by incorporating latent variables as in Albert and Chib (1993), resulting in a model that simultaneously incorporates a nonparametric regression function and a nonparametric link function.

3. HIERARCHICAL NONPARAMETRIC MODELS FOR FUNCTIONS

Section 2 described a flexible Bayesian model for a single random curve. Simultaneous inference on multiple curves can be accommodated using a similar construction; however, instead of a prior on a single multivariate distribution, we need to construct a prior on a collection of multivariate distributions. Dependence between distributions translates into dependence between the random curves. This section starts by reviewing models for collections of distributions based on the Dirichlet process, and then shows how these models can be used for multiple nonparametric regression in different settings.

3.1. Dependent Dirichlet processes. Much of the recent interest in Bayesian nonparametrics has focused on models for *collections* of distributions. Most extensions of the Dirichlet process achieve dependence through one of two strategies: either by forming convex combinations of independent processes (Müller et al., 2004; Dunson et al., 2007; Griffin and Steel, 2006a; Dunson, 2006; Pennell and Dunson, 2006), or by introducing dependence in the elements of the stick-breaking representation of the distribution (MacEachern, 1999, 2000; Teh et al., 2006; DeIorio et al., 2004; Gelfand et al., 2005; Griffin and Steel, 2006b; Rodriguez et al., 2006).

For example, given a set D , let $\{\boldsymbol{\eta}(t) \forall t \in D\}$ and $\{z(t) \forall t \in D\}$ be stochastic processes on D such that $z(t) \sim \text{Beta}(1, \alpha(t)) \forall t \in D$ and define

$$(7) \quad H_t(\cdot) = \sum_{l=1}^{\infty} w_l^*(t) \delta_{\boldsymbol{\eta}_l^*(t)}(\cdot),$$

where $\{\boldsymbol{\eta}_l^*(t)\}_{l=1}^{\infty}$ and $\{z_l^*(t)\}_{l=1}^{\infty}$ are collections of independent realizations of the stochastic processes $\{\boldsymbol{\eta}(t) \forall t \in D\}$ and $\{z(t) \forall t \in D\}$, and $w_l^*(t) = z_l^*(t) \prod_{s=1}^{l-1} (1 - z_s^*(t))$. The collection of probability measures $\mathcal{H}_D = \{H_t : t \in D\}$ is said to follow a dependent Dirichlet process (DDP). Note that, for any fixed t , G_t follows a Dirichlet process. One example of a dependent Dirichlet process is the nested Dirichlet process (nDP). A collection of probability measures $\{H_1, \dots, H_J\}$ is said to follow a *Nested Dirichlet Process* (nDP) with baseline measure H_0 and precision parameters α and β , denoted $\{H_1, \dots, H_J\} \sim \text{nDP}(\alpha, \beta, H_0)$, if

$$(8) \quad H_j(\cdot) \stackrel{iid}{\sim} \sum_{k=1}^{\infty} \pi_k^* \delta_{H_k^*}(\cdot)$$

$$(9) \quad H_k^*(\cdot) = \sum_{l=1}^{\infty} w_{lk}^* \delta_{\boldsymbol{\eta}_{lk}^*}(\cdot)$$

with $\boldsymbol{\eta}_{lk}^* \stackrel{iid}{\sim} H_0$, $w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1} (1 - u_{sk}^*)$, $\pi_k^* = v_k^* \prod_{s=1}^{k-1} (1 - v_s^*)$, $v_k^* \stackrel{iid}{\sim} \text{Beta}(1, \alpha)$ and $u_{lk}^* \stackrel{iid}{\sim} \text{Beta}(1, \beta)$.

The nDP uses the stick-breaking representation twice: first, it is used in (9) to construct the distributional atoms $\{H_k^*\}_{k=1}^{\infty}$, and then again in (8) to induce a prior on the collection of distributions using those atoms. Therefore, the nDP allows nonparametric estimation of the probability measures $\{H_1, \dots, H_J\}$, while borrowing information across them by creating clusters of identical measures. Exchangeability of the random measures is implicit in the definition of the nDP, and we can easily

prove that for any measurable set A and any j, j'

$$\begin{aligned}\mathbb{E}(H_j(A)) &= H_0(A), \\ \mathbb{V}(H_j(A)) &= \frac{H_0(A)(1 - H_0(A))}{\beta + 1}, \\ \mathbb{C}\text{or}(H_j(A), H_{j'}(A)) &= \frac{1}{1 + \alpha} = \mathbb{P}(H_j = H_{j'}) = \mathbb{C}\text{or}(H_j, H_{j'}).\end{aligned}$$

Therefore, H_0 can be interpreted as a mean distribution around which the condition-specific probability measures are centered, β controls the variability around that mean, and α controls the probability of two measures being assigned to the same cluster, and therefore the correlation among the probability measures. Efficient computation on the nDP can be carried out using truncations of the stick breaking representations (8) and (9). For a detailed discussion, see Rodriguez et al. (2006).

3.2. Hierarchical estimation of multiple functions. Consider now the problem of inferring multiple curves f_1, \dots, f_J , using the data $\mathbf{z}_1^{n_1}, \dots, \mathbf{z}_J^{n_J}$, where $\mathbf{z}_j^{n_j} = (\mathbf{z}_{1j}, \dots, \mathbf{z}_{j,n_j})$ is the set of n_j observations obtained under experimental condition j . The model described in section 2 could be extended to accommodate these multiple curves by using (conditionally) independent Dirichlet processes as the random measure on the mixing distributions. In order to borrow information, one can potentially include common unknown parameters in the baseline measures. However, such an approach can only borrow information globally under the parametric baseline model, so is quite inflexible.

Dependence across curves can also be incorporated by inducing dependence among the mixing distributions directly instead of through the baseline measure. Depending on the specific problem at hand, different types of processes can be used to induce such dependence. For example, if the goal is global functional clustering, the nested Dirichlet process (nDP) (Rodriguez et al., 2006) can be used as a prior on the collection of mixing distributions. On the other hand, if we are interested in local clustering of functions, the hierarchical Dirichlet process (Teh et al., 2006) (HDP) is a reasonable

choice. Finally, in a spatial data analysis setting, an extension of the spatial Dirichlet process (SDP) (Gelfand et al., 2005) could be used to enforce stronger dependence among curves obtained at closer geographical locations. Since for any fixed j the mixing distribution H_j derived from a dependent Dirichlet process follows a regular Dirichlet process, these models are marginally equivalent to that in Section 2.2.

As an illustration, consider a model for functional clustering using mixtures of nested Dirichlet Processes. Recall from section 3.1 that the nDP allows for simultaneous nonparametric estimation and clustering over a collection of distributions. Therefore, by using the nDP as a prior on the mixing distributions $\{H_1, \dots, H_J\}$ used to estimate the joint probability distributions $\{p_1(\mathbf{y}, \mathbf{x}), \dots, p_J(\mathbf{y}, \mathbf{x})\}$, we obtain a flexible model that allows for automatic nonparametric estimation of the regression functions, while partitioning the set of curves in groups of curves with similar shapes. Specifically, consider the following extension of the MEW model described in section 2.2, where

$$\begin{aligned}
 \mathbf{z}_{ij} = (\mathbf{y}_{ij}, \mathbf{x}_{ij}) &\sim \mathbf{N}_{p+q}(\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_{ij}) & i = 1, \dots, n_j; j = 1, \dots, J \\
 (\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_{ij}) &\sim H_j & \mathcal{H} = \{H_1, \dots, H_J\} \sim \text{nDP}(\alpha, \beta, H_0) \\
 (10) \quad H_0 &= \text{NIW}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0) & \kappa_0 \sim \text{G}(a_\kappa, b_\kappa) \\
 \boldsymbol{\theta}_0 &\sim \text{N}(\boldsymbol{\theta}_{00}, \mathbf{D}_{00}) & \boldsymbol{\Sigma}_{00} \sim \text{W}(\gamma, \boldsymbol{\Sigma}_{00}) \\
 \alpha &\sim \text{G}(a_\alpha, b_\alpha) & \beta \sim \text{G}(a_\beta, b_\beta).
 \end{aligned}$$

Let $n. = \sum_{j=1} n_j$ and $H_j^{n.}(\cdot | \mathbf{z}_1^{n_1}, \dots, \mathbf{z}_J^{n_J})$ be the posterior distribution of the parameters $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ under experimental condition j . Estimates of the mean regression functions $\{f_1, \dots, f_J\}$ can be

obtained from the posterior conditional expectations as

$$\begin{aligned}
 f_j^{n\cdot}(\mathbf{x}) &= \mathbb{E}_{H_j^{n\cdot}}(\mathbf{y}|\mathbf{x}, \mathbf{z}_1^{n_1}, \dots, \mathbf{z}_J^{n_J}) \\
 (11) \quad &= \int \frac{(\boldsymbol{\theta}_y + \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\theta}_x)) \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})}{\int \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx}) H_{0j}^{n\cdot}(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}_1^{n_1}, \dots, \mathbf{z}_j^{n_j})} H_{0j}^{n\cdot}(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}_1^{n_1}, \dots, \mathbf{z}_J^{n_J}).
 \end{aligned}$$

where, as before, $H_{0j}^{n\cdot}(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}_1^{n_1}, \dots, \mathbf{z}_J^{n_J})$ is the mean posterior mixing distribution in group j .

Although an explicit form is not available for the estimated regression function $f_j^{n\cdot}(\mathbf{x})$ or the estimated density $g_j^{n\cdot}(\mathbf{y}|\mathbf{x})$, they can be easily approximated for any \mathbf{x} (hence, for any dense grid of \mathbf{x} 's) using MCMC methods. Functional clustering in quantile or variance regression can be similarly approached by focusing on appropriate summaries of the posterior distribution. In addition to estimates of the underlying function for each of the experimental conditions, the model also generates a posterior distribution over all possible groupings of the J curves, which can be used to generate hypotheses about the scientific phenomena being studied.

From the definition of the nDP, it is clear that the model assumes that the curves are a priori exchangeable, and that there is a non-negative probability of multiple curves sharing the same mixture distribution, and therefore, the same shape. Note that the first level of nesting is used to estimate the regression functions non-parametrically, essentially reproducing the MEW model, while the second level induces clustering across the different functions. Curves j and j' are clustered together if $H_j = H_{j'} = H_k^*$ for some k , and such an event is given prior probability $1/(1 + \alpha)$. As $\alpha \rightarrow 0$, the model assumes a single cluster of curves (i.e. all the samples arise from the same underlying function), while $\alpha \rightarrow \infty$ implies different curves under each experimental condition. On the other hand, observations i and i' , respectively from distributions j and j' , are assigned to the same Gaussian component if and only if $H_j = H_{j'} = H_k^*$ and $(\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_{ij}) = (\boldsymbol{\theta}_{i'j'}, \boldsymbol{\Sigma}_{i'j'}) = (\boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^*)$ for some l . Therefore, the parameter β , in controlling the number of distinct $(\boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^*)$, controls the *non-linearity*

of the estimated functions by influencing the number of Gaussian distributions used to characterize the cluster-specific curves.

The hierarchical structure of the model implies that we borrow information across curves at two different levels. On one hand, curves assigned to the same cluster share the same set of regression lines and weights. On the other hand, curves assigned to different clusters borrow information through the parameters of the common baseline measure H_0 , which are in turn estimated by pooling information from all curves.

4. POSTERIOR CONSISTENCY

4.1. Posterior consistency of Dirichlet process mixtures. Posterior consistency and rates of convergence for nonparametric processes have been active areas of research in the last 20 years (Diaconis and Freedman, 1986a,b; Ghosal et al., 1999; Barron et al., 1999; Walker and Hjort, 2001), with seminal work dating back over 40 years (Doob, 1949; Schwartz, 1965). This section recalls some well known results on consistency of the Dirichlet Process that will be relevant later in the paper.

In what follows, we focus on the space of densities with respect to the Lebesgue measure on \mathbb{R}^p , which we denote $\mathfrak{m}(\mathbb{R}^p)$. Any element $g \in \mathfrak{m}(\mathbb{R}^p)$ has an associated absolutely continuous distribution G . There are a number of natural topologies on $\mathfrak{m}(\mathbb{R}^p)$, each one based on a different metric. For example, the Prokhorov-Lévy distance, defined as

$$\rho_w(g, g^0) = \inf \{ \epsilon > 0 : |G^0(\mathbf{z}) - G(\mathbf{z} - \epsilon)| \leq \|\epsilon\| \forall \mathbf{z} \in \mathbb{R}^p \},$$

induces the weak convergence topology. A weak ϵ -neighborhood of $g^0 \in \mathfrak{m}(\mathbb{R}^p)$ is defined as a set of the form,

$$U_\epsilon^w(g^0) = \left\{ g \in \mathfrak{m}(\mathbb{R}^p) : \left| \int \psi(\mathbf{z})g(\mathbf{z})d\mathbf{z} - \int \psi(\mathbf{z})g^0(\mathbf{z})d\mathbf{z} \right| < \epsilon \right\},$$

for all $\psi \in C_b(\mathbb{R}^p)$, the space of bounded continuous functions on \mathbb{R}^p .

Under this metric, the space $\mathfrak{m}(\mathbb{R}^p)$ is complete and separable and, under mild conditions on the kernel, the DPM model described in (2) is dense on $\mathfrak{m}(\mathbb{R}^p)$ (Ghosh and Ramamoorthi, 2003). Letting $\mathbf{z}_1, \dots, \mathbf{z}_n \sim g$ and $g \sim \mu$, with μ being a prior on $\mathfrak{m}(\mathbb{R}^p)$, the posterior probability of any measurable subset $A \subset \mathfrak{m}(\mathbb{R}^p)$ is given by

$$\mu_n(A) = \frac{\int_A \prod_{i=1}^n g(\mathbf{z}_i) \mu(dg)}{\int_{\mathfrak{m}(\mathbb{R}^p)} \prod_{i=1}^n g(\mathbf{z}_i) \mu(dg)},$$

and the optimal density estimate under square error loss is $g^n(\mathbf{z}) = \mathbb{E}(g(\mathbf{z})|\mathbf{z}^n)$, which reduces to (3) for the DPM prior. A prior μ on $\mathfrak{m}(\mathbb{R}^p)$ is said to be weakly consistent at g^0 iff, for almost every sequence $\mathbf{z}_1, \mathbf{z}_2, \dots$, $\int \psi(g) \mu_n(dg) \rightarrow \int \psi(g) \delta_{g^0}(dg)$ for every $\psi \in C_b(\mathbb{R}^p)$, which happens iff $\mu_n(U_\epsilon^w(g^0)) \rightarrow 1$, for all $\epsilon > 0$.

Note that, if a prior μ is weakly consistent at g^0 , the sequence of density estimates $\{g^n\}_{n=1}^\infty$ based on the sequence of posteriors $\{\mu_n\}_{n=1}^\infty$ converges pointwise to the true density g^0 with probability one. As we will see later, this implies that the estimates of the conditional distributions and conditional expectations also converge pointwise to the true underlying functions, ensuring consistency of our functional estimation method.

Sufficient conditions to ensure weak consistency were given by Schwartz (1965). As noted by Diaconis and Freedman (1986a,b), it is not enough to have g^0 in the weak support of μ , but g^0 needs to be in its Kullback-Leiber support, defined as the set

$$V^{KL}(g^0) = \{g : \pi(U_\epsilon^{KL}(g^0)) > 0 \forall \epsilon > 0\},$$

where $U_\epsilon^{KL}(g^0) = \{g : \int g^0 \log(g^0/g) < \epsilon\}$. The following result, which is an application of Schwartz's theorem, will be relevant for our consistency result,

Theorem 1 (Ghosal et al. (1999)). *Let $g^0 = \int \phi(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\Sigma})P_0(d\boldsymbol{\theta}, d\boldsymbol{\Sigma})$ be a location scale mixture of Gaussian distributions where P_0 is compactly supported and belongs to the weak support of μ . Then g^0 is in the Kullback-Leibler support of μ defined by the DP mixture in (2), and therefore the corresponding posterior distribution μ_n is weakly consistent at g^0 .*

4.2. Consistency of conditional function estimation. We focus now on the problem of assessing estimates $\{f_j^n(\mathbf{x})\}_{j=1}^J$ of the true regression functions $\{f_j^0(\mathbf{x})\}_{j=1}^J$ from estimates $\{g_j^n(\mathbf{y}, \mathbf{x})\}_{j=1}^J$ of the true joint distributions $\{g_j^0(\mathbf{y}, \mathbf{x})\}_{j=1}^J$ generating the data. We focus on the consistency of the sequence of functional estimates, rather than the more general problem of consistency of the posterior distribution on the space of random functions. In the sequel, we assume that the true mechanism generating the data for each curve $j = 1, \dots, J$ is as follows: 1) Covariates $\mathbf{x}_{1j}, \mathbf{x}_{2j}, \dots$ are drawn at random according to an absolutely continuous distribution with density $g_j^0(\mathbf{x})$ with compact support $D_{\mathbf{x}}$, and 2) Conditional on each \mathbf{x}_{ij} , the outcome \mathbf{y}_{ij} is sampled from the conditional density $g_j^0(\mathbf{y}|\mathbf{x})$, which is also absolutely continuous, with bounded support $D_{\mathbf{y}}$, and whose expectation is finite for every $\mathbf{x} \in D_{\mathbf{x}}$ and given by $\mathbb{E}_{g^0}(\mathbf{y}|\mathbf{x}) = f_j^0(\mathbf{x})$. Under this data generation mechanism, the joint true density $g_j^0(\mathbf{y}, \mathbf{x}) = g_j^0(\mathbf{y}|\mathbf{x})g_j^0(\mathbf{x})$ is absolutely continuous and defined on $D_{\mathbf{x}} \times D_{\mathbf{y}}$.

First, we consider the relationship between weak consistency of the prior on the joint distribution and pointwise consistency of the density estimates obtained from it.

Proposition 1. *If the prior μ_j on $\mathfrak{m}(\mathbb{R}^{p+q})$ is weakly consistent at $g_j^0(\mathbf{y}, \mathbf{x})$, then the estimates for the joint and marginal densities $g_j^n(\mathbf{y}, \mathbf{x})$ and $g_j^n(\mathbf{x})$ converge pointwise to $g_j^0(\mathbf{y}, \mathbf{x})$ and $g_j^0(\mathbf{x})$ respectively, for every $(\mathbf{x}, \mathbf{y}) \in D_{\mathbf{x}} \times D_{\mathbf{y}}$.*

Proof. Given that $(\mathbf{x}, \mathbf{y}) \in D_{\mathbf{x}} \times D_{\mathbf{y}}$, we note that both $g_j^n(\mathbf{y}, \mathbf{x})$ and $g_j^n(\mathbf{x})$ can be written as expectations of bounded functions with respect to the posterior measure μ_j^n (Antoniak, 1974; Lo, 1984).

Since the kernel $k(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\Sigma})$ is absolutely continuous, the result follows from the definition of weak consistency. \square

This pointwise consistency result can be extended to the density estimates of the conditional distributions.

Proposition 2. *Let $g_j^0(\mathbf{y}, \mathbf{x})$ be as described above and $\{g_j^n(\mathbf{y}, \mathbf{x}) = g_j^n(\mathbf{y} | \mathbf{x}) g_j^n(\mathbf{x})\}_{n=0}^\infty$ be a sequence of absolutely continuous density estimates arising from a prior μ_j on $\mathfrak{m}(\mathbb{R}^{p+q})$ that is weakly consistent at $g_j^0(\mathbf{y}, \mathbf{x})$. Then for any fixed \mathbf{x} , the estimate of the conditional density $g_j^n(\mathbf{y} | \mathbf{x})$ converges pointwise to $g_j^0(\mathbf{y} | \mathbf{x})$*

Proof. From Proposition 1 we know that for any $(\mathbf{x}, \mathbf{y}) \in D_{\mathbf{x}} \times D_{\mathbf{y}}$ it holds that

$$g^n(\mathbf{y}, \mathbf{x}) \rightarrow g^0(\mathbf{y}, \mathbf{x}) \quad \text{and} \quad g^n(\mathbf{x}) \rightarrow g^0(\mathbf{x})$$

Therefore, from Bayes' rule,

$$\lim_{n \rightarrow \infty} g^n(\mathbf{y} | \mathbf{x}) = \lim_{n \rightarrow \infty} \frac{g^n(\mathbf{y}, \mathbf{x})}{g^n(\mathbf{x})} = \frac{g^0(\mathbf{y}, \mathbf{x})}{g^0(\mathbf{x})} = g^0(\mathbf{y} | \mathbf{x})$$

for any $(\mathbf{x}, \mathbf{y}) \in D_{\mathbf{x}} \times D_{\mathbf{y}}$. \square

Corollary 1. *For any fixed $\mathbf{x} \in D_{\mathbf{x}}$, the functional estimate $f_j^n(\mathbf{x}) = \mathbb{E}_{g_j^n(\mathbf{y} | \mathbf{x})}(\mathbf{y})$ converges pointwise to $f^0(\mathbf{x})$.*

Remark 1. This is a result on pointwise convergence. Intuitively, uniform convergence is not to be expected. All functions in our sequence are continuous, but their limit might be a step function, as discussed below.

Remark 2. Since the true distribution $g_0(\mathbf{x})$ is assumed to be absolutely continuous over a compact set, Theorem 1 is an in-fill result, in the sense that it assumes that the function is observed on an finer and

finer grid as n increases. This suggests that, for designed experiments with repeated measurements, the behavior of the functional estimates can be unstable at points where no observations are made. This might potentially hold even if the true function is very smooth and the number of observations at the fixed design points is very large.

In the specific case of the MEW model described in Section 2, Corollary 1 can be made more specific.

Corollary 2. *Let S be the class of functions that arise as the conditional expectation of a countable mixture of normals, i.e.,*

$$S = \left\{ f(\mathbf{x}) : f(\mathbf{x}) = \mathbb{E}(\mathbf{y}|\mathbf{x}), (\mathbf{y}, \mathbf{x}) \sim \int \phi_{p+q}(\mathbf{y}, \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) P_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\}$$

where P_0 is compactly supported and almost surely discrete. Then, if $f^0 \in S$, the sequence of functional estimates from the MEW model is pointwise consistent, i.e., $f^n(\mathbf{x}) \rightarrow f^0(\mathbf{x})$ for every $\mathbf{x} \in D_{\mathbf{x}}$.

Proof. This is a consequence of theorem 1 and corollary 1. □

The following proposition shows that class S is large,

Proposition 3. *Under the L^1 metric, the closure of S is the space of bounded, integrable function on $D_{\mathbf{x}}$.*

Proof. First, recall that, under the L^1 metric, the space of step functions is dense on the space of integrable functions. That is, for any $\epsilon > 0$ and $f^0(\mathbf{x})$ that is bounded and absolutely continuous, there exists (at least) one step function $f^\epsilon(\mathbf{x})$ such that $\int |f^0(\mathbf{x}) - f^\epsilon(\mathbf{x})| d\mathbf{x} < \epsilon$. The problem reduces now to proving that S is dense on the space of step functions.

Note that any step function can be obtained as a conditional expectation of a joint distribution that is constant over hypercubes on \mathbb{R}^{p+q} (i.e., a tiled distribution). Let $g^\epsilon(\mathbf{y}, \mathbf{x})$ be the tiled distribution corresponding to $f^\epsilon(\mathbf{x})$.

Finally, note that any continuous distribution (and therefore, any tiled distribution) can be approximated arbitrarily well (in the total variation sense) by an infinite mixture of normals (Ghosh and Ramamoorthi, 2003). That is for any $g^\epsilon(\mathbf{y}, \mathbf{x})$ and any $\epsilon', \epsilon'' > 0$ there is a $g^*(\mathbf{y}, \mathbf{x})$ in the space of compactly supported mixtures of normals such that $\int |g^\epsilon(\mathbf{y}, \mathbf{x}) - g^*(\mathbf{y}, \mathbf{x})| d\mathbf{y} d\mathbf{x} < \epsilon'$ and $\int |g^\epsilon(\mathbf{x}) - g^*(\mathbf{x})| d\mathbf{x} < \epsilon''$. \square

The results for the MEW model can be extended to the nested Dirichlet Process. Rodriguez et al. (2006) provide simulation results suggesting consistency in multiple group density estimation for the nDP. The following theorem formally demonstrates consistency for a fixed number of groups and increasing number of observations per group.

Theorem 2. *Suppose that the true densities generating the data, $\{g_j^0\}_{j=1}^J$, each belong to the set $\{g_k^{*0}\}_{k=1}^K$, where g_k^{*0} is a compactly supported mixture of Gaussian densities with $K \leq J$ and J fixed. Then, the nDP mixture prior is weakly consistent as the sample sizes n_1, \dots, n_J all grow to infinity.*

Proof. First, note that the J true densities are clustered into K groups, and the allocation to groups defines a partition of $\{1, \dots, J\}$. The nDP induces a prior over the set of possible partitions. As this prior has full support, the posterior probability of the true partition will converge to one as the sample sizes in each of the groups increases. Conditional on the partition, the nDP implies independent Dirichlet process mixtures of Gaussian priors for the cluster-specific densities. Hence, posterior consistency follows automatically from the results of 1. \square

Corollary 3. *Each functional estimate arising from the model in 10 is consistent on the class of integrable functions on a compact set $D_{\mathbf{x}}$.*

General consistency results for the class of dependent Dirichlet processes are still an open problem. We note that, as these results become available, the propositions in this section can be used to establish consistency of the associated functional estimation model.

5. AN ILLUSTRATION: CLUSTERING TEMPERATURE PROFILES IN THE NORTH ATLANTIC

Conductivity and Temperature at Depth data (CTD) are regularly used in oceanography to study the physical properties of a water column. The CTD profiler is a torpedo-shaped instrument that is attached to a conducting wire and lowered to pre-specified depths. At each depth, information on pressure, temperature and conductivity is sent back to the ship through the wire. In some cases, water samples are also taken. The result from this measurement process is a sample from the functions relating conductivity and temperature with depths at each location and time.

Latitude plays the most important role in defining the shape of CTD profiles: the farther away from the equator, the lower the average temperature of the water column is. Seasonal effects are also very important; a difference of only 3 weeks can produce huge variations in the profile, particularly near the surface. However, these factors are not the only determinants of the profile shape. For example, oceanic currents and salinity gradients due to fresh water discharge can effectively become barriers preventing mixing. Therefore, CTD profiles can be highly non-linear, particularly in coastal regions.

Understanding the patterns of spatio-temporal evolution of the profiles can help scientists assess the magnitude and consequences of global phenomena like El Niño and the process of global warming. However, CTD profiles are obtained very sparsely (both in space and time) and do not necessarily

change smoothly with latitude or longitude due to the reasons discussed above, making regular spatio-temporal models hard to justify in many specific geographic regions. An alternative approach for the analysis of CTD profiles is to borrow information through probabilistic clustering in order to improve functional estimation and identify regions of the ocean with similar characteristics. We expect most of the clusters to agree with spatial locations, with inconsistencies signaling boundary regions.

As an illustration, we focus on 87 temperature profiles collected in the North Atlantic ocean between June 15 and June 22, 1986. The number of observations per curve varies between 31 and 83. Temperature measurements are usually collected every 10 m from a starting depth that varies with the location, but in some cases the separation between observations can be much larger. An exploratory analysis of the data shows four or five different types of profiles collected at three geographic regions: off the coast of Nova Scotia in Canada, off the coast of Portugal and 1000 km off from the coast of Africa. We apply the approach described in Section 3 to this data. Our goal is to assess clusters in the data and estimate the true profiles of temperature vs. depth by borrowing information across locations.

Computation was carried out using the algorithm described in appendix B, which is an extension of the Gibbs sampler described in Rodriguez et al. (2006). Hyperparameters were set according to the empirical distribution of the data, with θ_{00} equal to the overall sample mean and Σ_{00} equal to the sample covariance matrix. For the other parameters associated with the baseline measure, we chose $\mathbf{D} = \Sigma/100$, $a_{\kappa} = 1$ and $b_{\kappa} = 100$, in such a way that $\mathbb{E}(\kappa_0) = 0.01$, $\nu_0 = 3$ and $\gamma = 3$. For the precision parameters, and based on the discussion in Rodriguez et al. (2006), we pick $a_{\alpha} = b_{\alpha} = a_{\beta} = b_{\beta} = 3$.

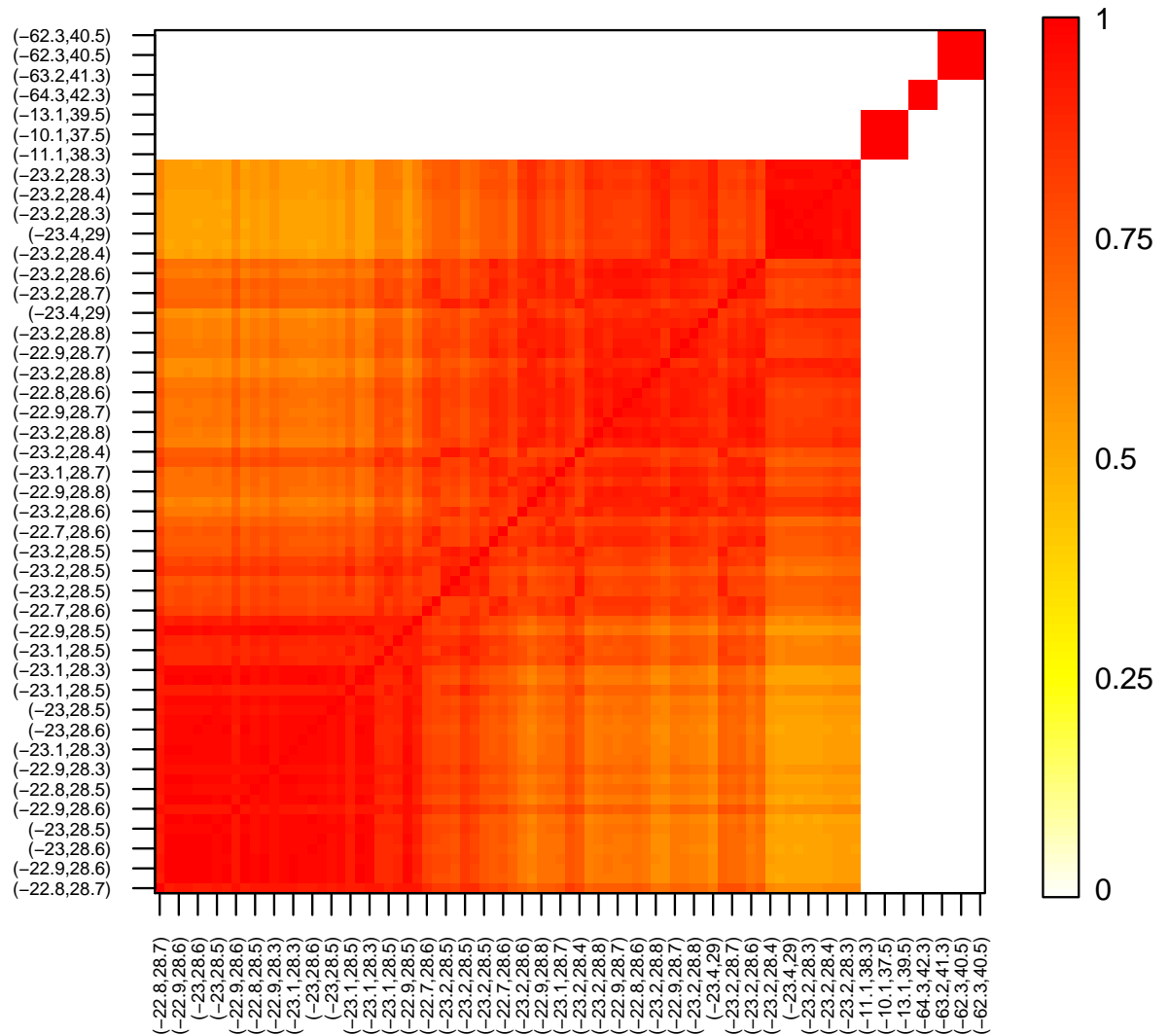


FIGURE 1. Heatmap with the probabilities of pairwise joint classification. Pixel (i, j) represents the posterior probability of locations i and j being clustered together. The axes correspond to the longitude/latitude where the data were collected.

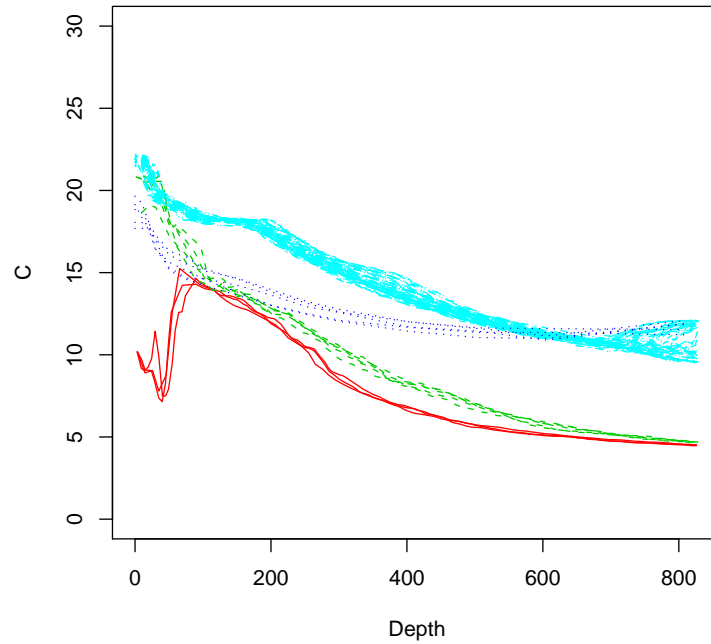


FIGURE 2. Raw profiles collected in the North Atlantic between June 15 and June 22, 1986. Colors and line types indicate cluster membership elicited from the pairwise posterior probability matrix.

All inferences are based on 40,000 samples obtained after a burn-in of 10,000 iterations. To obtain a reasonable starting cluster configuration, linear models were fitted separately to each of the locations in the sample. After hierarchical clustering was applied to the 87 pairs of parameters, a dendrogram was inspected to identify groups of curves with similar linear fits. When compared with a naive starting point, this heuristic was successful in speeding up convergence of the algorithm.

Figure 1 shows a heatmap of the probabilities of pairwise joint classification. In this figure, pixel (i, j) represents the posterior probability of locations i and j being clustered together. From the plot, it is clear that locations cluster in four groups, a large cluster composed of 75 locations, and

three smaller ones with 3, 4 and 5 observations each. Figures 2 and 3 show the raw curves and the location where the curves were collected, with colors corresponding to the clusters obtained from the heatmap. Note that the big cluster corresponds to the site off the coast of Africa, and the model shows a small probability (around 0.03) of this cluster being broken into two distinct groups based on the different behaviors observed after 750 m depth. One of the small clusters corresponds to the locations off Portugal, while the curves off the coast of Nova Scotia are classified in two groups, seemingly dependent on their distance to the coast. These two clusters have a straightforward explanation: two different currents, one flowing south from the Antarctic very close to the coast, and another running north from the Gulf of Mexico further away from the coast, meet by the coast of Nova Scotia. These two water masses do not mix, producing very different profiles in close geographic areas.

Figure 4 displays the estimated profiles at each of the 87 locations. These plots were obtained by estimating the value of the function on a grid of 200 points and doing linear interpolation. Since there is little uncertainty in the clustering, profiles overlap. The behavior of the profiles is clearly non linear and some of them are not even monotone, characteristics that are consistent with scientific knowledge. Indeed, the probability of a one component mixture is estimated to be zero for each one of the 87 curves, indicating that using linear models to approximate the functions would not be appropriate. The curves off Nova Scotia show a behavior that is consistent with our hypothesis about oceanic currents. The cluster closest to the coast is characterized by profiles with low surface temperature due to the influence of Arctic waters. On the other hand, the cluster farthest away from the coast is characterized by profiles with a very high surface temperature (almost as high as African profiles) that declines very fast. As is to be expected, the temperatures in both clusters seem to converge at depths over 600 m.

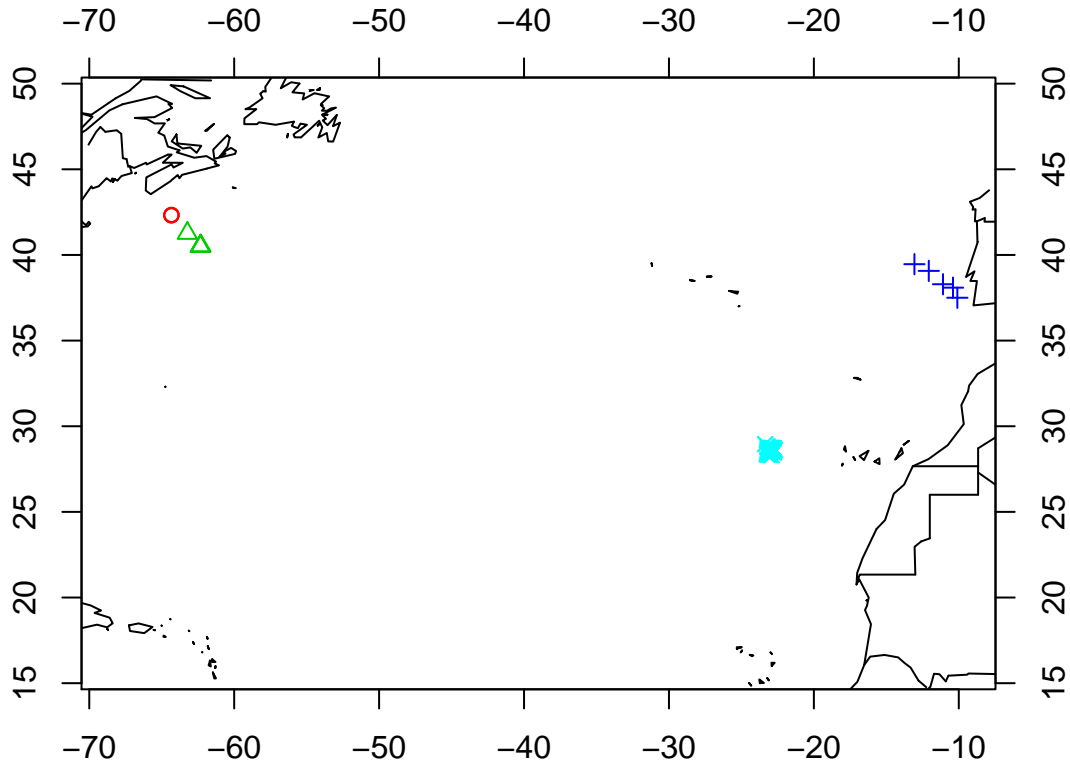


FIGURE 3. Geographic locations where the profiles were collected. Colors and symbols indicate cluster membership elicited from the pairwise posterior probability matrix.

The only unappealing feature of our functional estimates is the bump in the dark blue (dotted) curve appearing around 700m. This bump is due to the sparseness in the data off the Portuguese coast

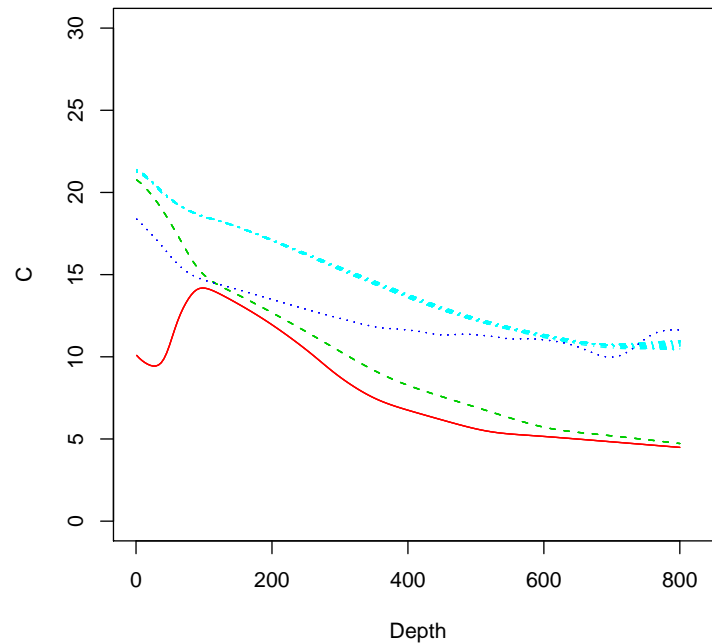


FIGURE 4. Fitted curves at each location obtained after model averaging. There are actually 87 distinct curves represented in the plot but, due to the tight cluster membership, most are undistinguishable. Colors and line types indicate cluster membership.

(in these locations, observations below the 300 m mark were collected only every 100 m). As was discussed in Section 4, large gaps in the predictor space can produce unstable functional estimates within the gaps. However, we do not expect this instability to affect the clustering results. In line with this comment, probability bands around the estimated function (not shown) become much wider in this section of the curve.

6. DISCUSSION

We have introduced a novel method to construct hierarchical models for functions. Central to our approach is the indirect estimation of the conditional distribution of outcomes given the predictors through the corresponding joint distribution. From this conditional distribution, the function of interest is obtained as the conditional expectation, yielding very flexible function estimates. We avoid parametric assumptions on the mean shape and error distributions, obtaining a method that balances non-linearity in the mean with non-normality in the residual distribution to obtain robust functional estimates. We also provide theoretical support for the methodology by establishing conditions for consistency of the function estimates. Our results link weak consistency in the density estimation problem and pointwise consistency of conditional expectations.

To demonstrate the advantages of the method, we focus on an application to functional clustering using the nested Dirichlet process as a prior on the collection of mixing distributions that define the joint density of outcomes and predictors. This model induces clustering on the joint distribution which is actually a stronger condition than clustering of the mean function. Although this can potentially produce more clusters than expected (either because multiple experiments have similar mean functions but different error structures, or because the sampling patterns for the covariates are different), we show that model performs well in practice and produces both interpretable clusters and sensible function estimates.

APPENDIX A. NOTATION

This appendix establishes the notation and parametrizations we used in the paper.

A.1. Normal-inverse-Wishart distribution. We say that $(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \sim \text{NIW}_p(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0)$ if the joint density can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+p+2)/2} \exp \left\{ -\frac{\kappa_0}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \frac{1}{2} \text{tr}(\nu_0 \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}^{-1}) \right\}$$

A.2. Gamma distribution. We denote $\tau \sim \text{G}(a, b)$ if

$$p(\tau) \propto \tau^{a-1} \exp \{-b\tau\}$$

A.3. Wishart distribution. We write $\mathbf{S} \sim \text{W}_p(\gamma, \mathbf{S}_0)$ if

$$p(\mathbf{S}) \propto |\mathbf{S}|^{(\gamma-p-1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\gamma \mathbf{S}_0^{-1} \mathbf{S}) \right\}$$

APPENDIX B. COMPUTATIONAL IMPLEMENTATION

We implement the nDP model using the two-level truncation algorithm described in Rodriguez et al. (2006). This algorithm uses a finite mixture to approximate each of the stick-breaking processes involved in the definition of the nDP. We used truncation levels set at $K = L = 55$ atoms. These truncation levels seem to yield reasonable approximations for the sample sizes involved in our oceanographic example.

We introduce latent variables ζ_j and ξ_{ij} such that $\zeta_j = k$ if $H_j = H_k^*$ and $\xi_{ij} = l$ if $(\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_{ij}) = (\boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^*)$. Once adequate starting values for the parameters have been chosen, computation proceeds through the following steps:

- (1) Sample the bottom-level indicators ζ_j for $j = 1, \dots, J$ from a multinomial distribution with probabilities

$$\mathbb{P}(\zeta_j = k | \dots) = q_k^j \propto w_k^* \prod_{i=1}^{n_j} \sum_{l=1}^L \pi_{lk} \phi_{p+q}(\mathbf{z}_{ij} | \boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^*), \quad k = 1, \dots, K.$$

- (2) Sample the top-level indicators ξ_{ij} for $j = 1, \dots, J$ and $i = 1, \dots, n_j$ from another multinomial distribution with probabilities

$$\mathbb{P}(\xi_{ij} = l | \dots) = b_{ij}^l \propto \pi_{l,\zeta_j}^* \phi_{p+q}(\mathbf{z}_{ij} | \boldsymbol{\theta}_{l,\zeta_j}^*, \boldsymbol{\Sigma}_{l,\zeta_j}^*), \quad l = 1, \dots, L.$$

- (3) Sample bottom-level probabilities π_k^* by generating

$$(u_k^* | \dots) \sim \text{Beta} \left(1 + m_k, \alpha + \sum_{s=k+1}^K m_s \right), \quad k = 1, \dots, K-1, \quad u_K^* = 1,$$

where m_k is the number of distributions assigned to component k , and constructing $\pi_k^* = u_k^* \prod_{s=1}^{k-1} (1 - u_s^*)$.

- (4) Sample the top-level probabilities w_{lk}^* by generating

$$(v_{lk}^* | \dots) \sim \text{Beta} \left(1 + n_{lk}, \beta + \sum_{s=l+1}^L n_{ls} \right), \quad l = 1, \dots, L-1, \quad v_{Lk}^* = 1,$$

where n_{lk} is the number of observations assigned to atom l of distribution k , and constructing $w_{lk}^* = v_{lk}^* \prod_{s=1}^{l-1} (1 - v_{sk}^*)$.

- (5) Sample the atoms $(\boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^*)$ from

$$(\boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^* | \dots) \sim \text{NIW}(\hat{\boldsymbol{\theta}}_{lk}, \hat{\kappa}_{lk}, \hat{\nu}_{lk}, \hat{\boldsymbol{\Sigma}}_{lk}),$$

where

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{lk} &= \frac{n_{lk}}{\kappa_0 + n_{lk}} \bar{\mathbf{z}}_{lk} + \frac{\kappa_0}{\kappa_0 + n_{lk}} \boldsymbol{\theta}_0 \\ \hat{\kappa}_{lk} &= \kappa_0 + n_{lk} \\ \hat{\nu}_{lk} &= \nu_0 + n_{lk} \\ \hat{\nu}_{lk} \hat{\boldsymbol{\Sigma}}_{lk} &= \nu_0 \boldsymbol{\Sigma}_0 + n_{lk} \bar{\mathbf{S}}_{lk} + \frac{\kappa_0 n_{lk}}{\kappa_0 + n_{lk}} (\bar{\mathbf{z}}_{lk} - \boldsymbol{\theta}_0)(\bar{\mathbf{z}}_{lk} - \boldsymbol{\theta}_0)' \\ \bar{\mathbf{z}}_{lk} &= \frac{1}{n_{lk}} \sum_{\{i,j:\zeta_j=k,\xi_{ij}=l\}} \mathbf{z}_{ij} \\ \bar{\mathbf{S}}_{lk} &= \frac{1}{n_{lk}} \sum_{\{i,j:\zeta_j=k,\xi_{ij}=l\}} (\mathbf{z}_{ij} - \bar{\mathbf{z}}_{lk})(\mathbf{z}_{ij} - \bar{\mathbf{z}}_{lk})'\end{aligned}$$

and n_{lk} is the number of observations assigned to atom (l, k) . Note that, if no observation is assigned to a specific cluster, then the parameters are drawn from the conditional prior distribution (baseline measure) $\text{NIW}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0)$.

(6) Sample the baseline mean $\boldsymbol{\theta}_0$ from

$$(\boldsymbol{\theta}_0 | \dots) \sim \text{N} \left([\mathbf{D}_{00}^{-1} + \bar{\mathbf{D}}]^{-1} [\mathbf{D}_{00}^{-1} \boldsymbol{\theta}_{00} + \bar{\mathbf{d}}], [\mathbf{D}_{00}^{-1} + \bar{\mathbf{D}}]^{-1} \right)$$

where

$$\bar{\mathbf{D}} = \kappa_0 \sum_{\{l,k:n_{lk} \neq 0\}} \boldsymbol{\Sigma}_{lk}^{*-1} \quad \bar{\mathbf{d}} = \kappa_0 \sum_{\{l,k:n_{lk} \neq 0\}} \boldsymbol{\Sigma}_{lk}^{*-1} \boldsymbol{\theta}_{lk}^*$$

(7) Sample the variance of the baseline measure, $\boldsymbol{\Sigma}_0$ from

$$(\boldsymbol{\Sigma}_0 | \dots) \sim \text{W} \left(\gamma + c\nu_0, \gamma \boldsymbol{\Sigma}_{00}^{-1} + \frac{\nu_0}{\kappa_0} \bar{\mathbf{D}} \right)$$

where c is the number of non-empty components.

(8) Sample the mean precision parameter κ_0 from

$$(\kappa_0 | \dots) \sim \text{G} \left(a_\kappa + \frac{c(p+q)}{2}, b_\kappa + \frac{1}{2} \sum_{\{l,k:n_{lk} \neq 0\}} (\boldsymbol{\theta}_{lk}^* - \boldsymbol{\theta}_0)' \boldsymbol{\Sigma}_{lk}^{*-1} (\boldsymbol{\theta}_{lk}^* - \boldsymbol{\theta}_0) \right)$$

(9) Sample the concentration parameters α and β from

$$(\alpha | \dots) \sim \text{G} \left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1} \log(1 - u_k^*) \right)$$

$$(\beta | \dots) \sim \text{G} \left(a_\beta + K(L-1), b_\beta - \sum_{l=1}^{L-1} \sum_{k=1}^K \log(1 - v_{lk}^*) \right)$$

REFERENCES

- Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.
- Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician* **46**, 175–185.
- Antoniak, C. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.
- Barron, A., M. Schervish, and L. Wasserman (1999). The consistency of distributions in nonparametric problems. *The Annals of Statistics* **27**, 536–561.
- Basu, S. and S. Chib (2003). Marginal likelihood and bayes factors for dirichlet process mixture models. *Journal of the American Statistical Association* **98**, 224–235.
- Behseta, S., R. E. Kass, and G. L. Wallstrom (2005). Hierarchical models for assessing variability among functions. *Biometrika* **92**, 419–434.
- Bigelow, J. L. and D. B. Dunson (2005). Semiparametric classification in hierarchical functional data analysis. Technical report, Institute of Statistics and Decision Sciences, Duke University.
- Bush, C. A. and S. N. MacEachern (1996). A semiparametric bayesian model for randomised block designs. *Biometrika* **83**, 275–285.
- Chu, C.-K. and J. S. Marron (1991). Choosing a kernel regression estimator. *Statistical Science* **6**, 404–419.
- Dahl, D. (2003). An improved merge-split sampler for conjugate dirichlet process mixture models. Technical report, Department of Statistics, University of Wisconsin.
- DeIorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An anova model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

- Diaconis, P. and D. Freedman (1986a). On inconsistent bayes estimates of location. *The Annals of Statistics* **14**, 68–87.
- Diaconis, P. and D. Freedman (1986b). On the consistency of bayes estimates. *The Annals of Statistics* **14**, 1–26.
- Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pp. 23–37. Centre National de la Recherche Scientifique.
- Dunson, D. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7**, 551–568.
- Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B*. *In press*.
- Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.
- Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* **90**, 577–588.
- Fan, J. Q., N. E. Hickman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of American Statistical Association* **90**, 141–150.
- Fan, J. Q., Q. Yao, and H. Tong (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.
- Fan, J. Q. and T. H. Yim (2004). A cross validation method for estimating conditional densities. *Biometrika* **91**, 819–834.
- Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.
- Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.
- Ghosal, S., J. Ghosh, and R. V. Ramamoorthi (1999). Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics* **27**, 143–158.
- Ghosh, J. K. and R. V. Ramamoorthi (2003). *Bayesian nonparametrics*. New York: Springer-Verlag.
- Green, P. and S. Richardson (2001). Modelling heterogeneity with and without the dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.
- Griffin, J. E. and M. F. J. Steel (2006a). Nonparametric inference in time series problems. In *Valencia Statistics 8*.
- Griffin, J. E. and M. F. J. Steel (2006b). Order-based dependent dirichlet processes. *Journal of the American Statistical Association* **101**, 179–194.
- Guo, W. (2002). Functional mixed effect models. *Biometrics* **58**, 121–128.
- Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.
- Jain, S. and R. M. Neal (2000). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. Technical report, Department of Statistics, University of Toronto.
- Lo, A. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *Annals of Statistics* **12**, 351–357.
- Luan, Y. and H. Li (2003). Clustering of time-course gene expression data using a mixed effects model with b-splines. *Bioinformatics* **19**, 474–482.

- MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics, Part B - Simulation and Computation* **23**, 727–741.
- MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.
- MacEachern, S. N. (2000). Dependent dirichlet processes. Technical report, Ohio State University, Department of Statistics.
- Morris, J. S. and R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68**, 179–199.
- Müller, P., A. Erkanli, and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.
- Müller, P., F. Quintana, and G. Rosner (2004). Hierarchical meta-analysis over related non-parametric bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.
- Pennell, M. L. and D. B. Dunson (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* **62**, 1044–1052.
- Ramoni, M., P. Sebastiani, and P. Kohane (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences* **99**, 9121–9126.
- Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian processes for machine learning*. The MIT Press.
- Ray, S. and B. K. Mallick (2006). Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society, Series B*. **68**, 305–332.
- Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53**, 233–243.

- Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2006). The nested dirichlet process. Technical report, Institute of Statistics and Decision Sciences, Duke University.
- Schwartz, L. (1965). On bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4**, 10–26.
- Sethuraman, J. (1994). A constructive definition of dirichelt priors. *Statistica Sinica* **4**, 639–650.
- Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Sharing clusters among related groups: Hierarchical dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.
- Truong, Y., C. Kooperberg, and C. Stone (2005). *Statistical modeling with spline functions: Methodology and theory*. Springer.
- Vidakovic, B. (1999). *Statistical modeling by wavelets*. New York: Wiley.
- Wakefield, J., C. Zhou, and S. Self (2003). Modelling gene expression over time: curve clustering with informative prior distributions. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *In Bayesian Statistics 7*, pp. 721–732. Oxford University Press.
- Walker, S. G. and N. L. Hjort (2001). On bayesian consistency. *Journal of the Royal Statistical Society, Series B* **63**, 811–821.
- Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B* **93**, 159–174.
- Wu, H. and J. T. Zhang (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of American Statistical Association* **97**, 883–897.