Published in final edited form as:
Biometrika. 2009 ; 96(1): 149–162.

Bayesian Nonparametric Functional Data Analysis Through Density Estimation

Abel Rodríguez,

Department of Applied Mathematics and Statistics, University of California, Mailstop SOE2, Santa Cruz, California 95064, U.S.A. abel@soe.ucsc.edu

David B. Dunson, and

Biostatistics Branch, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A. dunson1@niehs.nih.gov

Alan E. Gelfand

Department of Statistical Sciences, Duke University, Box 90251, Durham, North Carolina 27708, U.S.A. alan@isds.duke.edu

Summary

In many modern experimental settings, observations are obtained in the form of functions, and interest focuses on inferences on a collection of such functions. 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 Conductivity and Temperature at Depth data in the north Atlantic.

Keywords

Dependent Dirichlet process; Functional clustering; Nonparametric Bayes; Nonparametric regressions; Random probability measure

1 Introduction

Much current scientific interest centers on the functional relationships between variables, and how these change under different experimental conditions. Given pairs $\{(y_{ij}, x_{ij})\}$, where $j = 1, \dots, J$ indexes an experimental condition and $i = 1, \dots, n_j$ indexes an observation within the experiment, $y_{ij} \in \mathbb{R}^q$ and $x_{ij} \in \mathbb{R}^p$, we are interested in jointly estimating functions $f_1, \dots, f_J : \mathbb{R}^p \rightarrow \mathbb{R}^q$ that describe the relationship between predictors and outcomes, testing hypotheses about the dependence between the functions, and predicting the function under new experimental conditions. Depending on the application at hand, these functions might correspond to the conditional mean responses or quantiles of the conditional distributions, while the inference goal might be multiple comparison of curves, functional clustering or spatial prediction of the functional relationship. We develop a class of models that can tackle such joint inference problems from a Bayesian nonparametric perspective.

Popular approaches for nonparametric functional estimation include kernel regression methods, models that assume that the functions of interest can be represented as linear combinations, such as splines, wavelets and reproducing kernel methods, and methods that

assume that the functions in question are realizations of stochastic processes, with the Gaussian process being a common choice.

Different approaches have been used to extend these methodologies to collections of functions. For example, when the function of interest is modelled as a linear combination of basis functions, hierarchical models for the basis coefficients can be used to accommodate different types of dependence. This approach has been successfully exploited by authors such as Rice & Silverman (1991), Wu & Zhang (2002) and Morris & Carroll (2006) to construct analysis of variance and random effect models for curves. Along similar lines, Bigelow & Dunson (2008) and Ray & 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) develop a hierarchical model which treats individual curves as realizations of a Gaussian process centred on a Gaussian process mean function.

All these methods are based on specifications for the conditional densities $g_1(y/x), \dots, g_J(y/x)$, where $g_j(y/x)$ denotes the density of the outcome y given the predictor x under experimental condition j . Instead of modelling the conditional distributions directly, in this paper we induce a prior on the space of functions indirectly through a model on the collection of joint distributions $g_1(y, x), \dots, g_J(y, x)$ that uses mixtures of dependent Dirichlet processes (MacEachern, 1999; DeIorio et al., 2004; Gelfand et al., 2005; Rodríguez et al., 2008). This method is conceptually related to the double-kernel method of Fan & 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 the complicated and inefficient simulation algorithms typically required for adaptive function estimation. It also allows for different smoothing levels across different predictors.

As we obtain a joint posterior distribution for the conditional response distributions, we make inferences about regression functions characterized in terms of the mean or a quantile. 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 estimators on a dense subset of the space of bounded, integrable functions on compact subsets of $\mathbb{R}^p$.

Regression analysis through joint distributions is particularly useful in missing data and inverse regression problems, which require probability models on the covariate space. Typically, such models are specified as a sequence of generalized linear models, requiring the predictors to follow distributions in the exponential family. In contrast, our approach automatically models the joint distribution of the predictors as a flexible mixture of normals, which can be easily adapted for categorical predictors using a latent variable specification. Hence, we can allow for missing predictors automatically without imposing additional modelling assumptions or modifying the code for posterior computation. Similarly, inverse problems can be addressed by simply inverting the direction of the conditioning.

2 Single-function nonparametric regression

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 $\mathcal{X}$. Let $H \in \mathcal{H}$ be a probability measure on $(\mathcal{X}, \mathcal{B})$. A Dirichlet process (Ferguson, 1973) with baseline measure H_0 and precision α , denoted by $DP(\alpha H_0)$, 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 $\pi_i \sim \text{Be}(1, \alpha)$ and $\eta_i H_0$, for all i be independent. If H is defined as

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

where $w_i = \pi_i \prod_{j=1}^{i-1} (1 - \pi_j)$ and $\delta_x(\cdot)$ represents a degenerate distribution at x , then $H \sim DP(\alpha H_0)$ (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 models (Escobar, 1994; Escobar & West, 1995), where the Dirichlet process is used as the prior on the random mixing distribution over the parameters of a continuous distribution:

$$z \sim g(\cdot), \quad g(\cdot) = \int k(\cdot|\eta) H(d\eta), \quad H \sim DP(\alpha H_0). \quad (2)$$

Therefore, the Dirichlet process mixtures induce a prior on g indirectly through a prior on the mixing distribution H . A popular choice is the Dirichlet process mixture of Gaussian distributions, where $\eta = (\theta, \Sigma)$ and $k(\cdot|\eta) = \phi_p(\cdot|\theta, \Sigma)$ is the p -variate normal kernel with mean θ and covariance matrix Σ . Given an independent and identically distributed sample $z^n = (z_1, \dots, z_n)$, the posterior of the mixing distribution, $H_n(\cdot|z^n)$, is distributed as a mixture of Dirichlet processes (Antoniak, 1974),

$$H^n(\cdot|z^n) \sim \int DP \left(\alpha H_0 + \sum_{i=1}^n \delta_{\theta_i} \right) p(d\theta_1, \dots, d\theta_n | z^n),$$

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

$$g^n(z) = E_{H^n} \int k(z|\eta) H^n(d\eta|z^n) = \int k(z|\eta) H_0^n(d\eta|z^n), \quad (3)$$

where

$$H_0^n(\cdot|z^n) = \int \frac{\alpha H_0(\cdot) + \sum_{i=1}^n \delta_{\theta_i}(\cdot)}{\alpha + n} p(d\theta_1, \dots, d\theta_n | z^n)$$

is the posterior mean of the mixing distribution.

Computation for Dirichlet process mixture models is typically carried out using one of three different approaches: Pólya urn schemes that marginalize out the unknown distribution H (Escobar & West, 1995; Bush & MacEachern, 1996), truncation methods that use finite mixture models to approximate the Dirichlet process (Ishwaran & James, 2001; Green & Richardson, 2001) and reversible jump algorithms (Green & Richardson, 2001; Jain & Neal, 2004).

Consider the following application of the model for multivariate density estimation described in (2),

$$\begin{aligned} z_i &= (y_i, x_i) \sim N_{p+q}(\theta_i, \Sigma_i) & i=1, \dots, n \\ (\theta_i, \Sigma_i) &\sim H & H \sim DP(\alpha H_0) \end{aligned} \quad (4)$$

with $H_0 = NIW_{p+q}(\theta_0, \kappa_0, \nu_0, \Sigma_0)$, $\theta_0 \sim N_{p+q}(\theta_{00}; D_{00})$, $\Sigma_0 \sim Wi_{p+q}(\gamma, \Sigma_{00})$, $\kappa_0 \sim Ga(a_\kappa, b_\kappa)$, and $\alpha \sim Ga(a_\alpha, b_\alpha)$. In the above expression, NIW_p denotes the p -variate Normal-Inverse-Wishart distribution, Ga denotes the gamma distribution, Wi denotes the p -variate Wishart distribution; we say that $(\theta, \Sigma) \sim NIW_p(\theta_0, \kappa_0, \nu_0, \Sigma_0)$ if the joint density can be written as

$$p(\theta, \Sigma) \propto |\Sigma|^{-(\nu_0 + p + 2)/2} \exp \left\{ -\frac{\kappa_0}{2} (\theta - \theta_0)' \Sigma^{-1} (\theta - \theta_0) - \frac{1}{2} \text{tr}(\nu_0 \Sigma_0 \Sigma^{-1}) \right\}.$$

Similarly $\tau \sim Ga(a, b)$ if $p(\tau) \propto \tau^{a-1} \exp\{-b\tau\}$ and $S \sim Wig(\gamma, S_0)$ if

$$p(S) \propto |S|^{(\gamma - p - 1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\gamma S_0^{-1} S) \right\}.$$

In our model, the parameters at the top level are partitioned as

$$\theta = (\theta_y, \theta_x)' \quad \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{yx} & \Sigma_{xx} \end{pmatrix}.$$

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

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

In turn, a nonparametric estimate $f^n(x)$ of $f(x)$ can be obtained from (5) by calculating the conditional expectation,

$$f^n(x) = E(y|x, z^n) = \int \frac{\{\theta_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \theta_x)\} \phi_p(x|\theta_x, \Sigma_{xx})}{\int \phi_p(x|\theta_x, \Sigma_{xx}) H_0^n(d\theta, d\Sigma|z^n)} H_0^n(d\theta, d\Sigma|z^n). \quad (6)$$

Similarly, quantile regression can be performed by obtaining the quantile of the conditional distribution in (5). For any fixed 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.

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 Markov chain Monte Carlo algorithms, such as the method of Basu & 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 & Chib (1993), resulting in a model that simultaneously incorporates a non-parametric regression function and a non-parametric link function.

3 Hierarchical nonparametric models for functions

Section 2 deals with a single random curve. Simultaneous inference about 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.

Much of the recent interest in Bayesian nonparametrics has focused on models for collections of distributions. Most extensions of the Dirichlet process achieve dependence either by forming convex combinations of independent processes (Müller et al., 2004; Dunson, 2006; Pennell & Dunson, 2006; Dunson et al., 2007), or by introducing dependence in the elements of the stick-breaking representation of the distribution (MacEachern, 1999; Teh et al., 2006; DeIorio et al., 2004; Gelfand et al., 2005; Griffin & Steel, 2006; Rodríguez et al., 2008). For example, given a set D , let $\{\eta(t), t \in D\}$ and $\{z(t), t \in D\}$ be stochastic processes on D such that $z(t) \sim \text{Be}\{1, \alpha(t)\}$ for all $t \in D$ and define

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

where $\{\eta_l^*(t)\}_{l=1}^{\infty}$ and $\{z_l^*(t)\}_{l=1}^{\infty}$ are collections of independent realizations of the stochastic processes $\{\eta(t) \in D\}$ and $\{z(t), t \in D\}$, 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. Note that, for any fixed t , G_t follows a Dirichlet process. One example of a dependent Dirichlet process is the nested Dirichlet process. A collection of probability measures $\{H_1, \dots, H_J\}$ is said to follow a nested Dirichlet process with baseline measure H_0 and precision parameters α and β , denoted by $\{H_1, \dots, H_J\} \sim DP\{\alpha DP(\beta H_0)\}$, if

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

independently for all j and

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

with

$\eta_{lk}^* \sim H_0, w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1} (1 - u_{sk}^*), \pi_k^* = v_k^* \prod_{s=1}^{k-1} (1 - v_s^*), v_k^* \sim \text{Be}(1, \alpha)$ and $u_{lk}^* \sim \text{Be}(1, \beta)$ and the $\{\eta_{lk}^*\}$, $\{u^*\}$ and $\{v_k^*\}$ all independent.

The nested Dirichlet process uses the stick-breaking representation twice: 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 nested Dirichlet process allows nonparametric estimation of the probability measures $\{H_1, \dots, H_J\}$, while borrowing information across them by creating clusters of identical measures. The baseline measure H_0 can be interpreted as a mean distribution around which the condition-specific probability measures are centred, 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 for the nested Dirichlet process can be carried out using truncations of the stick-breaking representations (8) and (9). For a detailed discussion, see Rodríguez et al. (2008).

Consider now the problem of inferring multiple curves $f_1, \dots, f_J$, using the data $z_1^{n_1}, \dots, z_J^{n_J}$, where $z_j^{n_j} = (z_{1j}, \dots, z_{n_j j})$ is the set of n_j observations obtained under experimental condition j . The model described in §2 could be extended to accommodate these multiple curves by using conditionally independent Dirichlet processes as the random measures 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, and 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, different types of processes can be used to induce such dependence. For example, if the goal is global functional clustering, the nested Dirichlet process described above 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) is a reasonable choice. Finally, in a spatial data analysis setting, an extension of the spatial Dirichlet process (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 §2.

As an illustration, consider a model for functional clustering using mixtures of nested Dirichlet processes. By using the nested Dirichlet process as a prior on the mixing distributions $\{H_1, \dots, H_J\}$ used to estimate the joint probability distributions $\{p_I(y, x), \dots, p_J(y, x)\}$, 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. In particular, consider the following extension of the single curve model described in §2, with

$$\begin{aligned} z_{ij} &= (y_{ij}, x_{ij}) \sim N_{p+q}(\theta_{ij}, \Sigma_{ij}), & i=1, \dots, n_j; j=1, \dots, J, \\ (\theta_{ij}, \Sigma_{ij}) &\sim H_j, & \mathcal{H} = \{H_1, \dots, H_J\} \sim DP\{\alpha DP(\beta H_0)\} \end{aligned} \quad (10)$$

where, as before, $H_0 = NIW(\theta_0, \kappa_0, v_0, \Sigma_0)$, $\theta_0 \sim N(\theta_{00}; D_{00})$, $\Sigma_{00} \sim Wi(\gamma, \Sigma_{00})$, $\kappa_0 \sim Ga(a_\kappa, b_\kappa)$, $\alpha \sim Ga(a_\alpha, b_\alpha)$ and $\beta \sim Ga(a_\beta, b_\beta)$. Let $n = \sum_{j=1}^J n_j$ and let $H_j^n(\cdot | z_1^{n_1}, \dots, z_J^{n_J})$ be the posterior distribution of the parameters (θ, Σ) under experimental condition j . Estimates of the regression functions $\{f_1, \dots, f_J\}$ can be obtained from the posterior conditional expectations as

$$\begin{aligned}
 f_j^n(x) &= E_{H_j^n}(y|x, z_1^{n_1}, \dots, z_J^{n_J}) \\
 &= \int \frac{\{\theta_y + \Sigma_{yx} \Sigma_{xx}^{-1}(x - \theta_x)\} \phi_p(x|\theta_x, \Sigma_{xx})}{\int \phi_p(x|\theta_x, \Sigma_{xx}) H_{0j}^n(d\theta, d\Sigma|z_1^{n_1}, \dots, z_J^{n_J})} H_{0j}^n(d\theta, d\Sigma|z_1^{n_1}, \dots, z_J^{n_J}).
 \end{aligned} \tag{11}$$

where, as before, $H_{0j}^n(d\theta, d\Sigma|z_1^{n_1}, \dots, z_J^{n_J})$ is the mean posterior mixing distribution in group j .

Although an explicit form is not available for the estimated regression function $f_j^n(x)$ or the estimated density $g_j^n(y|x)$, they can be easily approximated for any x , and hence for any dense grid of x 's, using Markov chain Monte Carlo methods. Functional clustering in quantile 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.

From the definition of the nested Dirichlet process, it is clear that the model assumes that the curves are a priori exchangeable, and that there is a non zero 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 nonparametrically, reproducing the single curve model discussed in §2, 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 $(\theta_{ij}, \Sigma_{ij}) = (\theta_{i'j'}, \Sigma_{i'j'}) = (\theta_{lk}^*, \Sigma_{lk}^*)$ for some l . Therefore, the parameter β , in controlling the number of distinct $(\theta_{lk}^*, \Sigma_{lk}^*)$, controls the nonlinearity 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 the 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

Conditions for the consistency of the functional estimates obtained using the approach of §§2 and 3 have been discussed in unpublished work A. Rodríguez unpublished 2007 Ph.D. dissertation from Duke University. In brief, weak consistency in the nonparametric model used for estimation of the set of unknown joint densities $\{g_1(x, y), \dots, g_J(x, y)\}$ implies pointwise consistency of the functional estimates $\{f_1(x), \dots, f_J(x)\}$. Although general consistency results for the class of dependent Dirichlet processes are still an open problem, weak consistency can easily be proved in the class of nested Dirichlet process models.

Theorem 1

Suppose that the true densities generating the data, $\{g_j^0\}_{j=1}^J$, all 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 nested Dirichlet process mixture prior is weakly consistent as the sample sizes $n_1, \dots, n_J$ all grow to infinity.

The argument is straightforward. 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 nested Dirichlet process 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 size in each of the groups increases. Conditional on the partition, the nested Dirichlet process implies independent Dirichlet process mixtures of Gaussian priors for the cluster-specific densities. Hence, posterior consistency follows automatically from the results of Ghosal et al. (1999).

Corollary 1

Each functional estimate arising from the model in (10) is pointwise consistent on a dense subset of the class of bounded integrable functions with domain on a compact set D_x .

The compact support assumption is not particularly restrictive. Observations are virtually always collected over compact sets as there are physical limits to the values that the covariates can take. Also, the impact of the compact support assumption can typically be moderated by considering a support D_x that is larger than the range of covariates actually observed, making edge effects negligible. Moreover, the assumption of compact-support is sufficient but not necessary. It can be relaxed if additional assumptions, about the rate of growth of the function as $x \rightarrow \infty$, are introduced which rule out exponential growth to infinity.

5 An illustration: clustering temperature profiles in the North Atlantic

Conductivity and Temperature at Depth profiles are regularly used in oceanography to study the physical properties of a water column. The profiler is a torpedo-shaped instrument that is attached to a conducting wire and lowered to prespecified depths. At each depth, information about pressure, temperature and conductivity is sent back to the ship through the wire. The result from this measurement process is a sample from the functions relating temperature, our dependent variable y , with depth, our independent variable x , at each location and time.

Profiles can be highly nonlinear, and understanding their patterns of spatio-temporal evolution can help scientists assess the magnitude and consequences of global phenomena like El Niño and the process of global warming. However, the profiles are obtained very sparsely, both in space and time, and do not necessarily change smoothly with latitude or longitude, making standard spatio-temporal models hard to justify in many geographic regions. An alternative approach 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 signalling boundary regions.

We apply the approach discussed in §3 to 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 10m 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 profile 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.

Computation was carried out using the algorithm described in the Appendix, which is an extension of the Gibbs sampler described in Rodríguez et al. (2008). 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 $D = \Sigma/100$, $a_k = 1$, and $b_k = 100$ in such a way that $E(\kappa_0) = 0.01$, $v_0 = 3$ and $\gamma = 3$. For the precision parameters, and based on the discussion in Rodríguez et al. (2008), we pick $a_\alpha = b_\alpha = a_\beta = b_\beta = 3$. Sensitivity analysis shows that inference tends to be robust to moderate changes in most hyperparameters, as long as θ_{00} and Σ_{00} respect the scale of the data. All inferences are based on 50,000 samples obtained after a burn-in of 10,000 iterations. Different starting points for the sampling algorithm were tried, including assigning all curves to a common cluster, a random assignment of curves to clusters, and some heuristics based on linear fits and simple hierarchical clustering. Although all led to equivalent results, the 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. The big cluster corresponds to the site off the coast of Africa, and the model shows a small probability, of about 0.03, of this cluster being broken into two distinct groups based on the different behaviours observed below 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 correspond to 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, which meet by the coast of Nova Scotia. These two water masses do not mix, producing very different profiles in close geographic areas.

Figure 2 shows functional estimates and pointwise credible bands for each of the four groups. The behaviour of the profiles is clearly nonlinear and some of them are not even monotone, characteristics that are consistent with previous studies. 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. Note that the curves off Nova Scotia, corresponding Fig. 2 (b) and (c), show a behaviour that is consistent with our hypothesis about oceanic currents. The cluster closest to the coast, Fig. 2 (b) 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, Fig. 2 (c), 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 below 600 m. The only unappealing feature of our functional estimates is the bump in locations off Portugal, see Fig. 2 (d). This bump is accompanied by very wide credible bands, indicating a high level of uncertainty about the shape of the function. These features are due to the sparseness in the data for these locations, where observations below the 300 m mark were collected only every 100 m. This phenomenon is common with this type of model, where large gaps in the predictor space accompanied with repeated measurements at the observed locations can produce unstable functional estimates within the gaps, especially at covariate values where there is sign change in the slope of the function. However, we do not expect this instability to affect the clustering results.

As discussed in the introduction, one important feature of this class of models is its ability to deal automatically with inverse problems. Figure 3 shows density estimates of the depth of the 12°C isothermal layer corresponding to four curves in the sample, one for each of the clusters discussed above. There is little uncertainty in Fig. 3 (a) and (c), which is consistent with the functional estimates discussed above. In Fig. 3 (b) the density shows clear bimodality, which is again in line with Fig. 2. Finally, Fig. 3 (d) shows a main mode around 27m with a heavy left-hand tail and a second mode close to the 800 meters mark. Both of these features can be

readily explained by the additional uncertainty in the functional estimates due to the sparsity in the data for this group.

We compared our results with those obtained by fitting natural cubic splines with a variable number of equally spaced knots and then running an algorithm for hierarchical clustering on the coefficients. In addition to the obvious shortcomings of this simple approach, including its inability to provide adequate estimates of uncertainty in the clustering structure or the depth of isothermal layers, results were very sensitive to the number of knots used in this model, and clustering results differed significantly from ours for censored curves. The first phenomenon can be explained by the difference in smoothness across clusters, which makes the choice of a common and fixed number of knots across clusters inappropriate. If a small number of knots is used, estimation of the less smooth curves suffers, yielding wide confidence bands; if too many knots are used, then smoother curves are over-fitted, again leading to wider uncertainty bands. On the other hand, for censored curves, that is, curves that do not have observations at the lower depths, information about some of the spline coefficients is very limited. Hence, maximum likelihood estimates for these coefficients tend to have high variance and may not even exist. This problem tends to produce estimates of the spline coefficients for censored curves having some of the estimated coefficients very far from the estimates for other curves. By misinterpreting such estimates as providing evidence that the true curves are actually different, simple functional clustering procedures tend to allocate censored curves to singleton clusters. Such problems provide a compelling motivation for our flexible semiparametric Bayes approach, which automatically borrows information across the different curves to aid in interpolating across sparse data regions.

6 Discussion

We are currently considering variations of these ideas for local functional clustering and spatial functional data analysis. We would like to stress that these models are particularly appealing in the context of observational studies, where covariates are collected along with outcome variables and are therefore random. Indeed, covariates meant to be fixed by design are often measured with error. For example, in the context of our illustration, the depth is actually measured with error since, first, it is reconstructed from pressure measurements and, secondly, the calculations used in the reconstruction involve constants that vary slightly depending on unmeasured factors. However, these methods can also be useful in the context of designed experiments, where there is a history of related double-kernel approaches in the frequentist literature on conditional distribution estimation; see Fan & Yim (2004) and references therein. Indeed, having the marginal distribution of x influence the functional estimates has distinct advantages. As the formulae for the estimate of the conditional $p(y/x)$ show, the marginal distribution $p(x)$ enters our estimate only through the weights assigned to each mixture component. This actually allows the models to choose the number and locations of the knots adaptively, which is usually a complicated task in spline models.

Acknowledgement

This work is part of the first author's dissertation completed at Duke University and was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Science. The authors would like to thank Susan Lozier and Robert Wolpert for helpful comments, as well as the two anonymous referees and the Editor for their insights.

A Appendix: Computational implementation

We implement the nested Dirichlet process model using the two-level truncation algorithm described in Rodríguez et al. (2008). This algorithm uses a finite mixture to approximate each of the stick-breaking processes involved in the definition of the nested Dirichlet process. We

used truncation levels set at $K = L = 55$ atoms. These levels 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 $(\theta_{ij}, \Sigma_{ij}) = (\theta_{lk}^*, \Sigma_{lk}^*)$. Once adequate starting values for the parameters have been chosen, computation proceeds through the following steps:

Step 1. Sample ζ_j for $j = 1, \dots, J$ from a multinomial distribution with probabilities

$$\text{pr}(\zeta_j = k | \dots) = q_k^j \propto w_k^* \prod_{i=1}^{n_j} \sum_{l=1}^L \pi_{lk} \phi_{p+q}(z_{ij} | \theta_{lk}^*, \Sigma_{lk}^*), \quad k=1, \dots, K.$$

Step 2. Sample ξ_{ij} for $j = 1, \dots, J$ and $i = 1, \dots, n_j$ from a multinomial distribution with

$$\text{pr}(\xi_{ij} = l | \dots) = b_{ij}^l \propto \pi_{l, \zeta_j}^* \phi_{p+q}(z_{ij} | \theta_{l, \zeta_j}^*, \Sigma_{l, \zeta_j}^*), \quad l=1, \dots, L.$$

Step 3. Sample bottom-level probabilities π_k^* by generating

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

where m_k is the number of curves assigned to component k , and

$$\pi_k^* = u_k^* \prod_{s=1}^{k-1} (1 - u_s^*).$$

Step 4. Sample the top-level probabilities w_{lk}^* by generating

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

where n_{lk} is the number of observations assigned to atom l of curve k , and

$$w_{lk}^* = v_{lk}^* \prod_{s=1}^{l-1} (1 - u_{sk}^*).$$

Step 5. Sample the atoms $(\theta_{lk}^*, \Sigma_{lk}^*)$ from

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

where

$$\begin{aligned} \widehat{\theta}_{lk} &= \frac{n_{lk}}{\kappa_0 + n_{lk}} \bar{z}_{lk} + \frac{\kappa_0}{\kappa_0 + n_{lk}} \theta_0, & \widehat{\kappa}_{lk} &= \kappa_0 + n_{lk}, \\ \widehat{\nu}_{lk} \widehat{\Sigma}_{lk} &= \nu_0 \Sigma_0 + n_{lk} \bar{S}_{lk} + \frac{\kappa_0 n_{lk}}{\kappa_0 + n_{lk}} (\bar{z}_{lk} - \theta_0)(\bar{z}_{lk} - \theta_0)', & \widehat{\nu}_{lk} &= \nu_0 + n_{lk}, \\ \bar{S}_{lk} &= \frac{1}{n_{lk}} \sum_{\{i,j: \zeta_j=k, \xi_{ij}=l\}} (z_{ij} - \bar{z}_{lk})(z_{ij} - \bar{z}_{lk})', & \bar{z}_{lk} &= \frac{1}{n_{lk}} \sum_{\{i,j: \zeta_j=k, \xi_{ij}=l\}} z_{ij}, \end{aligned}$$

and n_{lk} is the number of observations assigned to atom (l, k) .

Step 6. Sample the baseline mean θ_0 from

$$\theta_0 | \dots \sim N \left\{ (D_{00}^{-1} + \bar{D})^{-1} (D_{00}^{-1} \theta_{00} + \bar{d}), (D_{00}^{-1} + \bar{D})^{-1} \right\}$$

where

$$\bar{D} = \kappa_0 \sum_{\{l,k:n_{lk} \neq 0\}} \Sigma_{lk}^{*-1}, \quad \bar{d} = \kappa_0 \sum_{\{l,k:n_{lk} \neq 0\}} \Sigma_{lk}^{*-1} \theta_{lk}^*.$$

Step 7. Sample the variance of the baseline measure, Σ_0 from

$$\Sigma_0 | \dots \sim Wi \left(\gamma + c\nu_0, \gamma \Sigma_{00}^{-1} + \frac{\nu_0}{\kappa_0} \bar{D} \right)$$

where c is the number of nonempty components.

Step 8. Sample the mean precision parameter κ_0 from

$$\kappa_0 | \dots \sim Ga \left(a_\kappa + \frac{c(p+q)}{2}, b_\kappa + \frac{1}{2} \sum_{\{l,k:n_{lk} \neq 0\}} (\theta_{lk}^* - \theta_0)' \Sigma_{lk}^{*-1} (\theta_{lk}^* - \theta_0) \right).$$

Step 9. Sample the concentration parameters α and β from

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

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

References

- Albert J, Chib S. Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc* 1993;88:669–679.
- Antoniak C. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist* 1974;2:1152–1174.
- Basu S, Chib S. Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Am. Statist. Assoc* 2003;98:224–235.
- Behseta S, Kass RE, Wallstrom GL. Hierarchical models for assessing variability among functions. *Biometrika* 2005;92:419–434.
- Bigelow JL, Dunson DB. Bayesian semiparametric joint models for functional predictors. *J. Am. Statist. Assoc.* To appear. 2008
- Bush CA, MacEachern SN. A semiparametric Bayesian model for randomized block designs. *Biometrika* 1996;83:275–285.
- DeIorio M, Müller P, Rosner GL, MacEachern SN. An anova model for dependent random measures. *J. Am. Statist. Assoc* 2004;99:205–215.
- Dunson D. Bayesian dynamic modeling of latent trait distributions. *Biostatistics* 2006;7:551–568. [PubMed: 16488893]
- Dunson DB, Pillai N, Park J-H. Bayesian density regression. *J. R. Statist. Soc. B* 2007;69:163–183.

- Escobar MD. Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Assoc* 1994;89:268–277.
- Escobar MD, West M. Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc* 1995;90:577–588.
- Fan JQ, Yim TH. A cross validation method for estimating conditional densities. *Biometrika* 2004;91:819–834.
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann. Statist* 1973;1:209–230.
- Gelfand AE, Kottas A, MacEachern SN. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Am. Statist. Assoc* 2005;100:1021–1035.
- Ghosal S, Ghosh J, Ramamoorthi RV. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist* 1999;27:143–158.
- Green P, Richardson S. Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Statist* 2001;28:355–375.
- Griffin JE, Steel MFJ. Order-based dependent Dirichlet processes. *J. Am. Statist. Assoc* 2006;101:179–194.
- Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc* 2001;96:161–173.
- Jain S, Neal RM. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comp. Graph. Statist* 2004;13:158–182.
- MacEachern, SN. *Proc. Set. Bayes. Statist. Sci. Alexandria, Virginia: Am. Statist. Assoc.; 1999. Dependent nonparametric processes; p. 50-55.*
- Morris JS, Carroll RJ. Wavelet-based functional mixed models. *J. R. Statist. Soc. B* 2006;68:179–199.
- Müller P, Erkanli A, West M. Bayesian curve fitting using multivariate normal mixtures. *Biometrika* 1996;83:67–79.
- Müller P, Quintana F, Rosner G. Hierarchical meta-analysis over related non-parametric Bayesian models. *J. R. Statist. Soc. B* 2004;66:735–749.
- Pennell ML, Dunson DB. Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* 2006;62:1044–1052. [PubMed: 17156278]
- Ray S, Mallick BK. Functional clustering by Bayesian wavelet methods. *J. R. Statist. Soc. B* 2006;68:305–332.
- Rice JA, Silverman BW. Estimating the mean and covariance structure nonpara-metrically when the data are curves. *J. R. Statist. Soc. B* 1991;53:233–243.
- Rodríguez A, Dunson DB, Gelfand AE. The nested Dirichlet process (with discussion). *J. Am. Statist. Assoc.* To appear. 2008
- Sethuraman J. A constructive definition of Dirichlet priors. *Statist. Sinica* 1994;4:639–650.
- Teh YW, Jordan MI, Beal MJ, Blei DM. Sharing clusters among related groups: Hierarchical Dirichlet processes. *J. Am. Statist. Assoc* 2006;101:1566–1581.
- Wu H, Zhang JT. Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Assoc* 2002;97:883–897.

Figure 1.

Oceanography study. 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. Four tight groups are clear from the plot, roughly agreeing with spatial location.

Figure 2.

Oceanography study. Fitted curves (solid line), along with pointwise credible bands (dashed lines) for each of the 4 groups of curves. Credible bands are tight even in groups with very few curves. Panel (d) is an exception, and the additional uncertainty is due to very sparse data.

Figure 3. Oceanography study. Density estimates for the depth of the isothermal layer at 12°C for representative curves on each cluster. These are preferable to more straightforward measures like the conditional mean when multi-modality is expected, like in panel (b). Results are in line with the functional estimates shown above.