SOME ADVANCES IN BAYESIAN NONPARAMETRIC MODELING

by

Abel Rodríguez

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Alan E. Gelfand, Supervisor

_____
Dr. David B. Dunson, Supervisor

_____
Dr. Mike West

_____
Dr. Robert Wolpert

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2007

ABSTRACT

(Statistics)

SOME ADVANCES IN BAYESIAN NONPARAMETRIC MODELING

by

Abel Rodríguez

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Alan E. Gelfand, Supervisor

_____
Dr. David B. Dunson, Supervisor

_____
Dr. Mike West

_____
Dr. Robert Wolpert

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2007

Abstract

Many modern scientific problems involve outcomes that are complex, infinite dimensional objects like curves and distributions. Often, no satisfactory statistical methodology is available for inference in these types of problems. Recent research in Bayesian nonparametric methods has focused on extending existing models to accommodate simultaneous inferences for multiple dependent distributions. This dissertation focuses on problems of density estimation on *collections of distributions* using extensions of the Dirichlet processes, as well as their application to nonparametric regression.

The dissertation can be broadly divided in three semi-autonomous pieces. In the first part, comprising chapters 2 to 4, we develop models for the joint estimation of collections of densities in two specific contexts: 1) multicenter studies, where distributions are assumed to form clusters indicating common underlying characteristics and 2) time series where distributions evolve in discrete-time. We demonstrate the versatility of the models through applications in epidemiology, public health and finance.

In the second part, which involves chapter 5 and 6, we frame nonparametric regression as a density estimation problem. First, we show that consistency of the density estimates automatically induces pointwise consistency of the functional estimates. From there, we develop methods for functional data analysis based on dependent Dirichlet processes. Specifically, we discuss applications to functional clustering and functional spatial data analysis. Examples of these methods are drawn from oceanography and public health.

Finally, chapter 7 introduces a novel nonparametric prior on the space of stochastic processes that provides a flexible alternative to the Gaussian process. This class of models has few precedents in the literature and is different from the models for

collection of distributions that we developed in the first part of the dissertation. As an application, we discuss a stochastic volatility model for option pricing.

Contents

List of Tables

List of Figures

Acknowledgements

There are many people that I am indebted to, but surely the most important is my wife Angela. She has been the light at the end of many dark days; she believed in me even when I did not believe in myself.

I also have a debt of gratitude to my committee, especially my supervisors, David Dunson and Alan Gelfand. They provided intellectual stimulation and were there for me every time I needed them. I cannot imagine finding better advisors and friends; I hope I can live up to their example.

My classmates and friends, present and past, have also been a profound source of strength and inspiration. Too many to name, from the UCAB and USB to those I met at Duke, your support has been fundamental for the completion of this work.

Finally I would like to thank my parents, who could not be here with me in this happy hour. In the end, they made me who I am, with my virtues and flaws.

Thanks to all!!!

Abel

Chapter 1

Introduction

Parametric statistical methods are inference procedures where the outcomes in the model are assumed to follow probability distributions that belong to a family that is determined except for a finite number of parameters, for example, the normal distribution with unknown mean and variance. Some of the statistical methods most widely used by practitioners like multiple linear regression, analysis of variance and generalized linear models, are parametric in nature.

The beauty of parametric methods is their relative simplicity: with a finite parameter space and some reasonable assumptions on the parametric families involved, deriving distributional theory, defining prior distributions and/or obtaining posterior distributions is relatively straightforward even for complicated and highly structured models.

In contrast, nonparametric methods try to avoid assumptions about the probability distributions in order to generate methods that can be used in settings where regular parametric assumptions do not work. Although applicable in more general circumstances, nonparametric models can lead to very complex mathematics in all but the simplest models. Also, there is an implicit tradeoff between the generality of nonparametric tests and the power to detect differences between populations.

From a frequentist perspective, a parametric t-test has a higher power if the normality assumption is indeed true, but might badly under perform the sign test if it is false, given the same type I error. From a Bayesian perspective, posterior distributions obtained from nonparametric models tend to have larger variances than their parametric counterparts.

Nonparametric methods have a long history in modern frequentist statistics, starting with Fisher's exact test (Fisher, 1922). One simple yet enlightening example of a classical nonparametric method is the sign test. Let $x_i \sim F$ for $i = 1, \ldots, n$, μ be the median of F, and suppose that we are interested in testing $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$, where μ_0 is some fixed number. Defining t as the number of values in the sample that are greater than μ_0, it is clear that $t \sim \mathsf{Bin}(n, 1/2)$, no matter what the true distribution F is.

In Bayesian statistics, nonparametric models are constructed through priors on rich families of distributions. Therefore, the term Bayesian nonparametrics is really a misnomer. Bayesian nonparametric models are not parameter free, but have an infinite number of parameters. Raiffa and Schlaifer (1961) and Ferguson (1973) in their seminal work on Bayesian nonparametrics mention some characteristics that should be kept in mind when constructing priors on spaces of distributions:

1. The class should be analytically tractable. Therefore, the posterior distribution should be easily computed, either analytically or through simulation.

2. The class should be rich, in the sense of having a large enough support.

3. The hyperparameters defining the prior should be easily interpreted.

Although it is not always possible to completely satisfy all of the requirements mentioned above, this dissertation will emphasize the importance of these features when developing our own nonparametric models.

This chapter makes a quick review of current Bayesian nonparametric models for distribution functions, making special emphasis on the Dirichlet process. Bayesian nonparametric methods in the context of regression models will be briefly discussed in chapters 3 and 5.

1.1 The Dirichlet Process

The Dirichlet Process (DP) (Ferguson, 1973, 1974; Blackwell and MacQueen, 1973; Sethuraman, 1994) is the base for the most widely used nonparametric models for random distributions in Bayesian statistics, mainly due to the availability of efficient computational techniques. Some recent applications of the Dirichlet Process include finance (Kacperczyk *et al.*, 2003), econometrics (Chib and Hamilton, 2002; Hirano, 2002), epidemiology (Dunson, 2005), genetics (Medvedovic and Sivaganesan, 2002; Dunson *et al.*, 2007a), medicine (Kottas *et al.*, 2002; Bigelow and Dunson, 2007) and auditing (Laws and O'Hagan, 2002).

1.1.1 Definition of the Dirichlet Process

Consider the probability spaces $(\mathbf{\Theta}, \mathcal{B}, P)$ and $(\mathbf{P}, \mathcal{C}, Q)$ such that $P \in \mathbf{P}$. In most applications, $\mathbf{\Theta} \subseteq \mathbb{R}^d$, $\mathcal{B}$ corresponds to the Borel σ-algebra of subsets of $\mathbb{R}^d$ and $\mathbf{P}$ is the space of probability measures over $(\mathbf{\Theta}, \mathcal{B})$, but most of the results mentioned in this section extend to any complete and separable metric space $\mathbf{\Theta}$. We will refer to $(\mathbf{\Theta}, \mathcal{B}, P)$ as the *base space* and to $(\mathbf{P}, \mathcal{C}, Q)$ as the *distributional space*. Given a finite, nonnegative, nonnull measure ν on $(\mathbf{\Theta}, \mathcal{B})$, the Dirichlet Process with base probability measure $H(\cdot) = \nu(\cdot)/\nu(\mathbf{\Theta})$ and precision $\alpha = \nu(\mathbf{\Theta})$, denoted as $\mathsf{DP}(\alpha H)$, is a probability measure Q over the space $(\mathbf{P}, \mathcal{C})$ such that $(P(B_1), \ldots, P(B_k)) \sim \mathsf{Dir}(\alpha H(B_1), \ldots, \alpha H(B_k))$ for any finite and measurable partition $B_1, \ldots, B_k$ of $\mathbf{\Theta}$, where $\mathsf{Dir}(\gamma_1, \ldots, \gamma_k)$ denotes the k-dimensional Dirichlet

distribution with parameters $\gamma_1, \ldots, \gamma_k$ (see appendix A).

The Dirichlet process can be alternatively characterized in terms of its predictive rule (Blackwell and MacQueen, 1973). If $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{n-1})$ is an iid sample from $P \sim \mathsf{DP}(\alpha H)$, we can integrate out the unknown P and obtain the conditional predictive distribution of a new observation,

$$\boldsymbol{\theta}_n | \boldsymbol{\theta}_{n-1}, \ldots, \boldsymbol{\theta}_1 \sim \frac{\alpha}{\alpha + n - 1} H + \sum_{l=1}^{n-1} \frac{1}{\alpha + n - 1} \delta_{\boldsymbol{\theta}_l}$$

where $\delta_{\boldsymbol{\theta}_l}$ is the Dirac probability measure concentrated at $\boldsymbol{\theta}_l$. Exchangeability of the draws ensures that the full conditional distribution of any $\boldsymbol{\theta}_l$ has this same form. This result, which relates the Dirichlet process to a Pólya urn, is the basis for the usual computational tools used to fit models based on the Dirichlet process.

The Dirichlet process can also be regarded as a type of *stick-breaking prior* (Sethuraman, 1994; Pitman, 1996; Ishwaran and James, 2001; Ongaro and Cattaneo, 2004). A stick-breaking prior on the space $\mathbf{P}$ has the form

$$P^K(\cdot) = \sum_{k=1}^{K} w_k \delta_{\boldsymbol{\theta}_k}(\cdot) \qquad \boldsymbol{\theta}_k \sim H$$

$$w_k = z_k \prod_{l=1}^{k-1} (1 - z_l) \qquad z_k \sim \begin{cases} \mathsf{beta}(a_k, b_k) & \text{if } k < K \\ 1 & \text{if } k = K \end{cases}$$

where the number of atoms K can be finite (either known or unknown) or infinite. For example, taking $K = \infty$, $a_k = 1 - a$ and $b_k = b + ka$ for $0 \leq a < 1$ and $b > -a$ yields the two-parameter Poisson-Dirichlet Process, also known as Pitman-Yor Process (Pitman, 1996), with the choice $a = 0$ and $b = \alpha$ resulting in the Dirichlet Process (Sethuraman, 1994).

The stick-breaking representation is probably the most versatile definition of the Dirichlet Process. It has been exploited to generate efficient alternative MCMC algorithms and as the starting point for the definition of many generalizations that allow

dependence across a collection of distributions, including the DDP (MacEachern, 2000), the πDDP (Griffin and Steel, 2006b) and the GSDP (Duan *et al.*, 2007).

Finally, the Dirichlet Process can be obtained as the asymptotic limit of certain finite mixture models (Green and Richardson, 2001; Ishwaran and Zarepour, 2002). In particular consider the finite-dimensional Dirichlet-Multinomial prior

$$P^K(\cdot) = \sum_{k=1}^{K} w_k \delta_{\boldsymbol{\theta}_k}(\cdot) \qquad \mathbf{w} \sim \mathsf{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \qquad \boldsymbol{\theta}_k \sim H$$

which differs from a truncated stick-breaking representation of the Dirichlet Process in the way the weights have been defined. Ishwaran and Zarepour (2002) prove that for each measurable function g which is integrable with respect to H, $\int g(\boldsymbol{\theta}) P^K(d\boldsymbol{\theta}) \xrightarrow{\text{p}} \int g(\boldsymbol{\theta}) P(d\boldsymbol{\theta})$ where $P \sim \mathsf{DP}(\alpha H)$, i.e., the finite-dimensional Dirichlet-Multinomial prior converges in distribution to the Dirichlet process. This result not only provides another useful approximation, but also justifies frequently used finite mixture models as approximating a DP.

1.1.2 Properties of the Dirichlet process

From the stick-breaking construction we can easily see that draws from a Dirichlet process are discrete distributions almost surely. It also provides a simple framework to calculate moments of the process. Note that for any measurable set $A \in \mathcal{B}$, $P(A)$ is a random quantity and

$$\begin{aligned}\mathbb{E}(P(A)) &= \sum_{l=1}^{\infty} \mathbb{E}(w_k)\mathbb{E}(\delta_{\boldsymbol{\theta}_k}(A)) \\ &= H(A)\sum_{l=1}^{\infty} E(w_k) = H(A).\end{aligned}$$

Using a similar argument

$$\mathbb{V}(P(A)) = \frac{H(A)(1-H(A))}{\alpha+1}.$$

In order to better understand the role of the parameters H and α, we show in Figure 1.1 approximate simulations of a Dirichlet Process with a standard Gaussian baseline measure and different values of the precision parameters. These were obtained by truncating the stick-breaking process when the leftover mass was smaller than $\epsilon = 10^{-6}$.

Note that in all cases the samples are centered on the baseline measure. However, for low values of α, the sampled distributions vary widely around the baseline measure and tend to have very few important atoms that concentrate most of the probability. As the precision parameter increases, the distributions look smoother and they tend to be very close to the standard Gaussian. These results justify the interpretation of H and α as location and precision/roughness parameters respectively.

Conjugacy is another appealing property of the Dirichlet process. If $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim P$ and $P \sim \mathsf{DP}(\alpha H)$, then

$$P|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim \mathsf{DP}\left(\alpha H + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}\right).$$

Therefore, the optimal estimator under squared error loss for P is

$$\hat{P}(\cdot) = \frac{\alpha}{\alpha + n} H(\cdot) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}, (\cdot)$$

which converges to the empirical distribution as $n \to \infty$.

Antoniak (1974) studies the properties of draws from a distribution that follow a Dirichlet process. In particular, he proves that, if H is nonatomic, the probability of k distinct values on a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ of size n is

$$\mathbb{P}(k) = c_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \tag{1.1}$$

for $k = 1, \ldots, n$, where $c_n(k)$ is a constant that can be obtained using recurrence formulas for Stirling numbers. The expected number of distinct values can be

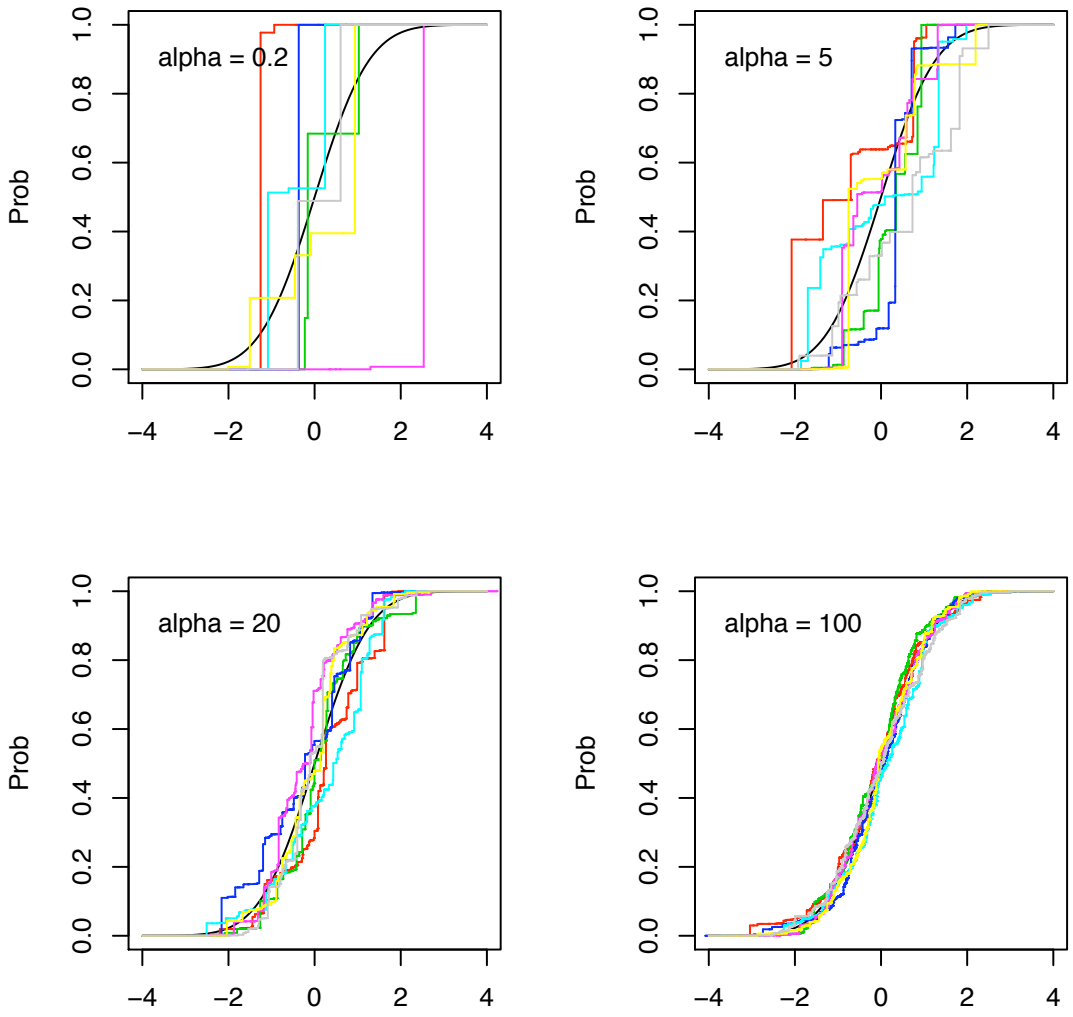


Figure 1.1: Samples of a DP process centered on a standard Gaussian distribution, for different precision parameters.

calculated as

$$\mathbb{E}(k|\alpha, n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \approx \alpha \log\left(\frac{\alpha + n}{\alpha}\right)$$

These results will be used later to construct computational algorithms that treat α as an unknown parameter and to elicit prior distributions for this parameter.

1.1.3 Mixtures of Dirichlet Processes

Antoniak (1974) also generalized the basic Dirichlet process by considering random baseline measures ν, resulting in the so-called Mixture of Dirichlet Process (MDP) model. Given an index space $(\mathbf{U}, \mathcal{A})$, let $\nu(\cdot, \cdot)$ be such that $\nu(\cdot, u)$ is a finite, non-negative and nonnull measure on $(\mathbf{\Theta}, \mathcal{B})$ for every $u \in \mathbf{U}$ and $\nu(B, \cdot)$ be measurable on $(\mathbf{U}, \mathcal{A})$ for every $B \in \mathbf{B}$. We say that P is distributed as a mixture of Dirichlet processes if for any measurable partition $(B_1, \ldots, B_k)$ of $\mathbf{\Theta}$

$$(P(B_1), \ldots, P(B_k)) \sim \int_U \mathsf{Dir}(\nu(B_1, u), \ldots, \nu(B_k, u)) T(du)$$

where the mixture distribution T is defined on $(\mathbf{U}, \mathcal{A})$. MDP priors, just like regular DP's, are almost surely discrete and conjugate. That is, if $X \sim P$ and P follows a mixture of Dirichlet processes, then $P|X$ is again an MDP with updated parameters.

1.2 Dirichlet Process Mixtures

Since the DP and MDP models put probability one on the space of discrete measures, they are typically not good choices for modeling continuous data. Instead, they are more naturally employed as priors on the random mixing distribution over the parameters of a continuous distribution K with density k,

$$\mathbf{z} \sim g(\cdot) \qquad g(\cdot) = \int k(\cdot|\boldsymbol{\theta}) H(d\boldsymbol{\theta}) \qquad H \sim \mathsf{DP}(\alpha H_0), \tag{1.2}$$

resulting in a DP mixture (DPM) model (Lo, 1984; Escobar, 1994; Escobar and West, 1995). 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 $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $k(\cdot|\boldsymbol{\theta}) = \phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the p-variate normal kernel with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

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

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

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

$$\begin{aligned} g^n(\mathbf{z}) &= \mathbb{E}\left[\int k(\mathbf{z}|\boldsymbol{\theta}) H^n(d\boldsymbol{\theta}|\mathbf{z}^n)\right] \\ &= \int k(\mathbf{z}|\boldsymbol{\theta}) \mathbb{E}\left[H^n(d\boldsymbol{\theta}|\mathbf{z}^n)\right] \\ &= \int k(\mathbf{z}|\boldsymbol{\eta}) \frac{\alpha H_0(\boldsymbol{\eta}) + \sum_{l=1}^{n} \delta_{\boldsymbol{\theta}_l}(\boldsymbol{\eta})}{\alpha + n} p(d\boldsymbol{\theta}_1, \ldots, d\boldsymbol{\theta}_n|\mathbf{z}^n). \end{aligned} \tag{1.3}$$

Density estimates arising from location-and-scale DP mixtures can be interpreted as Bayesian kernel density estimates with adaptive bandwidth selection. This interpretation is extremely appealing because it provides a direct link with well-known frequentist techniques and demonstrates the versatility of the model.

Due to the discrete nature of the DP prior, the DPM model divides the observations into independent groups, each one of them assumed to follow a distribution implied by the kernel k. Therefore, DPM models can be used for clustering as well as for density estimation. In this setting, the model automatically allows for an unknown number of clusters, with equation 1.1 providing the implicit prior distribution.

1.2.1 Computation for Dirichlet Process Mixtures

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; MacEachern and Müller, 1998; Neal, 2000), 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).

For computational purposes, it is convenient to rewrite model 1.2 using latent variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ corresponding to observations $\mathbf{z}_1, \ldots, \mathbf{z}_n$. In turn, these latent variables can be rewritten in terms of a set of $k \leq n$ unique values $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$ and a set of indicators $\zeta_1, \ldots, \zeta_n$, such that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{\zeta_i}^*$.

Pólya urn samplers, also called marginal samplers, are popular in practice because they are relatively easy to implement and produce exact samples from the posterior distribution of $\boldsymbol{\theta}$. However, they are more useful when the baseline measure H_0 is conjugate to the kernel k. Escobar and West (1995) original algorithm uses the Pólya urn directly to simultaneously sample group indicators and group parameters. They note that

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \mathbf{z}) = q_{i0}p(\boldsymbol{\theta}_i|\mathbf{z}_i, H_0) + \sum_{l=1, l\neq i}^{n} q_{il}\delta_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}_i)$$

where $q_{i0} = \alpha \int k(\mathbf{z}_i|\boldsymbol{\theta})H_0(d\boldsymbol{\theta})$, $q_{il} = k(\mathbf{z}_i|\boldsymbol{\theta}_l)$ for $l \geq 1$ and $p(\boldsymbol{\theta}_i|\mathbf{z}_i, H_0)$ is the posterior distribution for $\boldsymbol{\theta}_i$ based on the prior H_0 and a single observation $\mathbf{z}_i$. MacEachern (1994) points out that mixing can be slow in this setting, and proposes to add an additional step to the Gibbs sampler that resamples the group parameters conditional on the indicators. Taking this idea one step forward, Bush and MacEachern (1996) note that, in the conjugate case, the group parameters can be easily integrated out,

yielding a more efficient sampler. Finally, MacEachern and Müller (1998) propose an algorithm that can be used in the nonconjugate case. Neal (2000) provides an excellent review of marginal methods.

Blocked samplers are a more recent idea and are based on approximations to the Dirichlet process by finite mixture models. They are straightforward to code, tend to have better mixing properties than marginal samplers and, unlike them, directly produce (approximate) draws from the posterior distribution $H^n(d\boldsymbol{\theta}|\mathbf{z}^n)$. Their main drawback is that the samples only approximately follow the desired distribution. As an example, consider the truncation sampler of Ishwaran and James (2001), which starts with the finite stick breaking prior

$$P^K(\cdot) = \sum_{k=1}^{K} w_k \delta_{\boldsymbol{\theta}_k}(\cdot) \qquad \boldsymbol{\theta}_k \sim H$$

$$w_k = z_k \prod_{l=1}^{k-1}(1 - z_l) \qquad z_k \sim \begin{cases} \mathsf{beta}(a_k, b_k) & \text{if } k < K \\ \delta_1 & \text{if } k = K \end{cases}$$

After proving that P^K converges in distribution to a Dirichlet process when $K \to \infty$, the authors are able to construct a simple Gibbs sampler that exploits conjugacy between the generalized Dirichlet distribution and the multinomial distribution. A related approach is the retrospective sampler (Roberts and Papaspiliopoulos, 2007), who also use the stick breaking representation of the Dirichlet process to generate a sampler that avoids truncations but shares some of the advantages of the blocked sampler.

Finally, different authors (Jain and Neal, 2000; Dahl, 2003) have proposed samplers that use Reversible Jump steps (Green, 1995). These samplers can be very efficient in avoiding local modes on the posterior distribution, improving the mixing of the chain under some circumstances. However, they are typically much harder to implement.

This short review of methods is in no way exhaustive. Different authors have successfully implemented algorithms that cannot be readily categorized in any of the previous groups. For example, MacEachern *et al.* (1999) use sequential importance samplers, while Blei and Jordan (2006) adapt variational methods for the Dirichlet process. This last approach is particularly useful when large data sets are involved.

1.3 Models for collections of distributions based on the Dirichlet Process

Although most of the usual applications of Dirichlet Process models focus on problems with exchangeable samples from one unknown distribution, there is growing interest in extending the Dirichlet Process to accommodate multiple dependent distributions.

Most approaches in the literature accomplish dependence between the distributions either by introducing dependence in the elements of the stick-breaking representation of the distribution or by forming convex combinations of independent processes.

The dependent Dirichlet process (DDP) (MacEachern, 1999, 2000) induces dependence in a collection of distributions by replacing the elements of the stick-breaking representation (Sethuraman, 1994) with stochastic processes. It has been employed by DeIorio *et al.* (2004) to create ANOVA-like models for densities, and by Gelfand *et al.* (2005) to generate spatial processes that allow for non-normality and non-stationarity. This last class of models is extended in Duan *et al.* (2007) to create generalized spatial Dirichlet processes (GSDP) that allow different surface selection at different locations, among others.

Along similar lines, the hierarchical Dirichlet process (HDP) (Teh *et al.*, 2006) is another approach to introduce dependence. In this setting, multiple group-specific distributions are assumed to be drawn from a common Dirichlet Process whose base-

line measure is in turn a draw from another Dirichlet process. This allows the different distributions to share the same set of atoms but have distinct sets of weights. More recently, Griffin and Steel (2006b) proposed an order-dependent Dirichlet Process (πDDP), where the correspondence between atoms and weights is allowed to vary with the covariates. Also, Dunson and Park (2007) propose a kernel stick breaking that allows covariate dependent weights and fixed atoms.

An alternative approach to the DDP is to introduce dependence through linear combinations of realizations of independent Dirichlet processes. For example, Müller *et al.* (2004), motivated by a similar problem to Teh *et al.* (2006), define the distribution of each group as the mixture of two independent samples from a DP process: one component that is shared by all groups and one that is idiosyncratic. Dunson (2006) extended this idea to a time setting, and Dunson *et al.* (2007b) propose a model for density regression using a kernel-weighted mixture of Dirichlet Processes defined at each value of the covariate.

In what follows, we concentrate on the dependent Dirichlet process due to its theoretical appeal and computational simplicity.

1.3.1 Dependent Dirichlet processes

Given an index set D, let $\{\boldsymbol{\theta}(t) : t \in D\}$ and $\{z(t) : t \in D\}$ be stochastic processes over D such that $z(t) \sim \mathsf{beta}(1, \alpha(t)) \; \forall \; t \in D$ and define

$$H_t(\cdot) = \sum_{l=1}^{\infty} w_l^*(t) \delta_{\boldsymbol{\theta}_l^*(t)}(\cdot), \tag{1.4}$$

where $\{\boldsymbol{\theta}_l^*(t)\}_{l=1}^{\infty}$ and $\{z_l^*(t)\}_{l=1}^{\infty}$ are mutually independent collections of independent realizations of the stochastic processes $\{\boldsymbol{\theta}(t) : t \in D\}$ 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 (DDP) (MacEachern, 2000). Note that, for any fixed t, H_t follows a Dirichlet process.

DDP models are dense on a large class of distributions. Indeed, under mild conditions, the DDP assigns positive probability to every ϵ-ball centered on a finite collection of distributions that are absolutely continuous to the baseline measures corresponding to the same locations of the index space D (MacEachern, 2000).

One of the most popular variates of the DDP is the "single-p" model, where the weights are assumed to be constant over D while the atoms are allowed to vary. Models of this form can be rewritten as regular DP models with atoms arising from a stochastic process. Therefore, standard Gibbs sampling algorithms can be used to perform inferences for the "single-p" DDP models. The main drawback of this approach is its inability to produce a collection of independent distributions.

The hierarchical Dirichlet process (HDP) (Teh *et al.*, 2006) can also be recast as a DDP model. The HDP places a prior on a collection of *exchangeable* distributions $\{G_1, \ldots, G_J\}$. Conditional on a probability measure G_0, the distributions in the collection are assumed to be iid samples from a regular Dirichlet process centered around G_0. In order to induce dependence, G_0 is in turn given another Dirichlet process prior. In summary,

$$G_i | G_0 \sim \mathsf{DP}(\alpha G_0)$$

$$G_0 \sim \mathsf{DP}(\beta H)$$

Since G_0 is, by construction, almost surely discrete, the distributions G_i share the same set of random atoms (corresponding to those of G_0), but assign strictly different (although dependent) weights to each one of them. As is to be expected, H corresponds to the common expected value for each of the distributions in the collection, and α and β control the variance around H and the dependence between distributions. Computation for the HDP is performed using a generalized Pólya urn

scheme.

Another special case of the DDP is the order-dependent DP (πDP), proposed by Griffin and Steel (2006b). For the πDP, both atoms and weights are kept constant, but *the assignment* of weights to atoms (represented by a permutation on the indices) is allowed to vary on D. This construction has numerous theoretical advantages. By carefully constructing the stochastic processes driving the orderings, the πDDP allows for a set of independent distributions as a limiting case and for the posterior distribution at a new point $\mathbf{x}^* \in D$ to converge towards the baseline measure as $\mathbf{x}^*$ moves away from the observed points. Both of these features are not possible with the "single p" DDP. However, constructing the underlying stochastic processes necessary to define the process can be a complex task. Also, the algorithm used to fit this model is fairly complicated.

1.4 Other Bayesian nonparametric models

1.4.1 Pólya Trees

Pólya trees (PT) (Lavine, 1992; Mauldin *et al.*, 1992; Lavine, 1994; Paddock *et al.*, 2003) define random distributions on a space $(\boldsymbol{\Theta}, \mathcal{B})$ by first generating a sequence of binary partitions of the space and then assigning probability masses to each element of each partition in a hierarchically consistent way. The Dirichlet processes is obtained as a special case of a Pólya tree. We begin the description of Pólya Trees with a definition:

Definition 1. *A separating binary tree partition is a sequence of partitions* $\Pi = \{\pi_t : t = 0, 1, \ldots\}$ *such that* $\bigcup_{k=0}^{\infty} \pi_k$ *generates the measurable sets on* $\boldsymbol{\Theta}$ *and every* $B \in \pi_{k+1}$ *is obtained by splitting some* $B^* \in \pi_k$ *in two pieces.*

Let $D = \{0, 1\}$, $D_0 = \emptyset$, D_k be the k-fold product $D \times D \times \cdots \times D$ and $D^* =$

$\bigcup_{k=0}^{\infty} D_k$. If $\mathbf{\Theta}$ is a separable, measurable space and Π is a separating binary tree of partitions of $\mathbf{\Theta}$, the random probability measure H on $\mathbf{\Theta}$ has a Pólya tree with parameters $(\Pi, \mathcal{A})$, denoted $H \sim \mathsf{PT}(\Pi, \mathcal{A})$, if there exist a set of parameter $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in D^*\}$ and a collection of random parameters $\mathcal{Y} = \{Y_\epsilon : \epsilon \in D^*\}$ such that

1. The random variables in $\mathcal{Y}$ are independent.

2. For every $\epsilon \in D^*$, $Y_\epsilon \sim \mathsf{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$.

3. For every $k = 1, 2, \ldots$ and every $\epsilon_{1:k}$ we have

$$H(B_{\epsilon_{1:k}}) = \prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_{1:j-1}} \prod_{j=1;\epsilon_j=1}^{m} (1 - Y_{\epsilon_{1:j-1}})$$

Regular Pólya trees use a fixed partition (if $\mathbf{\Theta} = (0, 1]$, this is typically the canonical dyadic partition), while randomized Pólya trees assume the partition to be random and put a prior distribution on it. Unlike the Dirichlet process, Pólya tree priors can generate continuous distributions if the set $\mathcal{A}$ satisfies $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$. They can be centered on a distribution H_0, for example, by choosing $\Pi = \{\{(H_0(l2^{-k}, (l + 1)2^{-k}); l = 0, \ldots, 2^k - 1\}; k = 0, 1, \ldots\}$.

Pólya trees are conjugate priors, which can be exploited to design an efficient algorithm to sample from the posterior distribution. However, since the definition of Pólya trees involve a countable number of partitions, such computational approaches require the truncation of the process (typically, around 7 levels are used in practice). Implementations of Pólya tree priors on multidimensional spaces exist (Paddock, 1999), but it becomes harder to implement for more than two or three dimensions.

1.4.2 Neutral to the right processes

Neutral to the right processes (NTTR) were introduced by Doksum (1974). In Bayesian nonparametrics, they are used to model distribution functions with support

on $\mathbb{R}^+$ (typically, survival distribution functions). A random distribution H with support on $\mathbb{R}^+$ is said to follow a NTTR process if for every k and $0 < t_1 < \ldots < t_k$ there exist independent random variables $V_1, \ldots, V_k$ such that

$$(1 - H(t_1), 1 - H(t_2), \ldots, 1 - H(t_k)) \stackrel{d}{=} \left(V_1, V_1 V_2, \ldots, \prod_{j=1}^{k} V_k \right)$$

In this case, the distribution function $H(x)$ can be written as $H(x) = \exp\{S(x)\}$ where $S(\cdot)$ is an independent increments process. One particular example of a NTTR is the beta-Stacey process (Walker and Mulliere, 1997). NTTR are also conjugate priors on the space of distribution, in the sense that the posterior is another NTTR. This is true even under right-censored data, which reinforces its appeal for modeling survival data.

1.4.3 Lévy processes

A stochastic process $X = \{X(t)\}$ is said to be a Lévy process (LP) on a probability space $(\mathbf{\Omega}, \mathcal{F}, P)$ iif

1. X has independent increments.
2. $X(0) = 0$ almost surely.
3. X is stochastically continuous, i.e., for any $s, t \geq 0$, $X(t - s) - X(s) \stackrel{P}{\to} 0$ as $t \to 0$.
4. X is time homogeneous, i.e., for $s, t \geq 0$, the law of $X(t - s) - X(s)$ does not depend on s.
5. X is right continuous with left limits almost surely.

Brownian motion, compound Poisson processes and gamma process are some examples of Lévy process.

There is a well-known relationship between infinitely divisible distributions and Lévy processes. Indeed, there is a one-to-one relationship between the collection of all infinitely divisible distributions and the collection of all Lévy processes.

The characteristic function of any infinitely divisible measure has an integral representation in terms of a positive σ-finite measure $\nu(\cdot)$ on $\mathbb{R}$ 0, which is called the Lévy measure. This result is known as the Lévy-Khinchine theorem (Jacod and Shiryaev, 1987). As a consequence of this theorem, any LP can be represented as the sum of a Brownian motion with a drift and a pure jump process.

Strictly increasing Lévy processes (called subordinators) have been recently used in a Bayesian nonparametric context for tasks including time series analysis, nonparametric regression, spatio-temporal models, density estimation and solving integral equations (Wolpert and Ickstadt, 1998a; Wolpert *et al.*, 2003; Tu, 2006). A strictly increasing LP can be represented as a countable sum of point masses of the form

$$X(ds) = \sum_i v_i \delta_{\theta_i}(ds)$$

If the Lévy measure ν is finite, then the number of point masses is finite, otherwise it is countably infinite. In the second case, the measure is typically truncated for implementation purposes.

Computational tools employing MCMC schemes have been developed to fit models based on LP priors. In (Wolpert and Ickstadt, 1998b), the authors develop the Inverse Lévy Measure (ILM) algorithm, which employs data augmentation to obtain a conjugate model. This method generates a sample from the entire process using a Gibbs sampling scheme. On the other hand, Wolpert *et al.* (2003) develop a Reversible Jump MCMC scheme that directly samples the location and weights of the process. This second algorithm can be used in non-conjugate settings, but is generally slower than the ILM.

Lévy processes are an interesting generalization of the Dirichlet process and have been shown to be a powerful tool. However, LP priors are priors on measures (not necessarily probability measures), making them less attractive in density estimation problems. Besides choosing the truncation level is typically a complicated task that is usually problem specific and might dramatically affect the resulting estimates.

1.4.4 Other alternatives

The previous review just scratches the surface of the literature on Bayesian nonparametric methods. For example, the Dirichlet process, Pólya tress and NTTR are all examples of Tail Free processes. Gaussian processes (Rasmussen and Williams, 2006) are another class of nonparametric priors that has also been proposed to model the log density. Finally, generalized Pólya urn schemes and generalized stick breaking processes (Hjort, 2000; Ongaro and Cattaneo, 2004) constitute another set of alternatives.

In spite of this vast literature on nonparametric process, there is surprisingly little work on generating models for collections of distributions that does not exploit the Dirichlet process.

1.5 Posterior consistency

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 outlines some well-known results on consistency that will be relevant later.

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

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

induces the weak convergence topology. Weak ϵ-neighborhood of $g^0 \in \mathfrak{m}(\mathbb{R}^p)$ are sets of the form,

$$U_\epsilon^w(g^0) = \left\{ g \in \mathfrak{m}(\mathbb{R}^p) : \left| \int \psi_i(\mathbf{z}) g(\mathbf{z}) d\mathbf{z} - \int \psi_i(\mathbf{z}) g^0(\mathbf{z}) d\mathbf{z} \right| < \epsilon, \ i = 1, \ldots, k \right\},$$

for $\psi_i \in C_b(\mathbb{R}^p)$, 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 (1.2) is dense (in the L^1 sense) on $\mathfrak{m}(\mathbb{R}^p)$ (Ghosh and Ramamoorthi, 2003). Letting $\mathbf{z}_1, \ldots, \mathbf{z}_n \sim g$ 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 (1.3) 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 $\mathbf{z}_1, \mathbf{z}_2, \ldots$, $\int \psi(g) \mu_n(dg) \to \int \psi(g) \delta_{g^0}(dg)$ for every $\psi \in C_b(\mathbb{R}^p)$, which happens iff $\mu_n(U_\epsilon^w(g^0)) \to 1$, 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.

Sufficient conditions to ensure weak consistency were given by Schwartz (1965). As noted by Diaconis and Freedman (1986a,b), when the parameter space is infinite-dimensional (as in nonparametric models) it is not enough to have g^0 in the weak

support of μ, but g^0 needs to be in its Kullback-Leibler support, defined as the set

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

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 used later

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 a prior* μ *on* $\mathfrak{m}(\mathbb{R}^p)$*. Then* g^0 *is in the Kullback-Leibler support of* μ *defined by the DPM model in (1.2), and therefore the corresponding posterior distribution* μ_n *is weakly consistent at* g^0*.*

Chapter 2

The Nested Dirichlet Process

In multicenter studies, subjects in different centers may have different outcome distributions. This chapter is motivated by the problem of nonparametric modeling of these distributions, borrowing information across centers while also allowing centers to be clustered. This type of model can be used to generate relevant hypotheses that can be further explored in subsequent studies.

As a motivating example, consider assessing quality of care across hospitals in the US. The outcomes of patients in each institution define a hospital-specific distribution, which can be non-normal, presenting skewness, multi-modality and/or heavy tails. In this setting, it is of interest to cluster centers according to the full distribution of patients outcomes, and to identify outlying centers. On the other hand, it is also interesting to simultaneously cluster patients within the centers, and to do so by borrowing information across centers that present clusters with similar characteristics. This task is different from clustering patients within and across centers, which could be accomplished using the approaches discussed in Teh *et al.* (2006) and Müller *et al.* (2004).

In order to build our model, we start with a stick-breaking representation of the Dirichlet process (DP) and replace the random atoms with random probability

measures drawn from a DP. This results in a nested Dirichlet process (nDP) prior, which can be placed on the collection of distributions for the different centers, with centers drawn from the same DP component automatically clustered together.

This chapter is organized as follows. We start in section 2 with the motivation, definition and properties of the nested Dirichlet process (nDP). In section 3 we discuss truncations of the nDP and their application in deriving efficient computational schemes is discussed in section 4. Sections 6 and 7 present examples that illustrate the advantages of our methodology. Finally, we close in section 8 with a brief discussion.

2.1 The Nested Dirichlet Process

2.1.1 Definition and basic properties

Suppose y_{ij}, for $i = 1, \ldots, n_j$ are observations for different subjects within center j, for $j = 1, \ldots, J$. For example, $\mathbf{y}_j = (y_{1j}, \ldots, y_{n_j j})'$ may represent patient outcomes within the jth hospital or hospital-level outcomes within the jth state. Although covariates, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ are typically available, we initially assume that subjects are exchangeable within centers, with $y_{ij} \overset{iid}{\sim} F_j$, for $j = 1, \ldots, J$.

In analyzing multi-center data, there are a number of customary strategies, with the most common being (1) pool the data from the different centers; (2) analyze the data from the different centers separately; and (3) fit a parametric hierarchical model to borrow information. The first approach is too restrictive, as subjects in different centers may have different distributions, while the second approach is inefficient. The third approach parameterizes F_j in terms of the finite-dimensional parameter $\boldsymbol{\theta}_j$, and then borrows information by assuming $\boldsymbol{\theta}_j \overset{iid}{\sim} F_0$, with F_0 a known distribution (most commonly normal), possibly having unknown parameters (mean, variance). One can potentially cluster centers having similar random effects, $\boldsymbol{\theta}_j$, though clustering may be sensitive to F_0 (Verbeke and Lesaffre, 1996). Assuming that F_0 has an arbitrary

discrete distribution having k mass points provides more flexible clustering, but the model is still dependent on the choice of k and the specific parametric form for F_j.

Furthermore, clustering based on the random effects has the disadvantage of only borrowing information about aspects of the distribution captured by the parametric model. For example, clustering centers by mean patient outcomes ignores differences in the tails of the distributions. Our motivation is to borrow information and cluster across distributions $\{F_j, j = 1, \ldots, J\}$ nonparametrically to enhance flexibility, and we use a Dirichlet type of specification to enable clustering of random distributions.

In what follows, a collection of distributions $\{F_1, \ldots, F_J\}$ is said to follow a *Nested Dirichlet Processes Mixture* if

$$F_j(\cdot|\boldsymbol{\phi}) = \int_{\boldsymbol{\Theta}} p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi}) G_j(d\boldsymbol{\theta}) \tag{2.1}$$

$$G_j(\cdot) \sim \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*(\cdot)} \tag{2.2}$$

$$G_k^*(\cdot) = \sum_{l=1}^{\infty} w_{lk}^* \delta_{\boldsymbol{\theta}_{lk}^*}(\cdot) \tag{2.3}$$

with $\boldsymbol{\theta}_{lk}^* \sim H$, H is a probability measure on $(\boldsymbol{\Theta}, \mathcal{B})$, $w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1}(1 - u_{sk}^*)$, $\pi_k^* = v_k^* \prod_{s=1}^{k-1}(1 - v_s^*)$, $v_k^* \sim \mathsf{beta}(1, \alpha)$ and $u_{lk}^* \sim \mathsf{beta}(1, \beta)$. In expression (2.1), $p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi})$ is a distribution parameterized by the finite dimensional vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, whose specific choice depends on the application at hand. For example, in the case of a univariate response, if the collection $\{F_1, \cdots, F_J\}$ is assumed exchangeable, an attractive choice would be $\boldsymbol{\theta} = (\mu, \sigma)$ and $p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathsf{N}(\cdot|\mu, \sigma^2)$, which yields a class that is dense on the space of absolutely continuous distributions (Lo, 1984). On the other hand, if a vector $\mathbf{x}$ of covariates is available, we could opt for a random effects model where $\boldsymbol{\theta} = \mu$, $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \sigma^2)$ and $p(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathsf{N}(\cdot|\mu + \mathbf{x}'\boldsymbol{\gamma}, \sigma^2)$, similar in spirit to Mukhopadhyay and Gelfand (1997) and Kleinman and Ibrahim (1998). Extensions to multivariate or discrete outcomes are immediate using the standard

Bayesian machinery.

The collection $\{G_1, \ldots, G_J\}$, used as the mixing distribution, is said to follow a *Nested Dirichlet Process* with parameters α, β and H, and is denoted $\mathsf{nDP}(\alpha, \beta, H)$. In a more concise notation, the model for our clustering problem can be rewritten as

$$y_{ij} \sim p(y_{ij}|\boldsymbol{\theta}_{ij}) \qquad \boldsymbol{\theta}_{ij} \sim G_j \qquad \{G_1, \ldots, G_J\} \sim \mathsf{nDP}(\alpha, \beta, H)$$

Since *apriori* $\mathbb{P}(G_j = G_{j'}) = \frac{1}{1+\alpha} > 0$, the model naturally induces clustering in the space of distributions. Also, for any measurable set $A \in \mathcal{B}$

$$\mathbb{E}(G_j(A)) = H(A) \qquad \text{and} \qquad \mathbb{V}(G_j(A)) = \frac{H(A)(1 - H(A))}{\beta + 1}$$

Since our goal is to create a collection of *dependent* distributions, it is natural to also consider the correlation induced by the model between the probabilities assigned by two members of the collection to a given set $A \in \mathcal{B}$, i.e., $\mathbb{C}\mathrm{or}(G_j(A), G_{j'}(A))$. It is shown in appendix B that for the nDP,

$$\mathbb{C}\mathrm{or}(G_j(A), G_{j'}(A)) = \frac{1}{1+\alpha} = \mathbb{P}(G_j = G_{j'})$$

Note that this result, which provides a natural interpretation for the additional parameter in the nDP, is independent of the set A. Therefore, from now on we will refer to it as the prior correlation between distributions, denoted $\mathbb{C}\mathrm{or}(G_j, G_{j'})$. The correlation between draws from the process can also be calculated (see again appendix B), yielding

$$\mathbb{C}\mathrm{or}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) = \begin{cases} \frac{1}{(1+\beta)} & j = j' \\ \frac{1}{(1+\alpha)(1+\beta)} & j \neq j' \end{cases}$$

This shows that the a priori correlation between observations arising from the same center is larger than the correlation between observations from different centers, which is an appealing feature. Given a specific form for $p(\cdot|\boldsymbol{\theta}_j, \boldsymbol{\phi})$, the previous

expression allows us to calculate the prior correlation that the model induces on the observations.

Note that as $\alpha \to \infty$, each distribution in the collection is assigned to a distinct atom of the stick-breaking construction. Therefore, the distributions become a priori independent given the baseline measure H, which agrees with the fact that $\lim_{\alpha\to\infty} \mathbb{C}\text{or}(G_j, G_{j'}) = 0$. On the other hand, as $\alpha \to 0$ the a priori probability of assigning all the distributions to the same atom G^* goes to 1, and thus the correlation goes to 1. Hence, approaches (1) and (2) for the analysis of multiple centers described above are limiting cases of the nDP. Moreover, since $F_j(\cdot) \to p(\cdot|\boldsymbol{\theta}_j^*, \boldsymbol{\phi})$ as $\beta \to 0$, the nDP also encompasses the natural parametric-based clustering (option (3) above) as a limiting case.

Since every G_k^* is almost surely discrete, the model simultaneously enables clustering of observations within each center along with clustering the distributions themselves. For example, we can simultaneously group hospitals having the same distribution of patient outcomes, while also identifying groups of patients within a hospital having the same outcome distribution. Indeed, centers j and j' are clustered together if $G_j = G_{j'} = G_k^*$ for some k, while patients i and i', respectively from hospitals j and j', are clustered together if and only if $G_j = G_{j'} = G_k^*$ and $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j'} = \boldsymbol{\theta}_{lk}^*$ for some l.

2.1.2 Alternative characterizations of the nDP

Just as the Dirichlet Process is a distribution on distributions, the nDP can be characterized as a *distribution on the space of distributions on distributions.* Recall the original definition of the Dirichlet Process (Ferguson, 1973, 1974) stated in section 1.1. The choice $\boldsymbol{\Theta} \subset \mathbb{R}^n$ for the base space of the Dirichlet Process is merely a practical one, and the results mentioned above extend in general to any complete and

separable metric space $\mathbf{\Theta}$. In particular, since the space of probability distributions is complete and separable under the weak topology metric (see Ghosh and Ramamoorthi (2003), page 13), we could have started by taking $(\mathbf{P}, \mathcal{C}, Q)$ (defined before) as our base space and defining a new distributional space $(\mathbf{Q}, \mathcal{D}, S)$ such that $\mathcal{D}$ is the smallest σ-algebra generated by all weakly open sets in $\mathbf{Q}$ and $Q \in \mathbf{Q}$. In this setting, $\mathbf{Q}$ is the space of *distributions on probability distributions* on $(\mathbf{\Theta}, \mathcal{B})$.

By requiring S to be such that $(Q(C_1), \ldots, Q(C_k)) \sim \mathsf{Dir}(\alpha\nu(C_1), \ldots, \alpha\nu(C_k))$ for any partition $(C_1, \ldots, C_k)$ of $\mathbf{P}$ generated under the weak topology and some α and suitable ν, we have defined a new Dirichlet Process $S \sim \mathsf{DP}(\alpha\nu)$, this time on an abstract space, that satisfies the usual properties. The nested Dirichlet process is a special case of this formulation in which ν is taken to be a regular $\mathsf{DP}(\beta H)$. Therefore, the nDP is an example of a DP where the baseline measure is a stochastic process generating probability distributions. An alternative notation for the nDP corresponds to $G_j \overset{iid}{\sim} Q$ with $Q \sim \mathsf{DP}(\alpha\mathsf{DP}(\beta H))$.

The nDP can also be characterized as a dependent Dirichlet process (MacEachern, 2000) where the stochastic process generating the elements of the stick-breaking representation corresponds to a Pólya urn. Indeed, we can write

$$G_j = \sum_{l=1}^{\infty} w'_{lj} \delta_{\boldsymbol{\theta}'_{lj}}$$

where

$$\left((\mathbf{w}'_j, \boldsymbol{\theta}'_j) | (\mathbf{w}'_1, \boldsymbol{\theta}'_1), \ldots, (\mathbf{w}'_{j-1}, \boldsymbol{\theta}'_{j-1})\right) \sim \sum_{k<j} \frac{1}{1+\alpha} \delta_{(\mathbf{w}'_k, \boldsymbol{\theta}'_k)} + \frac{\alpha}{1+\alpha} H^{\infty} \times S_{\beta}^{\infty}$$

where $H^{\infty} = H \times H \times \cdots$ and S_{β}^{∞} is the prior distribution on the (infinite) collection of weights induced by the stick breaking construction with parameter β.

Finally, the NDP can be viewed as a way to simultaneously define a prior on a random partition of the collection $\{G_1, \ldots, G_J\}$ (in the style of Quintana and Iglesias

(2003)) and each of the resulting unique distributions.

2.1.3 Comparing the nDP with other nonparametric models

It is important to note that, although both approaches generalize the DP to allow hierarchical data structures, the dependence induced by the NDP is fundamentally different from that induced by the HDP. Figure 2.1 illustrates these differences. In the HDP, *one draw* from a Dirichlet process is used as the baseline measure G_0 of the process generating the members of the collection. As discussed in Teh *et al.* (2006), this implies that $\{G_1, \ldots, G_J\}$ share the same atoms (the atoms of G_0) but assign them different weights. Therefore $\mathbb{P}(G_j = G_{j'}) = 0$ under the HDP and clustering happens only at the level of the observations.

On the other hand, the baseline measure in the nDP is not a draw from a Dirichlet Process, but the whole process itself. In particular, we have already shown that such a construction implies that two given distributions either share both atoms and weights (making them exactly equal, as G_1 and G_3 in the right panel of figure 2.1), or do not share any of the features. This induces clustering on both observations and distributions.

The nDP is also different from the linear combination models in Müller *et al.* (2004), which allow for a limited form of clustering across distributions. In Müller *et al.* (2004), an unknown distribution G_i is represented as a linear combination

$$G_i = \epsilon_i H_0 + (1 - \epsilon_i) H_i$$

where each H_i is an independent draw from a regular Dirichlet Process. H_0 is called the common component, while the H_i's are called the idiosyncratic components. Note that, for two distributions G_i and $G_{i'}$ to be equal in this model, they must correspond to the common component in the mixture, i.e., $\epsilon_i = \epsilon_{i'} = 0$ implying that $G_i = G_{i'} = H_0$. Thus, there is at most one cluster having more than one member.

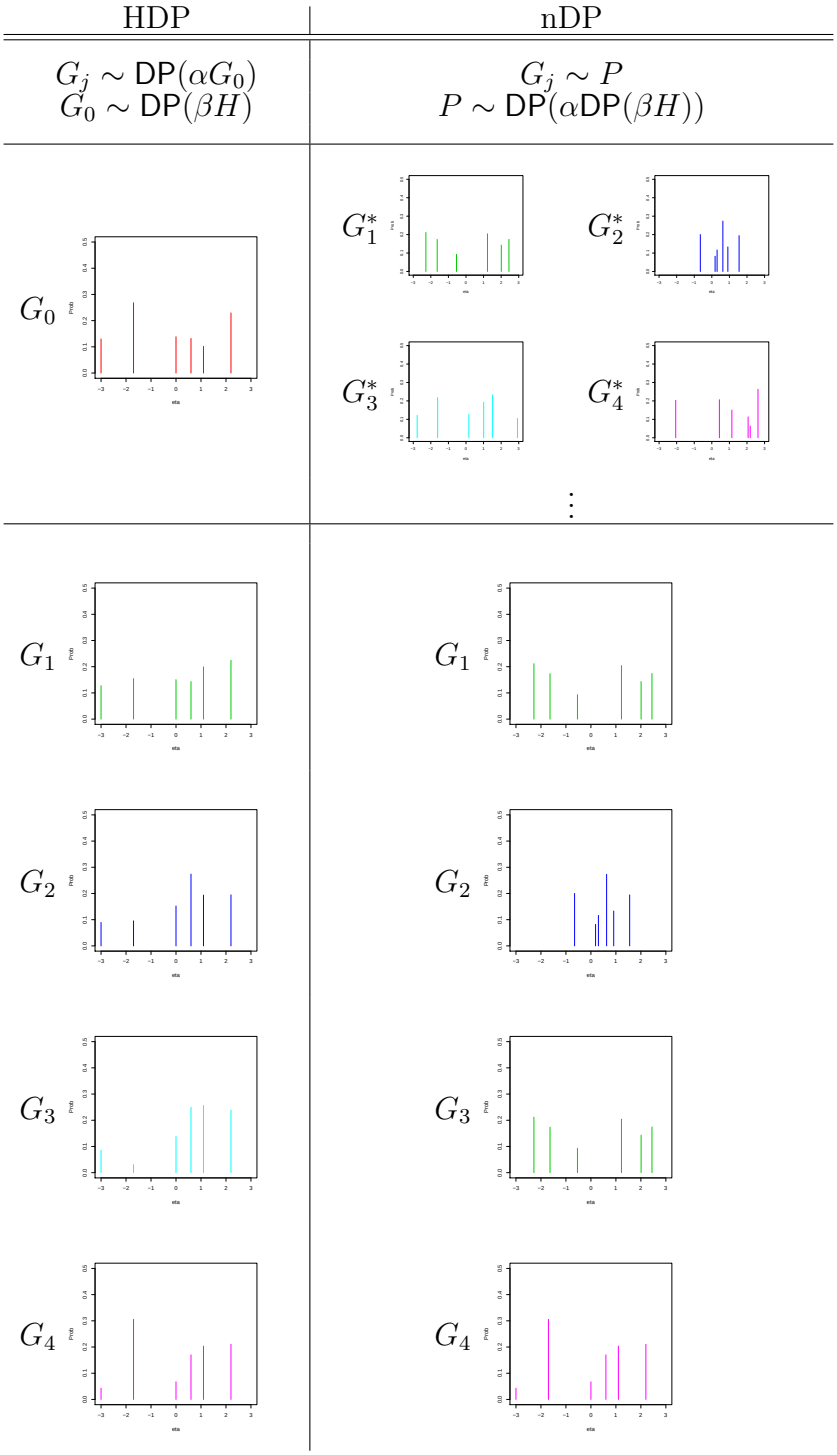


Figure 2.1: Comparing the nDP and the HDP. For the HDP, the distributions $\{G_1, \ldots, G_J\}$ share the same atoms but assign them different weights. For the nDP the different distributions either have the same atoms with the same weights, of completely different atoms and weights.

2.2 Truncations

In this section, we consider finite-mixture versions of the nDP. Finite mixtures are usually simpler to understand, and can help provide insights into the more complicated, infinite dimensional models. Additionally, they provide useful approximations that can be used for computation.

Definition 2. *An LK truncation of an* $\mathsf{nDP}(\alpha, \beta, H)$ *is defined by the finite-mixture model*

$$G_j^K(\cdot) \sim \sum_{k=1}^{K} \pi_k^* \delta_{G_k^{L*}(\cdot)}$$

$$G_k^{L*}(\cdot) = \sum_{l=1}^{L} w_{lk}^* \delta_{\boldsymbol{\theta}_{lk}^*}(\cdot) \quad \pi_k^* = v_k^* \prod_{s=1}^{l-1}(1 - v_s^*) \qquad v_K^* = 1$$

$$v_k^* \sim \mathsf{beta}(1, \alpha)\ k = 1, \ldots, K-1$$

$$\boldsymbol{\theta}_{lk}^* \sim H \qquad w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1}(1 - u_{sk}^*) \quad u_{Lk}^* = 1$$

$$u_{lk}^* \sim \mathsf{beta}(1, \beta)\ l = 1, \ldots, L-1$$

We refer to this model as a bottom-level truncation or $\mathsf{nDP}^{L\infty}(\alpha, \beta, H)$ *if* $K = \infty$ *and* $L < \infty$*, whereas if* $K < \infty$ *and* $L = \infty$ *we refer to it as a top-level truncation or* $\mathsf{nDP}^{\infty K}(\alpha, \beta, H)$*. Finally, if both* L *and* K *are finite we have a two-level truncation or* $\mathsf{nDP}^{LK}(\alpha, \beta, H)$*.*

The total variation distance between an nDP and its truncation approximations can be shown to have decreasing bounds as $L, K \to \infty$. For simplicity, we consider the case when $n_j = n\ \forall\ j$.

Theorem 2. *Assume that samples of* n *observations have been collected for each of*

J distributions and are contained in vector $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_J')$. *Also, let*

$$P^{\infty\infty}(\boldsymbol{\theta}) = \int\int P(\boldsymbol{\theta}|G_j)P^{\infty}(dG_j|Q)P^{\infty}(dQ)$$

$$P^{LK}(\boldsymbol{\theta}) = \int\int P(\boldsymbol{\theta}|G_j)P^{L}(dG_j|Q)P^{K}(dQ)$$

be, respectively, the prior distribution of the model parameters under the nDP model and its corresponding LK truncation after integrating out the random distributions, and $P^{\infty\infty}(\mathbf{y})$ *and* $P^{LK}(\mathbf{y})$ *be the prior predictive distribution of the observations derived from these priors. Then*

$$\int \left|P^{LK}(\mathbf{y}) - P^{\infty\infty}(\mathbf{y})\right| d\mathbf{y} \leq \int \left|P^{LK}(d\boldsymbol{\theta}) - P^{\infty\infty}(d\boldsymbol{\theta})\right| \leq \epsilon^{LK}(\alpha, \beta)$$

where

$$\epsilon^{LK}(\alpha,\beta) = \begin{cases} 4\left(1 - \left[1 - \left(\frac{\alpha}{1+\alpha}\right)^{K-1}\right]^{J}\right) & \text{if } L = \infty, K < \infty \\ 4\left(1 - \left[1 - \left(\frac{\beta}{\beta+1}\right)^{L-1}\right]^{nJ}\right) & \text{if } L < \infty, K = \infty \\ 4\left(1 - \left[1 - \left(\frac{\alpha}{1+\alpha}\right)^{K-1}\right]^{J}\left[1 - \left(\frac{\beta}{\beta+1}\right)^{L-1}\right]^{nJ}\right) & \text{if } L < \infty, K < \infty \end{cases}$$

The proof of this theorem is presented in appendix C. Note that the bounds approach zero in the limit, so the truncation approximations and its predictive distribution converge in total variation (and therefore in distribution) to the nDP. Even more, the bounds are strictly decreasing in both L and K. As a consequence of this observation we have the following corollary.

Corollary 1. *The posterior distribution under a LK truncation and the corresponding nDP converge in distribution as both* $L, K \to \infty$.

The proof is presented in appendix E. It is straightforward to extend the previous results and show that $\lim_{L\to\infty} \mathsf{nDP}^{LK} = \mathsf{nDP}^{\infty K}$ and $\lim_{K\to\infty} \mathsf{nDP}^{LK} = \mathsf{nDP}^{L\infty}$ in distribution.

In order to better understand the influence of the truncation levels on the accuracy of the approximation we show in Figure 2.2 the error bounds for a $\mathsf{nDP}(3, 3, H)$ in various sample size settings. The value $\alpha = \beta = 3$ in this simulation, which will typically lead to a relatively large number of components in the mixtures, was chosen as a worst case scenario since the bounds are strictly decreasing in both α and β.

The first three examples have a total of 5,000 observations, which have been split in different ways. Note that, as the number of groups J increases, K needs to be increased to maintain accuracy. The fourth example has the same number of observations per group as the first, but double the number of groups. In every case, increasing K over 35 seems to have little effect on the error bound. These results suggest that for moderately large sample sizes ($n \leq 500$ and $J \leq 50$), and typical values of the concentration parameters α and β, a choice of $K = 35$ and $L = 55$ seems to provide an adequate approximation.

2.3 Posterior computation

Broadly speaking, there are three strategies for computation in standard DP models: (1) Employ the Pólya urn scheme to marginalize out the unknown infinite-dimensional distribution(s) (MacEachern, 1994; Escobar and West, 1995; MacEachern and Müller, 1998), (2) Employ a truncation approximation to the stick-breaking representation of the process and then resort to methods for computation in finite mixture models (Ishwaran and Zarepour, 2002; Ishwaran and James, 2001) and (3) Use reversible-jump MCMC (RJMCMC) algorithms for finite mixtures with an unknown number of components (Dahl, 2003; Green and Richardson, 2001; Jain and Neal, 2000). In this section, we explore the use of these strategies to construct efficient algorithms for inference in the nDP setting. In the sequel, let $\zeta_j = k$ and $\xi_{ij} = l$ iff $G_j = G_k^*$ and $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{l\zeta_j}^*$.

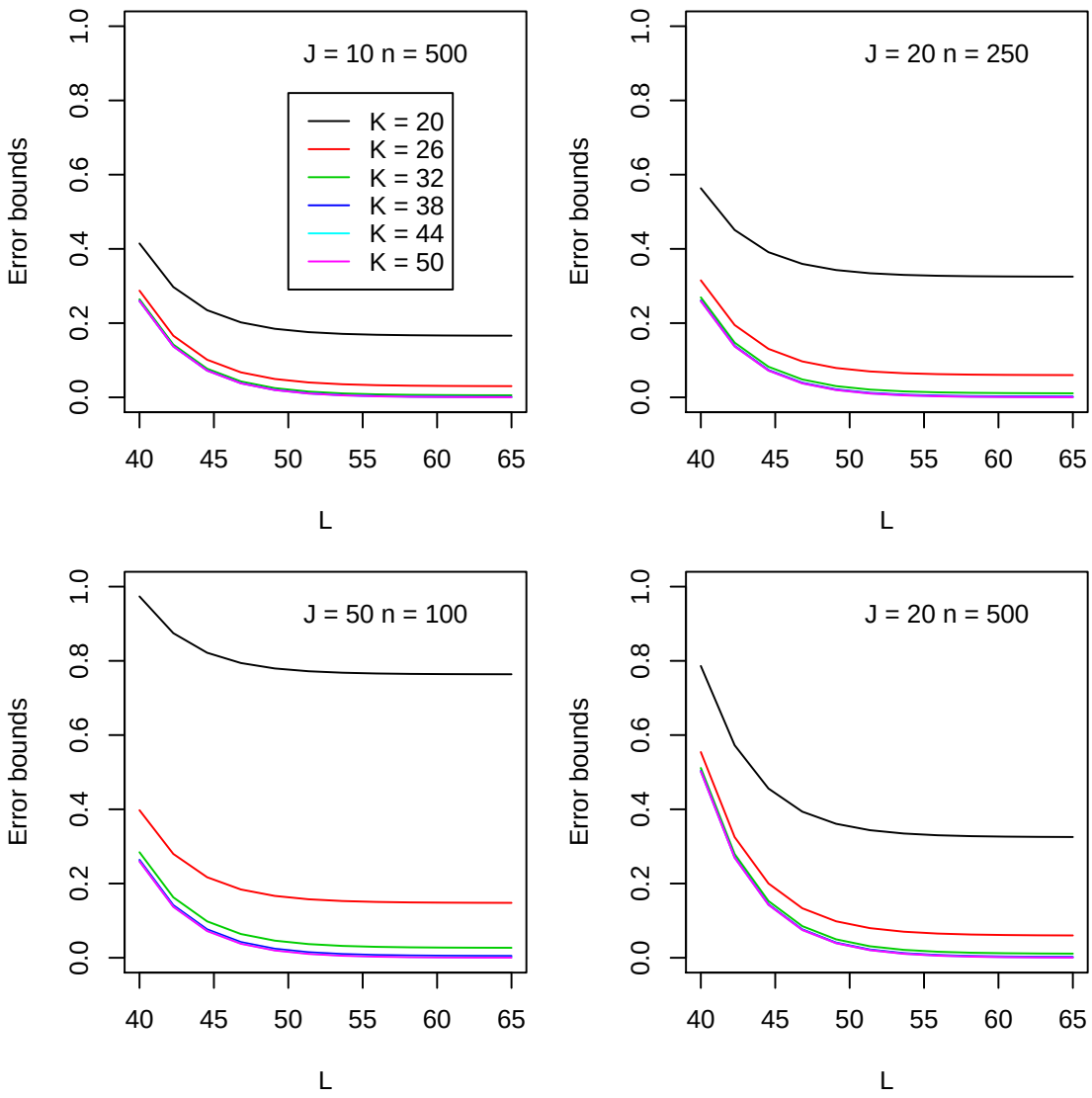


Figure 2.2: Approximate error bounds for the LK truncation of a $\mathsf{nDP}(3, 3, H)$. Top left corner corresponds to $n = 500$ and $J = 10$, top right to $n = 250$ and $J = 20$, bottom left to $n = 100$ and $J = 50$ and bottom right to $n = 500$ and $J = 20$.

Implementations of the nDP based on (1) are, in general, infeasible. Although sampling ξ_{ij} given $(\zeta_1, \ldots, \zeta_J)$ using a Pólya urn scheme is straightforward, sampling ζ_j requires the evaluation of the predictive distributions $p(\mathbf{y}_j|H)$ or $p(\mathbf{y}_j|\{\mathbf{y}_s|\zeta_s = k\})$ (both of which are finite mixtures with a number of terms that grows exponentially with n_j), or the conditional $p(\mathbf{y}_j|G_s^*)$ (whose evaluation requires an infinite sum since $G_s^* \sim \mathsf{DP}(\beta H)$). Details can be seen in appendix D. Algorithms using RJMCMC in the nDP are likely to run into similar problems, with the added disadvantage of low acceptance probabilities due to the large number of parameters that need to be proposed at the same time, without any obvious way to construct efficient proposals. Hence, we focus on combinations of truncation approximations.

2.3.1 Sampling by double truncation

The obvious starting place is to consider a two-level truncation of the process using values of K and L elicited from plots like those shown in Figure 2.2. Once adequate values of K and L have been chosen, computation proceeds through the following steps:

1. Sample the center indicators ζ_j for $j = 1, \ldots, J$ from a multinomial distribution with probabilities

$$\mathbb{P}(\zeta_j = k|\cdots) = q_k^j \propto w_k^* \prod_{i=1}^{n_j} \sum_{l=1}^{L} \pi_{lk} p(y_{ij}|\boldsymbol{\theta}_{lk}^*)$$

2. Sample the group indicators ξ_{ij} for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$ from another multinomial distribution with probabilities

$$\mathbb{P}(\xi_{ij} = l|\cdots) = b_{ij}^l \propto \pi_{l\zeta_j}^* p(y_{ij}|\boldsymbol{\theta}_{l\zeta_j}^*)$$

3. Sample w_{lk}^* by generating

$$(v_{lk}^*|\cdots) \sim \mathsf{beta}\left(1 + m_{lk}, \beta + \sum_{s=l+1}^{L} m_{ls}\right) \quad l = 1, \ldots, L-1, \qquad v_{Lk}^* = 1$$

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

4. Sample π_k^* by generating

$$(u_k^*|\cdots) \sim \mathsf{beta}\left(1 + m_k, \alpha + \sum_{s=k+1}^{K} m_s\right) \quad k = 1, \ldots, K-1 \qquad u_K^* = 1$$

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

5. Sample $\boldsymbol{\theta}_{lk}^*$ from

$$p(\boldsymbol{\theta}_{lk}^*|\cdots) \propto \left[\prod_{\{i,j|\zeta_j=k,\xi_{ij}=l\}} p(y_{ij}|\boldsymbol{\theta}_{lk}^*)\right] p(\boldsymbol{\theta}_{lk}^*)$$

Note that if no observation is assigned to a specific cluster, then the parameters are drawn from the prior distribution (baseline measure) $p(\boldsymbol{\theta}_{lk}^*)$. Also, if the prior is conjugate to the likelihood then sampling is greatly simplified. However, non-conjugate priors can be accommodated using rejection sampling or Metropolis-Hastings steps.

6. Sample the concentration parameters α and β from

$$p(\alpha|\cdots) \propto \alpha^{K-1} \exp\left\{\alpha \sum_{k=1}^{K-1} \log(1 - u_k^*)\right\} p(\alpha)$$

$$p(\beta|\cdots) \propto \beta^{K(L-1)} \exp\left\{\beta \sum_{l=1}^{L-1} \sum_{k=1}^{K} \log(1 - v_{lk}^*)\right\} p(\beta)$$

If conditionally conjugate priors $\alpha \sim \mathsf{Gam}(a_\alpha, b_\alpha)$ and $\beta \sim \mathsf{Gam}(a_\beta, b_\beta)$ are chosen then,

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

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

Note that the accuracy of the truncation depends on the values of α and β. Thus, the hyperparameters (a_α, b_α) and (a_β, b_β) should be chosen to give little prior probability to values of α and β larger than those used to calculate the truncation level.

Besides the simplicity of its implementations, an additional advantage of this truncation scheme is that implementation in parallel computing environments is straightforward, which is useful for large sample sizes. Note that the most computationally expensive steps are (1), (2) and (5). However, $(\zeta_j|\cdots)$ and $(\zeta_{j'}|\cdots)$ are independent, and so are the pairs $(\xi_{ij}|\cdots)$ and $(\xi_{i'j'}|\cdots)$, and $(\boldsymbol{\theta}^*_{lk}|\cdots)$ and $(\boldsymbol{\theta}^*_{l'k'}|\cdots)$. Hence, steps (1), (2) and (5) can be divided into subprocesses that can be run in parallel.

2.3.2 Sampling by one-level truncation

In order to compute predictive probabilities needed to sample the center indicators, only the top-level truncation is strictly necessary. If this level is truncated, $\zeta_1, \ldots, \zeta_J$ can be sampled using a regular Pólya Urn scheme avoiding the need for the second truncation. However, even a prior $p(\boldsymbol{\theta}_{ij})$ conjugate to the likelihood $p(y_{ij}|\boldsymbol{\theta}_{ij})$ does not imply a conjugate model on the distributional level. Hence, Pólya urn methods for *non-conjugate distributions* (MacEachern and Müller, 1998; Neal, 2000) need to be employed in this setup, greatly reducing the computational advantages of the Pólya urn over truncations.

The resulting algorithm, inspired by algorithm 8 in Neal (2000), goes through the steps described above with the following modifications

(1) Choose $m \geq 1$. For each j, let J^{*-} be the number of distinct ζ_s for $s \neq j$, label these ζ_s with values in $\{1, \ldots, J^{*-}\}$, and let $h = J^{*-} + m$. If $\zeta_j = \zeta_s$ for some $s \neq j$ draw new distributions $G^*_{J^{*-}+1}, \cdots, G^*_h$ from their prior by sampling L atoms from H and L weights from the truncated stick-breaking process for each of them. If $\zeta_j \neq \zeta_s$ for all $j \neq s$ let $\zeta_j = J^{*-} + 1$ and draw distributions $G^*_{J^{*-}+2}, \cdots, G^*_h$ from their prior. Then ζ_j is drawn from the a multinomial distribution with

$$\mathbb{P}(\zeta_j = k | \cdots) \propto \begin{cases} \frac{r_k^-}{J-1+\alpha} \prod_{i=1}^{n_j} \sum_{l=1}^{L} w^*_{lk} p(y_{ij}|\boldsymbol{\theta}_{lk}) & \text{for } k \leq J^{*-} \\ \frac{\alpha/m}{J-1+\alpha} \prod_{i=1}^{n_j} \sum_{l=1}^{L} w^*_{lk} p(y_{ij}|\boldsymbol{\theta}_{lk}) & \text{for } J^{*-} < k \leq h \end{cases}$$

where r_k^- is the number of distributions assigned to atom G^*_k once observation j has been removed.

(3) This step is unnecessary.

(6) The concentration parameter α should be sampled from

$$p(\alpha | \cdots) \propto \alpha^{J^*} \frac{\Gamma(\alpha)}{\Gamma(\alpha + J)} p(\alpha)$$

where J^* is the number of distinct distributions in the current iteration of the algorithm. Again under a $\mathsf{Gam}(a_\alpha, b_\alpha)$ prior, posterior sampling can be accomplished through the data augmentation method discussed in Escobar and West (1995).

Although this algorithm saves memory and computation time (since it is not necessary to update empty components), mixing could be a concern. On the other hand, most of the comments we made before on parallel implementation hold for

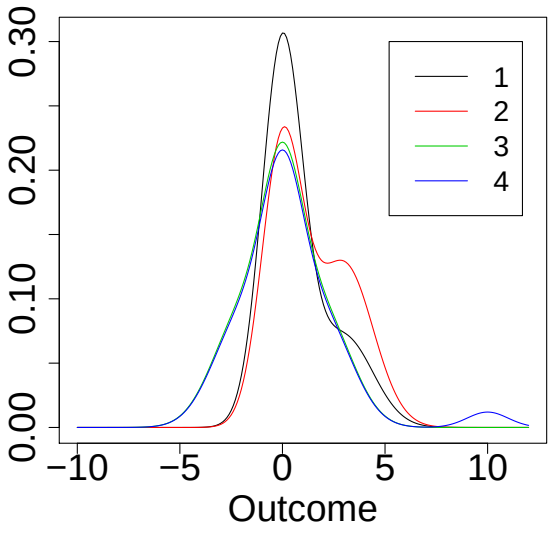


Figure 2.3: True distributions used in the simulation study on the nDP.

this sampling algorithm. However, $\zeta_j|\cdots$ now depends on the other supergroup indicators, so parallelization in this step is not straightforward.

2.4 Simulation study

In this section we present a simulation study designed to provide insight into the discriminating capability of the nDP, as well as its ability to provide more accurate density estimates by borrowing strength across centers. The set up of the study is as follows: J samples of size n are obtained from four mixtures of four Gaussians defined in table 2.1 and plotted in Figure 2.3. These distributions have been chosen to reflect situations that are conceptually hard: T1 and T2 are asymmetric and composed of the same two Gaussian components which have been weighted differently, while T3 and T4 share three distributions located symmetrically around the origin, differing only in an additional bump that T4 presents on the right tail.

The value of J and n was varied across the study in order to assess the influ-

Table 2.1: Parameters for the true distributions $p_T(\cdot) = \sum_i w_i \mathsf{N}(\cdot|\mu_i, \sigma_i^2)$ used in the simulation study on the nDP.

Distrib	Comp 1			Comp 2			Comp 3			Comp 4		
	w	μ	σ^2	w	μ	σ^2	w	μ	σ^2	w	μ	σ^2
T1	0.75	0.0	1.0	0.25	3.0	2.0	-	-	-	-	-	-
T2	0.55	0.0	1.0	0.45	3.0	2.0	-	-	-	-	-	-
T3	0.40	0.0	1.0	0.30	−2.0	2.0	0.30	2.0	2.0	-	-	-
T4	0.39	0.0	1.0	0.29	−2.0	2.0	0.29	2.0	2.0	0.03	10.0	1.0

ence of the sample size on the discriminating capability of the model. The precision parameters α and β were both fixed to 1 and a Normal Inverse-Gamma distribution $\mathsf{NIG}(0, 0.01, 3, 1)$ was chosen as the baseline measure H, implying that a priori $\mathbb{E}(\mu|\sigma^2) = 0$, $\mathbb{V}(\mu|\sigma^2) = 100\sigma^2$, $\mathbb{E}(\sigma^2) = 1$ and $\mathbb{V}(\sigma^2) = 3$. The algorithm described in section 2.3.1 was used to obtain samples of the posterior distribution under the nDP. Following the discussion in section 2.2, truncation levels were chosen as $K = 35$ and $L = 55$. All results shown below are based on 50,000 samples obtained after a burn-in period of 5,000 iterations.

Visualization of high dimensional clustering structures is a hard task. A summary commonly employed looks at the set of $J(J-1)/2$ possible pairs of populations and, for each pair, obtains the probability that the two of them fall in the same cluster. Estimates of these probabilities are easily obtained from the output of our MCMC algorithm and can be effectively displayed using heatmaps, like those in Figure 2.4. To simplify interpretation of the plot, samples from the same mixture distribution are adjacent. Other possible summaries are discussed in section 2.5.

For small values of n, the nDP is able to roughly separate T1 and T2 from T3 and T4, but not to discriminate between T1 and T2 or T3 and T4. This is not really surprising: the method is designed to induce clustering. Therefore, when differences are highly uncertain, it prefers to create less rather than more clusters. However, as n increases, the model is able to distinguish between distributions and correctly identify

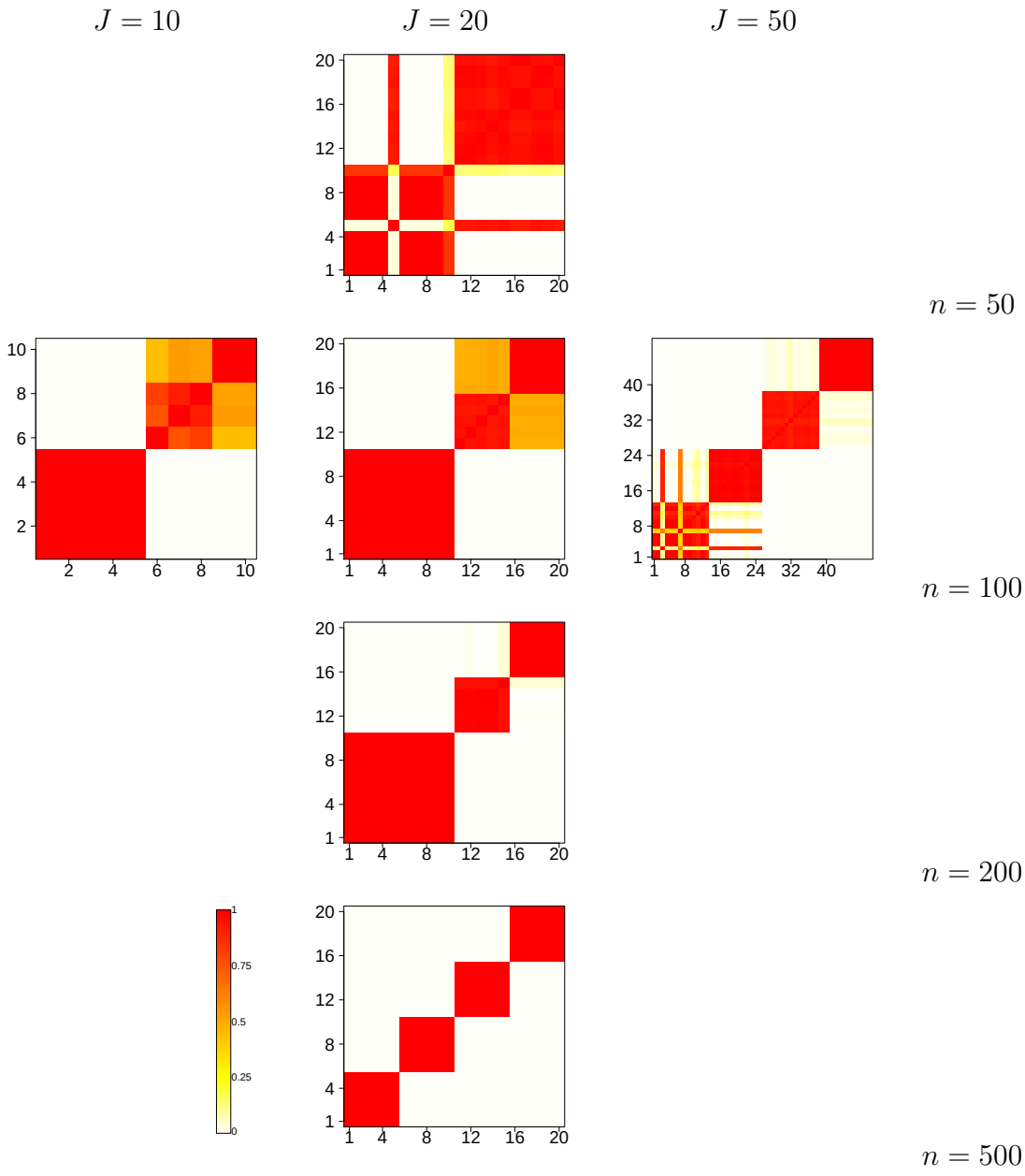


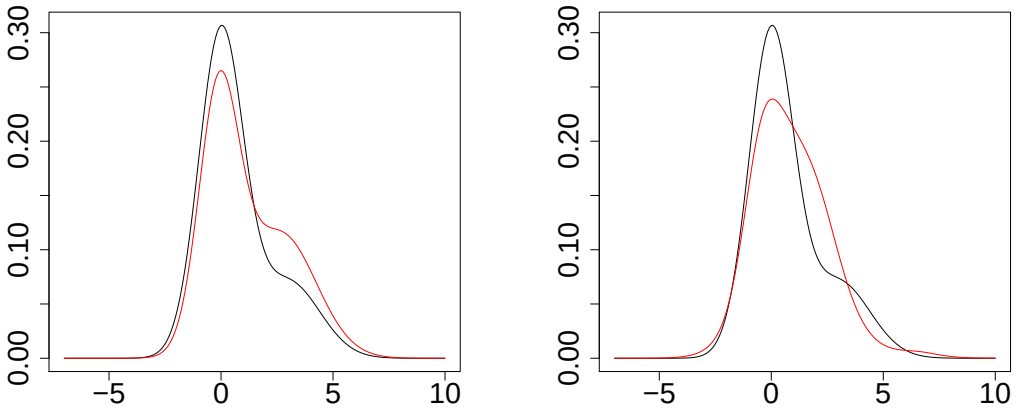
Figure 2.4: Pairwise probabilities of joint classification for the simulation study on the nDP

both the number of groups and the membership of the distributions. It is particularly interesting that the model finds it easier to discriminate between distributions that differ just in one atom rather than in weights. On the other hand, as J increases the model is capable of discovering the underlying groups of distributions, but the uncertainty on the membership is not reduced without increasing n.

In Figure 2.5 we show density estimates obtained for sample 1 of the example $J = 20$, $n = 100$. The left panel shows the one obtained from the nDP (which borrows information across all samples), while the right panel was obtained by fitting a regular DPM model with the same precision parameter $\beta = 1$ and baseline measure. We note that, although the nDP borrows information across samples that actually come from a slightly different data-generation mechanism, the estimate is more accurate: it not only captures the small mode to the right more clearly, but it also emphasizes the importance of the main mode. Indeed the Kullback-Leibler of the density estimate relative to the true distribution for the estimate of T1 under the nDP is 0.011, while under the regular DPM it was 0.017.

2.5 An application: Health care quality in United States

Data on quality of care in hospitals across the United States and associated territories is made publicly available by the Department of Health and Human Services at the website `http://www.hhs.gov/`. Twenty measures are recorded for each hospital, comprising aspects like proper and timely application of medication, treatment and discharge instructions. In what follows we focus on one specific measure: the proportion of patients that were given the most appropriate initial antibiotic(s), transformed through the logit function. Four covariates are available for each center: type of hospital (either acute care or critical access), ownership (nine possible levels, including



(a) (b)

Figure 2.5: True (black) and estimated (red) densities for distribution 1 of the simulation with $J = 20$ and $n = 100$. Panel (a) corresponds to an estimate based on the nDP, which borrows information across all samples, while panel (b) corresponds to an estimate based only on sample 1.

government at different levels, proprietary and different types of voluntary non-profit hospitals), whether the hospital provides emergency services (yes or no) and whether it has an accreditation (yes or no). Location, in the form of the ZIP code, is also available. Hospitals with less than 30 patients treated and territories with less than 4 hospitals were judged misrepresentative and removed from the sample, yielding a final sample size of 3077 hospitals in 51 territories (the 50 states plus the District of Columbia). Number of hospitals per state varies widely, with 5 in Delaware, 10 in Alaska, 13 in Idaho, 164 in Florida, 205 in Texas and 254 in California. The number of patients per hospital varies between 30 and 1175, with quartiles at 76, 130 and 197 patients. Since the value tends to be large, we perform our analysis on the observed proportion without adjusting for sample sizes.

We wish to study differences in quality of care across states after adjusting for the effect of the available covariates. Specifically, we are interested in clustering states according to their quality rather than getting smoothed quality estimates. Indeed,

differences in quality of care are probably due to a combination of state policies and practice standards, and clustering patterns can be used to identify such factors. Therefore, there is no reason to assume a priori that geographically neighboring states have similar outcomes.

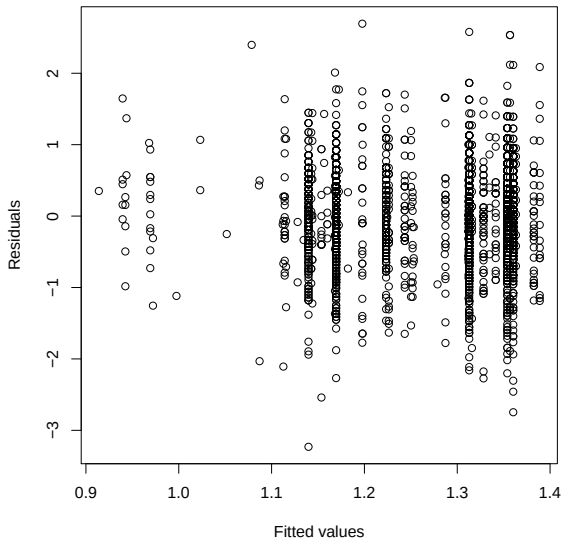
In order to motivate the use of the nDP, we consider first a simple preliminary analysis of the data. To adjust for the covariates, an ANOVA model containing only main effects was fitted to the data. Of these effects, only the presence of an emergency service and the ownership seem to affect the quality of the hospital (p-values 0.011 and 1.916×10^{-8}). Residual plots for this model show some deviation from homocedasticity and normality (see Figure 2.6), but given the large sample size it is unlikely that this has any impact on the results so far.

It is clear from Figure 2.7 that residual distributions vary across states. At this point, one possible course of action is to assume normality within each state and cluster states according to the mean and/or variance of its residual distribution. However, the density estimates in Figure 2.8 (obtained using Gaussian kernels with a bandwidth chosen with the rule of thumb described in Silverman (1986)) show that state-specific residual distributions can be highly non-normal and that changes across states can go beyond location and scale changes to affect the whole shape of the distribution. Invoking asymptotic arguments at this point is not viable since sample sizes are small and we are dealing with the shape of the distribution (rather than the parameters), for which no central limit theorem can be invoked.

Figure 2.8 also shows that states located in very different geographical areas can have similar error distributions, like California and Minnesota or Florida and North Carolina.

To improve the analysis, we resort to a Bayesian formulation of the main-effects ANOVA and use the nDP to model the state-specific error distributions. The model

(a)



(b)

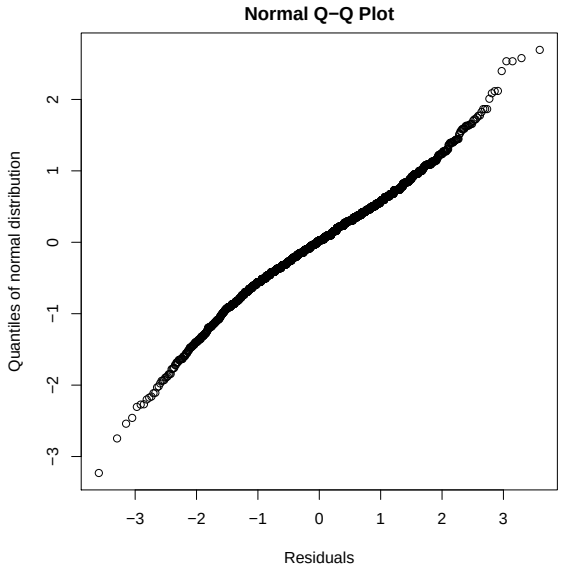


Figure 2.6: Residual plots for the ANOVA model on the initial antibiotic data: (a) Residuals vs. fitted values, (b) Quantile-quantile plot

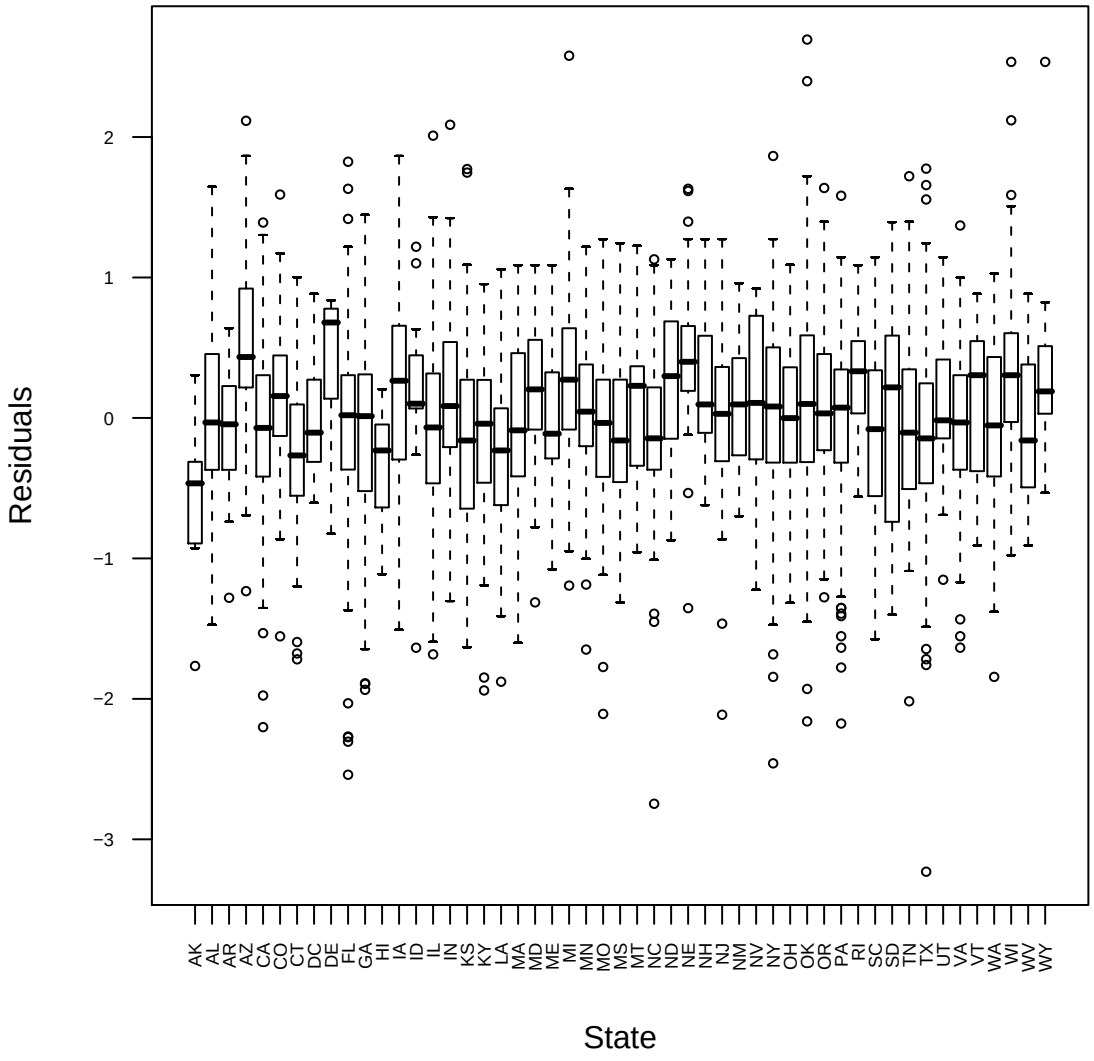


Figure 2.7: State-specific residual boxplots for the ANOVA model on the initial antibiotic data.

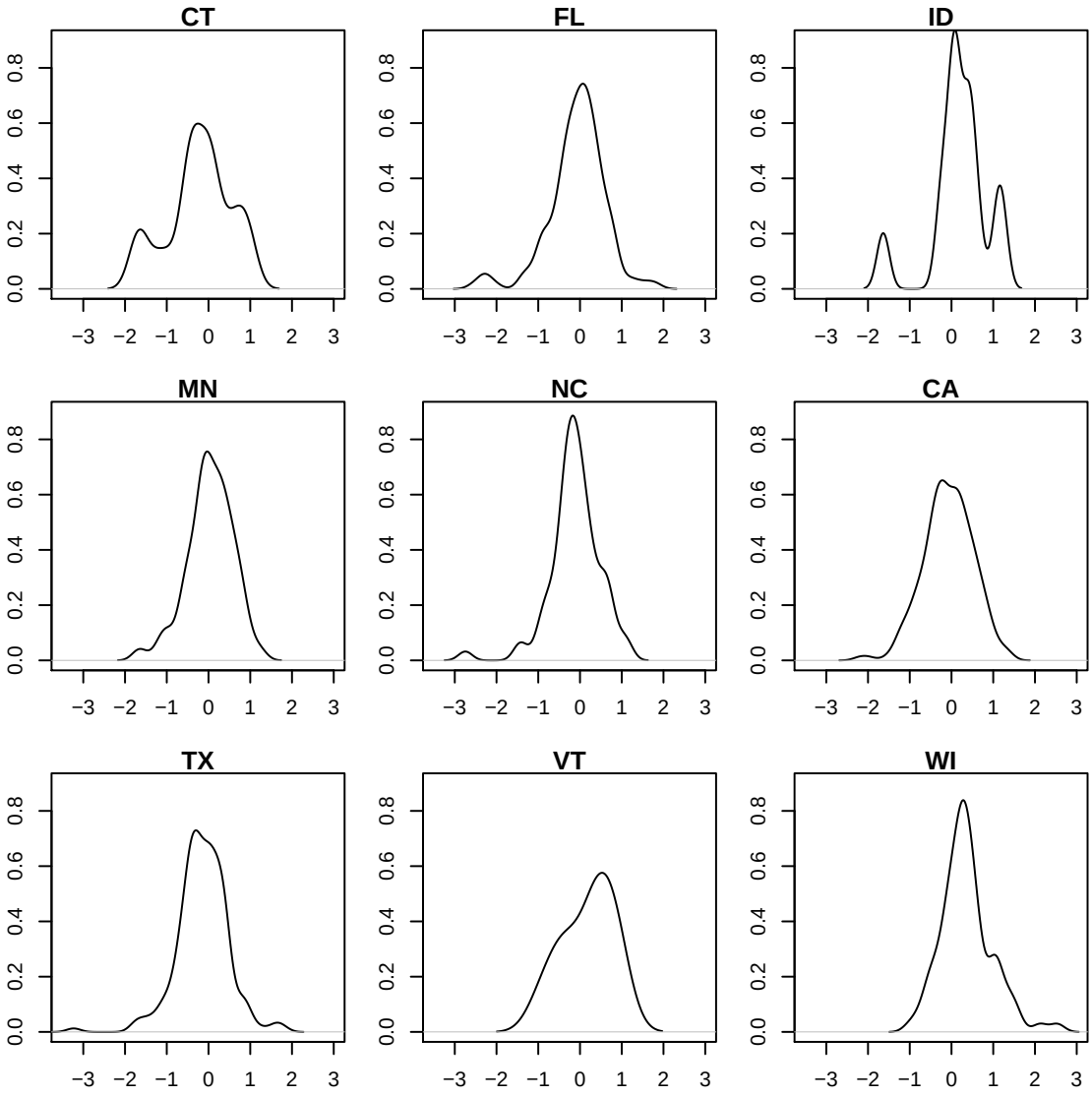


Figure 2.8: Density estimates for the residual distribution in selected states. Note that distributions seem clearly non-normal and that their shape can have important variations, making any parametric assumption hard to support.

is similar in spirit to those in West *et al.* (1998) and Burgess *et al.* (2000), who use the *non-returns to follow-up care* as a measure of quality. Specifically, if we let y_{ij} be the response of hospital i in state j after subtraction of the global mean:

$$y_{ij} = \mu_{ij} + \mathbf{x}_{ij}\boldsymbol{\gamma} + \epsilon_{ij} \qquad \epsilon_{ij} \sim \mathsf{N}(0, \sigma_{ij})$$
$$(\mu_{ij}, \sigma^2_{ij}) \sim G_j \qquad \{G_1, \ldots, G_J\} \sim \mathsf{nDP}(\alpha, \beta, H)$$

where $\mathbf{x}_{ij}$ is the vector of covariates associated with the hospital. Prior elicitation is simplified by centering the observations. We pick $H = \mathsf{NIG}(0, 0.01, 3, 3)$, which implies $\mathbb{E}(\mu|\sigma^2) = 0$, $\mathbb{V}(\mu|\sigma^2) = 100\sigma^2$, $\mathbb{E}(\sigma^2) = 1$ and $\mathbb{V}(\sigma^2) = 3$. This choice reflects the natural scale (logit) of the data, which, on a Gaussian linear model, would be expected to have mean zero and variance close to unit after adjusting for covariates. We use a standard reference (flat) prior on $\boldsymbol{\gamma}$. Finally, we set $\alpha, \beta \sim \mathsf{Gam}(3, 3)$ a priori, implying that $\mathbb{E}(\alpha) = \mathbb{E}(\beta) = 1$ (a common choice in the literature) and $\mathbb{P}(\alpha > 3) = \mathbb{P}(\beta > 3) \approx 0.006$. Note that this choice implies that $\mathbb{P}(\mathbb{C}\text{or}(G_j, G_{j'}) > 0.25) \approx 0.994$.

Posterior computation is a straightforward using the algorithm presented in section 2.3.1. As described there, the model is a regular ANOVA with known variance conditional on $\boldsymbol{\theta} = (\mu_{ij}, \sigma_{ij})$, and the full conditional posterior distribution of $\boldsymbol{\gamma}$ (which corresponds to $\boldsymbol{\phi}$ in our general notation) following a normal distribution. On the other hand, conditional on $\boldsymbol{\gamma}$, we can use the nDP sampler on the pseudo-observations $z_{ij} = y_{ij} - \mathbf{x}_{ij}\boldsymbol{\gamma}$. Results below are based on 50,000 iterations obtained after a burn-in period of 5,000 iterations. As with the simulation study, we choose $K = 35$ and $L = 55$ as the truncation levels. Results seem to be robust to reasonable changes in prior specification and different initial states for the variables in the sampler. There was no evidence of lack of convergence from visual inspection of trace plots.

The posterior distribution on the number of distinct states shows strong evidence

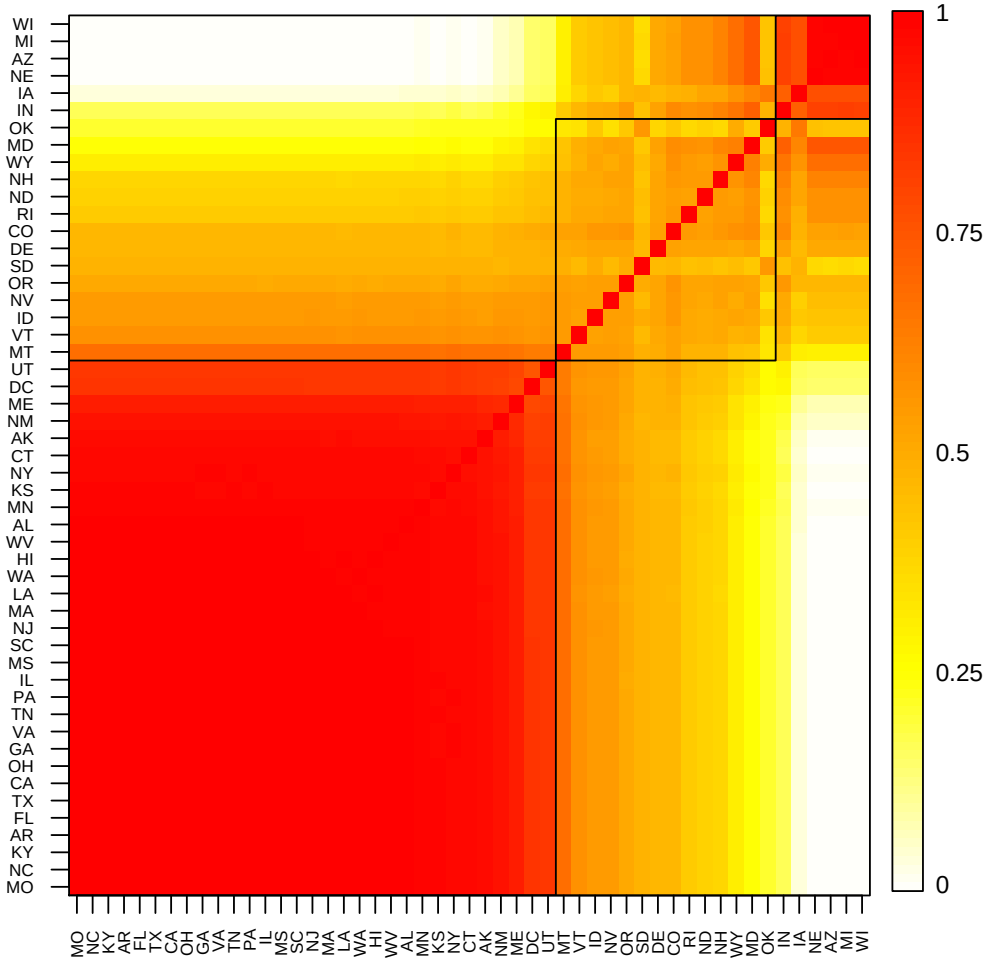


Figure 2.9: Residual plots for the ANOVA model on the initial antibiotic data.

in favor of either 2 or 3 components (posterior probabilities 0.616 and 0.363 respectively), and little support for either 1, 4 or 5 distributions (posterior probabilities 0.00, 0.02 and 0.001 respectively). As with the simulated example, we visualize the matrix of pairwise probabilities using a heatmap, which is shown in Figure 2.9. In order to make sense of the plot, we first reorder the states using an algorithm inspired by those used for hierarchical clustering.

This heatmap provides additional insight into the clustering structure. It shows

three well defined groups: (1) a large homogenous clusters of 31 members (lower left corner of the plot); (2) a small homogenous cluster of 6 states (upper right corner); and (3) an heterogeneous group made of the remaining 15 states, which are not clear members of any of the two previous clusters and do not seem to form a coherent cluster among themselves.

A few different approaches can be used to choose one specific partition of the set of States. One appealing option is to choose $\hat{\mathbf{p}}$ such that it minimizes a given loss functions. Following Binder (1978, 1981); Lau and Green (2006), we chose the label-invariant loss function

$$\Psi(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{\{(j,j'):j<j'\leq J\}} \left(a\mathbf{1}_{(\zeta_j=\zeta_{j'},\hat{\zeta}_j\neq\hat{\zeta}_{j'})} + b\mathbf{1}_{(\zeta_j\neq\zeta_{j'},\hat{\zeta}_j=\hat{\zeta}_{j'})} \right) \tag{2.4}$$

where $\mathbf{p}$ denotes the true (unknown) partition of states, $\mathbf{1}_A$ is the indicator function on the set A, ζ_j and $\hat{\zeta}_j$ denote the true and estimated clustering indicators induced by $\mathbf{p}$ and $\hat{\mathbf{p}}$, and a and b are pairwise misclassification penalties. Minimizing the posterior expected loss under Ψ is equivalent to picking a partition $\hat{\mathbf{p}}$ such that the function

$$\sum_{\{(j,j'):j<j'\leq J\}} \mathbf{1}_{(\hat{\zeta}_j=\hat{\zeta}_{j'})}(\rho_{jj'} - \tau)$$

is maximized, where $\rho_{jj'}$ is the probability of joint classification for states i and j (which are the values depicted in Figure 2.9) and $\tau = b/(a+b) \in [0, 1]$. For $\tau = 0$, the optimal partition places all states in a single cluster (since we are only worried about not putting together states that should be in the same cluster). On the other hand, if $\tau = 1$, the optimal allocation creates individual clusters for each state (since we are only worried about erroneously putting together states that should be separated). Intermediate values of τ correspond to a compromise between both types of errors. Hence, for $\tau = 0.3$, the optimal partition divides the 51 states in two groups, a small one comprising 8 states (AZ, IA,IN, MD, MI, NE, OK and WI) and a large one

containing all the remaining states. For $\tau = 0.5$, the optimal allocation corresponds to three clusters: a very small one comprising only OK and SD, an intermediate one comprising AZ, CO, DE, IA, IN, MD, MI, ND, NE, NH, RI, WI and WY (note the similarities with $\tau = 0.3$) and a large cluster with the remaining states. Finally, for $\tau = 0.75$, the optimal clustering agrees with the one depicted in Figure 2.9, with two tight groups and 14 single-state clusters. The posterior probabilities for each of these partitions estimated from the MCMC are 4×10^{-5}, 0 and 0 respectively. In contrast, the most frequent configuration sampled by the model (posterior probability 7×10^{-4}, much larger but still rather small), divides the sample in two groups, a small one with 17 states (AZ, CO, IA, ID, IN, MD, MI, ND, NE, NH, NV, OK, OR, RI, SD, WI and WY), and another with the rest.

We can also study the clustering of hospitals within states, but meaningful interpretations have to be done *conditionally* on the state-level partition. As an illustration, consider conditioning in the optimal clustering suggested by taking $\tau = 0.75$. The small cluster (comprising AZ, IA, IN, MI, NE and WI) is made of one (posterior probability 0.81) or two (posterior probability 0.18) groups of hospitals, while the large cluster (comprising 31 states including TX and NC) comprises between 2 (probability 0.89) or three different groups of hospitals (probability 0.10). This shows that low/high quality groups of hospitals can be identified within each group of states and state-specific distributions are non-normal as expected.

Indeed, Figure 2.10 shows posterior predictive density estimates for four representative states: North Carolina (cluster 1), Wisconsin (cluster 2), and South Dakota and Oklahoma, which belong to the third group. North Carolina (and, in general, the states in group 1) presents a lower mean and a heavier-than-Gaussian left tail, indicating that each of those states contains some underperforming hospitals and few or none over performing hospitals. The situation for Wisconsin and cluster 2

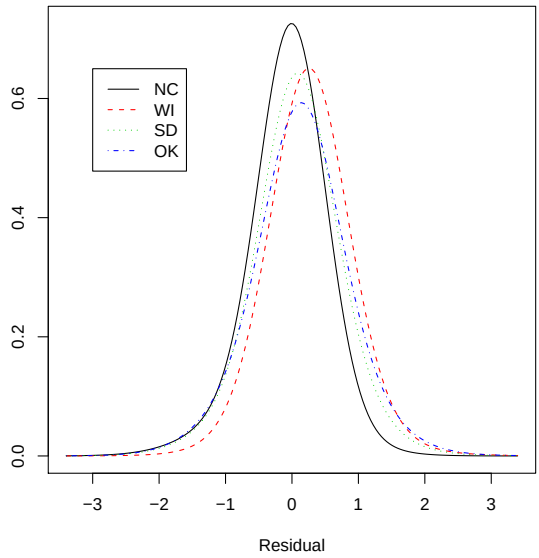


Figure 2.10: Mean predictive density for four representative states: North Carolina (NC), Wisconsin (WI), South Dakota (SD) and Oklahoma (OK).

is reversed: these seem to be states with a higher average performance, quite a few hospitals that have an excellent record in the application of antibiotics and few or no low-quality hospitals. Finally, South Dakota and Oklahoma present a mixed behavior, showing evidence for both under and over performing hospitals. Note that these density estimates are much smoother than those in Figure 2.8. This is not surprising for three reasons: 1) as discussed in Escobar and West (1995), location-scale mixtures act as adaptive-bandwidth kernel estimators, 2) we are borrowing information across estates; and 3) our estimates average over a large number of alternative models, which induces smoothness. All of these features tend to produce smoother estimates than those obtained from standard kernel density estimates.

It is interesting to contrast these results with those obtained from a similar model that uses the HDP instead of the nDP to induce dependence among residual distributions. Although density estimates (not shown) for the different states look similar to those in Figure 2.10, the HDP does not provide an equivalent to Figure 2.9, as

it only clusters hospitals and not states. Indeed, the HDP-based model divides the 3077 hospitals in roughly 3 groups, which we could easily label as average (for the largest, central group), underperformer and over performers (both containing a relatively small number of observations). The density estimates are then obtained by weighting these groups differentially for each state.

2.6 Discussion

We have formulated a novel extension of the Dirichlet process for a family of a priori exchangeable distributions that allows us to simultaneously cluster groups and observations within groups. Moreover, the groups are clustered by their entire distribution rather than by particular features of it. After examining some of the theoretical properties of the model, we describe a computationally efficient implementation and demonstrate the flexibility of the model through both a simulation study and an application where the nDP is used to jointly model the random effect and error distribution of an ANOVA model. We also offer heatmaps to summarize the clustering structure generated by the model. Attractively, while being nonparametric, the nDP encompasses a number of typical parametric and nonparametric models as limiting cases. Therefore the model is flexible while avoiding issues of model specification that be hard in practical applications

One natural generalization of the nDP is to replace the $\mathsf{beta}(1, \alpha)$ and $\mathsf{beta}(1, \beta)$ stick-breaking densities with more general forms. In the setting of stick-breaking priors for a single random probability measure, Ishwaran and James (2001) considered general $\mathsf{beta}(a_k, b_k)$ forms, with the DP corresponding to the special case $a_k = 1$, $b_k = \alpha$. Similarly, by using $\mathsf{beta}(a_k, b_k)$ and $\mathsf{beta}(c_k, d_k)$ respectively, we can obtain a rich class of nested stick-breaking priors that encompasses the nDP as a particular case.

Another obvious generalization of the nDP is to enrich the stick breaking process. In principle, the random u_k and v_{lk} can be drawn independently from an arbitrary distribution on $[0, 1]$ (see Hjort (2000)). For example, Ishwaran and Zarepour (2002) suggest $\mathsf{beta}(a, b)$ distributions, while Ishwaran and James (2001) discuss the general case $\mathsf{beta}(a_k, b_k)$.

Including hyperparameters in the baseline measure H is another straightforward extension. We note that, conditional on H, the distinct atoms $\{G_k^*\}_{k=1}^{\infty}$ are assumed to be independent. Therefore, including hyperparameters in H allows us to parametrically borrow information across the distinct distributions.

Chapter 3

Multilevel clustering in reproductive function studies

Infertility and early pregnancy loss (EPL) are currently major public health issues in the US that can be financially and emotionally costly, both for couples and society as a whole. Studies carried out during the 90's showed that around 7% of married couples report difficulties in achieving a pregnancy, while about 6,000,000 women (10% of the women in the 15-44 age range) reported impaired fecundity or the inability to establish or maintain a pregnancy (Fidler and Bernstein, 1999). On the other hand, early pregnancy loss could represent around 30% of all pregnancies (Wilcox *et al.*, 1998).

Reproductive function studies aim to explore biological and environmental causes for female infertility and EPL. Detailed reproductive function studies record, along with demographic characteristics and outcome variables, daily hormonal levels along multiple menstrual cycles for each woman in the study. Certain characteristics of these hormonal profiles (like baseline or peak levels) are known to be correlated with important outcomes like fertility and EPL (see Venners *et al.* (2004), Baird *et al.* (1997) and references therein). Therefore, it is important to develop methods that allow for response variables in the form of random functions.

Some popular tools for modeling random functions include Gaussian processes (Rasmussen and Williams, 2006), kernel methods (Altman, 1992; Chu and Marron, 1991; Fan *et al.*, 1995), wavelet decompositions (Vidakovic, 1999) and splines (Truong *et al.*, 2005). Splines are smooth, piecewise polynomial functions that provide remarkable flexibility and can typically be fitted using tools borrowed from multivariate linear regression. Splines allow us to represent the unknown random function as a linear combination of simpler basis functions. That way, the problem of finding a prior on an infinite-dimensional space is reduced to a finite dimensional problem, which is, putting a prior on the basis coefficients. From this point of view, splines, as well as wavelet and kernel methods, can be cataloged together as "basis expansion" procedures.

Splines have been previously used to model hormonal profiles. For example, Brumback and Rice (1998) develop a model for nested and crossed samples of curves based on natural cubic splines. In the context of functional clustering, Bigelow and Dunson (2007) have used Dirichlet process priors as part of the hierarchical specification of the model coefficients in order to induce clustering across curves. These methods are designed for problems where only one curve is obtained for each woman and the goal is to cluster women. However, when multiple curves are obtained for each individual, the results obtained from these methods can be misleading since they end up using an average curve to represent the group.

In this chapter, we use the nested Dirichlet process to construct a functional clustering algorithm that uses the *distribution of curves* to construct groups of women with similar hormonal profiles. Our motivation comes from the Early Pregnancy Study (Wilcox *et al.*, 1998), where daily measurements of progesterone levels were collected over a six month period for 221 women. Our goal is to cluster women, identifying typical and unusual clusters of women and generate hypotheses about

the mechanisms underlying the hormonal process by using woman-specific covariates. As a byproduct of the model, we are also able to impute unobserved levels for some women. Our methods are conceptually related to Ray and Mallick (2006), who developed a model for *one-level clustering* of curves using Dirichlet processes and wavelet basis functions, and to Heard *et al.* (2006), who present a model that uses agglomerative clustering and truncated power spline basis. Instead, the models presented here use splines basis and the nested Dirichlet Processes as priors on the distributions of the coefficients of the spline bases, leading to models that allow for *two-level clustering* of women and curves within women. Extending our methods to other basis functions is straightforward.

In order to illustrate the differences between mean and distribution-based clustering, consider the hormone profiles depicted in figure 3.1. Frames (a) to (c) depict the hormone profiles for 3 women in our data set, while frame (d) shows the mean profile corresponding to each one of them, obtained by simply averaging all available observations at a given day within the cycle. When looking at the mean profiles in (d), women 43 and 36 seem to have very similar hormonal responses, which are different from those of woman 3. However, when the individual profiles are considered, it is clear that most of the cycles of woman 43 look like those of woman 3 and that the big difference in the means is driven by the single abnormal cycle.

The rest of the chapter is organized as follows. Section 3.1 reviews the basic theory behind Bayesian spline models. Section 3.2 describes our model, while section 3.3 develops the Markov Chain Monte Carlo algorithm we employ to fit the model. Section 3.4 shows the results of our method in the (Wilcox *et al.*, 1998). Finally, section 3.5 presents our closing comments and discusses possible extensions and novel applications.

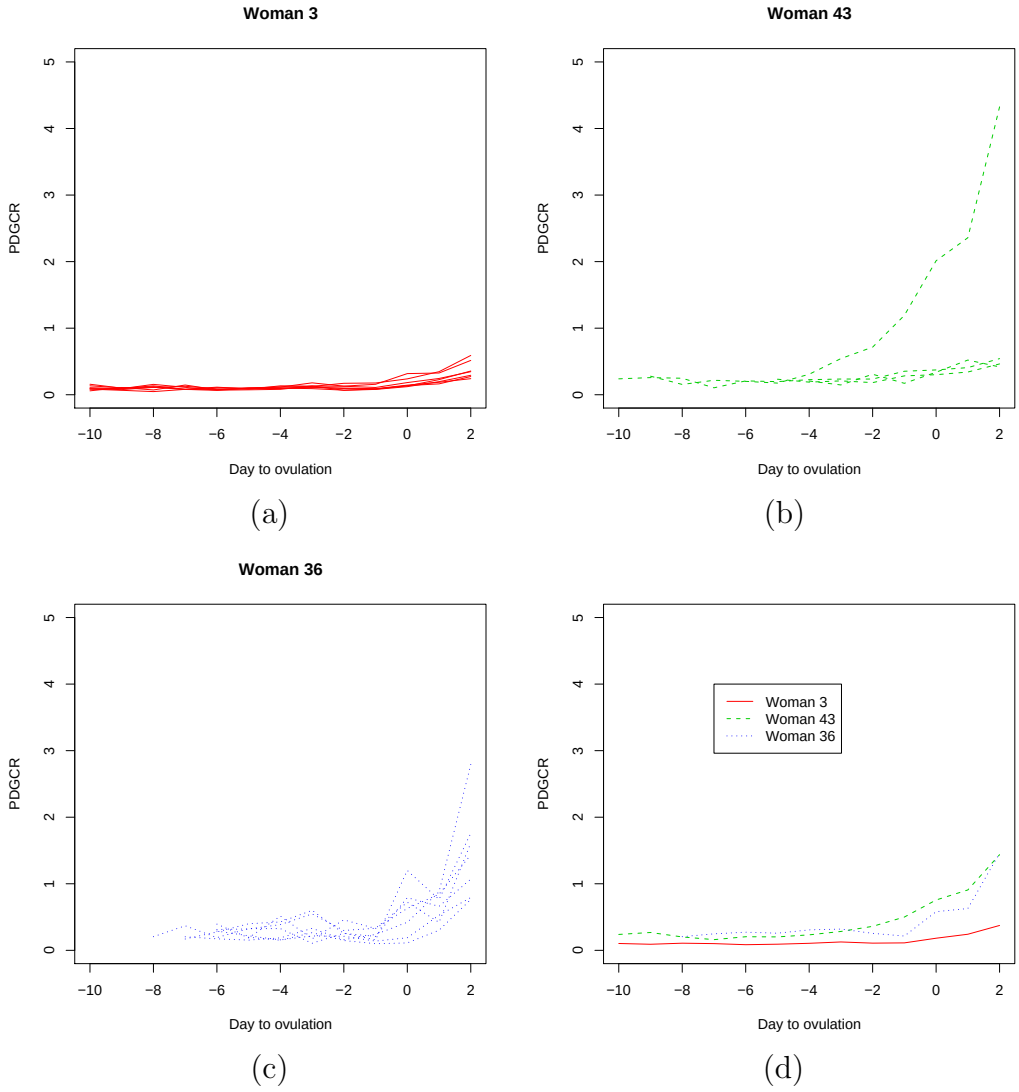


Figure 3.1: Comparison of hormone profiles for three women in the Early Pregnancy Study. Frames (a) to (c) show multiple profiles for each woman, while frame the (d) shows the average profile for each woman.

3.1 Splines and Bayesian nonparametric regression

Given a set of $m+1$ knots $\tau_0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_m$, an order q spline $f(x)$ is a piecewise polynomial function defined on the interval $[\tau_0, \tau_m)$ such that,

$$f(x) = \begin{cases} r_1(x) & \tau_0 \leq x < \tau_1 \\ r_2(x) & \tau_1 \leq x < \tau_2 \\ \vdots & \\ r_m(x) & \tau_{m-1} \leq x < \tau_m \end{cases}$$

where $\{r_i(x)\}_{i=1}^m$ are polynomials of degree q. Typically, continuity and differentiability conditions are attached to the spline, introducing constraints in the coefficients of the different pieces. If the knots are equidistantly distributed in the interval we say the spline is uniform, i.e., $\tau_{j+1} - \tau_j = h$ for every j; otherwise we say it is non-uniform.

For any given knot set, the corresponding set of order q B-splines provide a basis system for order q splines with $q-1$ continuous derivatives. Starting with piecewise constant functions, such basis system can be obtained recursively using De Boors' formula (De Boor, 1978),

$$b_{k,q}(x) = \frac{x - \tau_k}{\tau_{k+q} - \tau_k} b_{k,q-1} + \frac{\tau_{k+q+1} - x}{\tau_{k+q+1} - \tau_{k+1}} b_{k+1,q-1} \qquad k = 0, \ldots, m-q-1$$

$$b_{k,0}(x) = \begin{cases} 1 & \tau_k \leq x < \tau_{k+1} \\ 0 & \text{otherwise} \end{cases}$$

For uniform cubic B-splines, the basis functions reduce to

$$b_{k,q}(x) = b_q^*\left(\frac{x - \tau_k}{h}\right)$$

where h is the distance between knots and

$$b_q^*(x) = \begin{cases} \frac{1}{6}x^3 & 0 \leq x < 1 \\ \frac{1}{6}(-3x^3 + 12t^2 - 12t + 4) & 1 \leq x < 2 \\ \frac{1}{6}(3x^3 - 24t^2 + 60t - 44) & 2 \leq x < 3 \\ \frac{1}{6}(4 - x)^3 & 3 \leq x < 4 \end{cases}$$

and the unknown function can be written as

$$f(x) = \sum_{k=1}^{m-q-1} \theta_k b_{k,q}(x)$$

for some vector of real coefficients $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{m-q-1})$.

In practical applications, deciding on the number of knots is a hard problem, typically requiring complex computational algorithms. A simple approach is to use a large number of knots, together with a penalty term to prevent overfitting (too many knots typically lead to "bumpy" functions). For $q \geq 3$, one commonly used such term is proportional to the total bending energy required to fit the function, given by

$$\int_{\tau_0}^{\tau_m} \left(\frac{\partial^2 f}{\partial x^2} \right)^2 dx = \boldsymbol{\theta}' \boldsymbol{\Omega} \boldsymbol{\theta}$$

$$[\boldsymbol{\Omega}]_{kl} = \int_{\tau_0}^{\tau_m} b_{k,q}(x) b_{l,q}(x) dx$$

This penalty term can be interpreted from a Bayesian perspective as a prior on the spline coefficients. Following Eilers and Marx (1996), we can write the bending energy as

$$\boldsymbol{\theta}' \boldsymbol{\Omega} \boldsymbol{\theta} = \frac{1}{h^2} \left[c_1 \boldsymbol{\theta}' \left(\sum_{j=1}^{m-q-1} \mathbf{d}_j \mathbf{d}_j' \right) \boldsymbol{\theta} + c_2 \boldsymbol{\theta}' \left(\sum_{j=2}^{m-q-1} \mathbf{d}_j \mathbf{d}_{j-1}' \right) \boldsymbol{\theta} \right]$$

where

$$c_1 = \int_{\tau_0}^{\tau_m} b_{k,q-2}^2 dx$$

$$c_2 = \int_{\tau_0}^{\tau_m} b_{k,q-2} b_{k-1,q-2} dx$$

$$\mathbf{d}_j' = (\underbrace{0, \ldots, 0}_{j-1}, 1, -2, 1, \underbrace{0, \ldots, 0}_{m-q-j-3})$$

For cubic B-splines, this reduces to $c_1 = \frac{2h^2}{6}$ and $c_2 = \frac{h^2}{6}$.

Unit-information priors (Paciorek, 2006) or mixtures of g-priors (Liang *et al.*, 2005) are also popular alternatives for prior distributions on the spline coefficients. Finally, other authors, starting with Carter and Kohn (1994), have used independent zero-inflated priors and other mixture priors to induce sparsity.

3.2 Models for functional clustering

In this section, we develop two hierarchical models for functional clustering based on spline representations. The first model uses a regular Dirichlet Process and is intended for clustering women based on their mean profile. It is very similar the model in Bigelow and Dunson (2007), although they allow for uncertainty in the basis functions selection. The second model uses the nested Dirichlet Process and allows us to cluster women according to the distribution of the curves. The models are described as a function of a pre-specified set of basis functions $\{b_k(\cdot)\}_{k=1}^{m-q-1}$. Although we concentrate in our applications on piecewise constant splines, extending the methods to handle other basis systems (even wavelets or kernel basis) is straightforward.

3.2.1 Mean-curve clustering

In the sequel, let y_{ijt} be the progesterone level on the t-th day of the j-th menstrual cycle of woman i, with $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ and $t = 1, \ldots, T_{ij}$. We model

the expected evolution of progesterone in time as a linear combination of piecewise constant B-spline bases. In order to allow for outliers, which are commonly present in this type of data, we use a Student t distribution with a small number ν of degrees of freedom to model the error distribution of the observations. Specifically, we let

$$y_{ijt} \sim \mathsf{N}\left(\sum_{k=1}^{m-q-1} b_k(x_{ijk})\theta_{ijk}, \frac{\sigma^2}{\lambda_{ijt}}\right)$$

$$\lambda_{ijt} \sim \mathsf{Gam}(\nu/2, \nu/2)$$

where, for computational purposes, we have used the well known representation of the t distribution as a scale-mixture of normals. In principle, we allow a different set of basis coefficients for each cycle of each woman. In order to borrow information across curves within each woman, we use a random effects model and assume that the coefficients for each cycle arise from woman-specific distribution

$$\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \ldots, \theta_{ij,m-q-1}) \sim \mathsf{N}(\boldsymbol{\theta}_i^*, \sigma^2\boldsymbol{\Omega})$$

By taking $\boldsymbol{\Omega} \to 0$ we force a unique set of coefficients $\boldsymbol{\theta}_i^*$ for all cycles within a given woman. Larger values of $\boldsymbol{\Omega}$ allow for increasingly larger deviations of individual cycles from this average. In order to borrow information across women, we need a hyperprior for the woman specific parameters $\{\boldsymbol{\theta}_i\}_{i=1}^I$. Instead of the more standard normal prior, we use a Dirichlet process centered around a normal,

$$\boldsymbol{\theta}_i^* \sim G$$

$$G \sim \mathsf{DP}(\alpha H)$$

$$H = \mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$$

As a byproduct of the DP prior, we obtain clusters of women with similar "average" curves, given by the $\boldsymbol{\theta}_i^*$ parameters. Note that our specification does not require for all the curves to be observed at exactly the same times. By borrowing information across cycles and women, our model implicitly imputes any missing values.

The model is completed by assigning priors for the hyperparameters. The observational variance is given a conditionally conjugate inverse gamma prior, $\sigma^2 \sim \mathsf{IGam}(a_\sigma, b_\sigma)$. The degrees of freedom ν are given a fixed value, which we take as 4 in our applications. For the random effect variances we take inverse-Wishart priors.

$$\boldsymbol{\Omega} \sim \mathsf{IW}(\gamma_\Omega, \boldsymbol{\Omega}_0)$$

$$\boldsymbol{\Sigma} \sim \mathsf{IW}(\gamma_\Sigma, \boldsymbol{\Sigma}_0)$$

Finally, the concentration parameter α is given a gamma prior $\alpha \sim \mathsf{Gam}(a_\alpha, b_\alpha)$.

3.2.2 Multilevel clustering

Our model for multilevel clustering is similar to the one described in section 3.2.1 above. As before the top level of the hierarchy is given by

$$y_{ijt} \sim \mathsf{N}\left(\sum_{k=1}^{m-q-1} b_k(x_{ijk})\theta_{ijk}, \frac{\sigma^2}{\lambda_{ijt}}\right)$$

$$\lambda_{ijt} \sim \mathsf{Gam}(\nu/2, \nu/2)$$

$$\boldsymbol{\theta}_{ij} \sim G_i$$

In section 3.2.1 we took the collection $\{G_i\}_{i=1}^I$ to be made of (conditionally independent) normal distributions $G_i = \mathsf{N}(\boldsymbol{\theta}_i^*, \sigma^2\boldsymbol{\Omega})$. In this section we use a nested Dirichlet process instead, with

$$G_i \sim \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*}$$

$$G_k^* \sim \sum_{l=1}^{\infty} w_{lk} \delta_{\boldsymbol{\theta}_{lk}^*}$$

$$\boldsymbol{\theta}_{kl}^* \sim H$$

In words, our model generates groups of women, assigning to each of those groups a common random effect distribution G_k^*. In turn, each one of these distributions is

a mixture of different curve shapes, given by the atoms $\{\boldsymbol{\theta}^*_{lk}\}_{l=1}^{\infty}$. All curve shapes are assigned a common prior distribution H.

As with other models based on the nDP construction, interesting special cases can be obtained by considering the limit of the precision parameters. For example, letting $\beta \to 0$ induces a model where all menstrual cycles within a woman are assumed to have the same profile, and women are clustered according to their mean cycle. Such a model is equivalent to taking $\boldsymbol{\Omega} \to 0$ in section 3.2.1. On the other hand, by letting $\alpha \to \infty$, we obtain a model where all women are treated as different and menstrual cycles are clustered within each women. Therefore, information is borrowed across the menstrual cycles of each women, but not across women.

As before, we employ an inverse gamma prior for the observational variance, $\sigma^2 \sim \mathsf{IGam}(a_\sigma, b_\sigma)$. The stick-breaking weights are constructed as,

$$\pi_k = u_k \prod_{s<k}(1-u_s) \qquad\qquad u_k \sim \mathsf{beta}(1,\alpha)$$

$$w_{lk} = v_{lk} \prod_{s<l}(1-v_{sk}) \qquad\qquad v_{lk} \sim \mathsf{beta}(1,\beta)$$

and the baseline measure H is set as a $\mathsf{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \sim \mathsf{IW}(\gamma_\Sigma, \boldsymbol{\Sigma}_0)$. Finally, gamma priors are used for both concentration parameters,

$$\alpha \sim \mathsf{Gam}(a_\alpha, b_\alpha)$$

$$\beta \sim \mathsf{Gam}(a_\beta, b_\beta)$$

3.3 Inference

We only describe a computational implementation for the model in section 3.2.2, which is based on a finite truncation of the nested Dirichlet process. Similarly to the mean curve clustering model, we introduce latent variables ζ_j and ξ_{ij} such that $\zeta_j = k$ if $H_j = H^*_k$ and $\xi_{ij} = l$ if $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}^*_{lk}$. Once adequate starting values for the parameters have been chosen, computation proceeds through the following steps:

1. Sample the bottom-level indicators ζ_i for $i = 1, \ldots, I$ from a multinomial distribution with probabilities

$$\mathbb{P}(\zeta_i = k|\cdots) = q_k^i \propto w_k^* \prod_{j=1}^{n_i} \sum_{l=1}^{L} \pi_{lk} \left[\prod_{s=1}^{T_{ij}} \phi(y_{ijs}|\mathbf{b}(x_{ijs})'\boldsymbol{\theta}_{lk}^*, \sigma^2/\lambda_{ijs}) \right],$$

$$k = 1, \ldots, K.$$

2. Sample the top-level indicators ξ_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$ from another multinomial distribution with probabilities

$$\mathbb{P}(\xi_{ij} = l|\cdots) = b_{ij}^l \propto \pi_{l,\zeta_i}^* \prod_{s=1}^{T_{ij}} \phi(y_{ijs}|\mathbf{b}(x_{ijs})'\boldsymbol{\theta}_{l,\zeta_i}^*, \sigma^2/\lambda_{ijs}), \quad l = 1, \ldots, L.$$

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

$$(u_k^*|\cdots) \sim \mathsf{beta}\left(1 + r_k, \alpha + \sum_{s=k+1}^{K} r_s\right), \quad k = 1, \ldots, K-1, \quad u_K^* = 1,$$

where r_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

$$(v_{lk}^*|\cdots) \sim \mathsf{beta}\left(1 + r_{lk}, \beta + \sum_{s=l+1}^{L} r_{ls}\right), \quad l = 1, \ldots, L-1, \quad v_{Lk}^* = 1,$$

where r_{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

$$(\boldsymbol{\theta}_{lk}^*|\cdots) \sim \mathsf{N}(\hat{\boldsymbol{\theta}}_{lk}, \sigma^2\hat{\boldsymbol{\Gamma}}_{lk}),$$

where

$$\hat{\mathbf{\Gamma}}_{lk} = \left[\mathbf{\Sigma}^{-1} + \sum_{\{(i,j,s):\zeta_j=k,\xi_{ij}=l\}} \lambda_{ijs} \mathbf{b}(x_{ijs}) \mathbf{b}(x_{ijs})' \right]^{-1}$$

$$\hat{\boldsymbol{\theta}}_{lk} = \hat{\mathbf{\Gamma}}_{lk} \left[\sum_{\{(i,j,s):\zeta_j=k,\xi_{ij}=l\}} \lambda_{ijs} y_{ijs} \mathbf{b}(x_{ijs}) \right]$$

Note that, if no observation is assigned to a specific cluster, then the parameters are drawn from the conditional prior distribution (baseline measure) $\mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$.

6. Sample the variance of the observations from $\sigma^2 | \cdots \sim \mathsf{IGam}(\hat{a}_\sigma, \hat{b}_\sigma)$ where

$$\hat{a}_\sigma = a_\sigma + \frac{\sum_{(i,j)} T_{ij}}{2} + \frac{pKL}{2}$$

$$\hat{b}_\sigma = b_\sigma + \frac{\sum_{(i,j,s)} \lambda_{ijs} (y_{ijs} - \mathbf{b}(x_{ijs})' \boldsymbol{\theta}_{\zeta_i,\xi_{ij}})^2}{2} + \frac{\sum_{(l,k)} \boldsymbol{\theta}'_{lk} \mathbf{\Sigma}^{-1} \boldsymbol{\theta}_{lk}}{2}$$

7. Sample the auxiliary variables λ_{ijt} from

$$\lambda_{ijt} \sim \mathsf{Gam}\left(\frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(y_{ijt} - \mathbf{b}(x_{ijt})' \boldsymbol{\theta}_{\zeta_i,\xi_{ij}})^2}{2\sigma^2} \right)$$

8. Sample the variance of the baseline measure from $(\mathbf{\Sigma} | \cdots) \sim \mathsf{IW}(\hat{\gamma}_\Sigma, \hat{S})$ where

$$\hat{\gamma}_\Sigma = \gamma_\Sigma + \sum_{\{l,k:r_{lk} \neq 0\}} 1$$

$$\hat{\gamma}_\Sigma \hat{S} = \gamma_\Sigma S + \frac{1}{\sigma^2} \sum_{\{l,k:r_{lk} \neq 0\}} \boldsymbol{\theta}_{lk} \boldsymbol{\theta}'_{lk}$$

and $\sum_{\{l,k:r_{lk} \neq 0\}} 1$ is the number of non-empty components.

9. Sample the concentration parameters α and β from

$$(\alpha | \cdots) \sim \mathsf{Gam}\left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1} \log(1 - u^*_k) \right)$$

$$(\beta | \cdots) \sim \mathsf{Gam}\left(a_\beta + K(L-1), b_\beta - \sum_{l=1}^{L-1} \sum_{k=1}^{K} \log(1 - v^*_{lk}) \right)$$

3.4 An application to the Early Pregnancy Study

Progesterone plays a crucial role in controlling different aspects of the reproductive function in women, from fertilization to early development and implantation. Therefore, understanding the influence of different variables on the hormonal profile along the menstrual cycle is critical to identify risk factors for infertility and early pregnancy loss. The main difficulties are the functional nature of the outcome variable and the wide variability in hormonal profiles, not only across women, but also among different menstrual cycles for the same women.

Our data, extracted from the Early Pregnancy Study (Wilcox *et al.*, 1998), consists of daily creatinine-corrected concentrations of pregnanediol-3-glucuronide (PdG) for 60 women along multiple menstrual cycles, measured in micrograms per milligram of creatinine (μg/ml Cr). We focus on a 13 days periods extending from 10 days before ovulation to 2 days after ovulation. We considered only non-conceptive cycles and women with at least four cycles in record. Therefore, the number of curves per woman varies between 4 and 9. Available woman-specific variables include age, weight, last birth-control method, number of previous pregnancies, age of first menses, regular length of menses, smoking habits and marijuana intake during study.

We use the algorithm described on section 3.3 with $K = L = 60$ to fit a piecewise constant spline model. We use $m = 13$ nodes, corresponding to each of the days considered in the study. This yields basis functions of the form

$$b_k(x) = \begin{cases} 1 & k - 10 \leq x < k - 9 \\ 0 & \text{otherwise} \end{cases} \qquad\qquad k = 0, \ldots, 12$$

Prior distributions were set as follows. For the observational variance, we used a vague but proper prior $\sigma^2 \sim \mathsf{Gam}(0.001, 0.001)$. For the concentration parameters, we used proper priors $\alpha \sim \mathsf{Gam}(3, 3)$ and $\beta \sim \mathsf{Gam}(3, 3)$, such that $\mathbb{P}(\alpha > 3) = \mathbb{P}(\beta > 3) \approx 0.006$. Finally the hyper-prior for the variance-covariance matrix of the spline

coefficients is set such that $\gamma_\Sigma = 3$ and $\mathbf{\Sigma} = 4 \cdot 13 \cdot \mathbf{I}_{13}$, where $\mathbf{I}_d$ represents a $d \times d$ identity matrix.

All inferences are based on 80,000 samples obtained after a burn-in period of 10,000 iterations. Results seem robust to reasonable changes in the parameter values. However, mixing can be an issue, as the algorithm seems sensitive to the starting point. This is probably due to the small number of cycles available for each woman. In order to improve mixing, we first run the mean-based clustering model obtained by letting $\alpha \to 0$ (which is computationally equivalent to fix $\xi_{ij} = \xi_i^*$ for some set $\{\xi_i^*\}_{i=1}^I$). The results from this run are then used as to generate random starting configuration for our algorithm. Although posterior clusters obtained from both models are different (as we illustrate later), mean-based clustering seems to provide a reasonable starting point for distribution-based clustering.

Figure 3.2 shows a heatmap of the posterior average incidence matrix generated by our sample. Entry (i, j) of this matrix corresponds to the marginal posterior probability of women i and j being assigned to the same cluster. There is evidence of nine groups (labeled 0 to 8 in the figure) of varying size. Clusters 1, 3 and 4 are the largest and best separated clusters, comprising 45 out of 60 women. Clusters 0 and 2 are not clearly separated from the other seven groups, and probably correspond to women with cycles that conform to more than one of the big clusters. Clusters 5 to 8 are small clusters of one or two women, and correspond to patients with consistently anomalous hormonal profiles. Note that, as was suggested in the introduction, women 3 and 43 are clearly clustered together. In contrast, in our preliminary run (not shown) using mean-based clustering, woman 43 was tightly clustered with woman 36.

Figure 3.3 shows reconstructed profiles for some representative women on each of the groups. Since piecewise constant splines were used, there is very little smoothing.

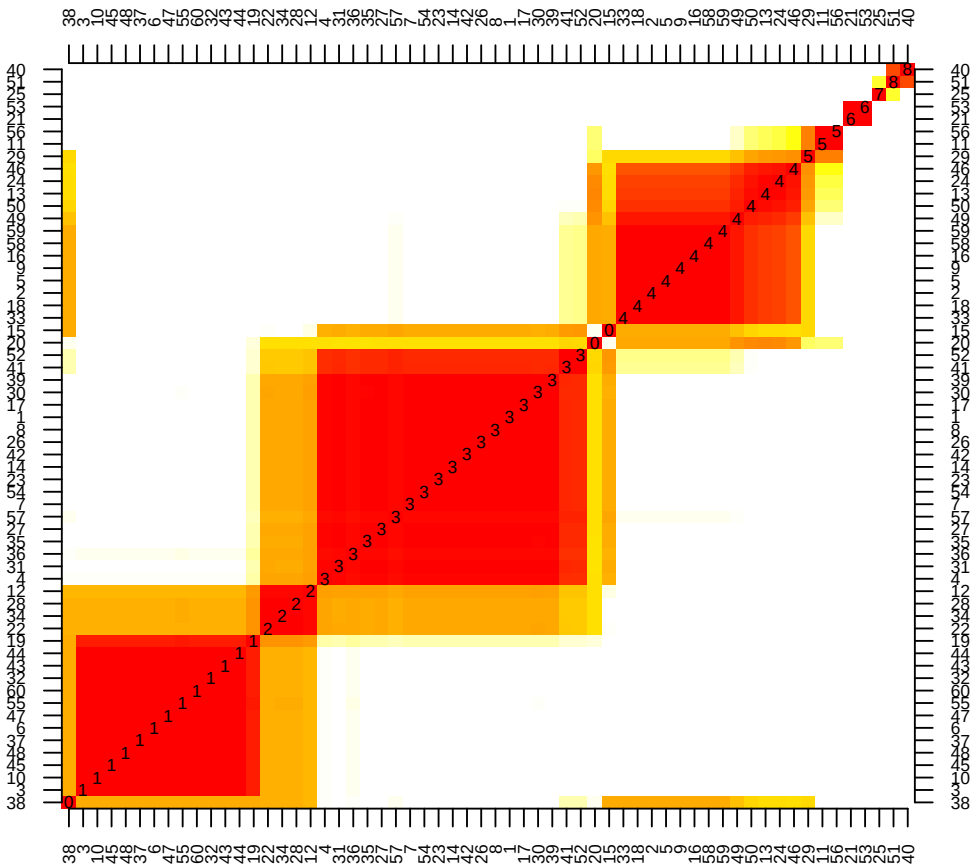


Figure 3.2: Average incidence matrix, illustrating probabilities of joint pairwise classification for the 60 women in the EPS. White corresponds to zero probability, while red corresponds to 1. Numbered labels correspond to clusters of women.

Most profiles are relatively flat before ovulation, when hormone levels start to increase. Also, profiles tend to be relatively consistent for any single woman. However, we can see some outliers, typically corresponding to elevated post-ovulation levels and/or early increases in the hormone levels.

The reconstructed profiles also provide insight into the characteristics of the groups. Cluster 1 corresponds to women with very low hormonal levels, even after ovulation. These group presents few outliers, and those present are characterized by a slightly larger increase after ovulation. Group 3 has a slightly higher baseline level and also present a larger increase in PdG in the luteal phase, reaching a concentration 1.35 μg/ml Cr. Outlier cycles in this groups are characterized by earlier increases in PdG during the follicular phase, sometimes as early as 5 days before ovulation. Group 4 shows the highest baseline hormonal level and the larger increase after ovulation, and outliers include both cycles with a higher-than-normal luteal-phase hormonal levels and follicular-phase increases in hormones. Group 2 is a transition cluster, with curves that resemble those in both cluster 1 and 3. As we mentioned above, the other 4 groups comprise women that are characterized by very high hormonal levels. For example, woman 53 consistently presents a PdG concentration over 2 μg/ml Cr, while patient 29 has final PdG values of up to 5 μg/ml Cr.

3.5 Discussion

We have developed a method for multilevel clustering of curves. For small sample sizes like the ones in our hormone example, we think of this model as generating clusters of women based on the most frequent (modal) profile instead of the mean profile, leading to more reasonable results than those obtained from standard methods. The clustering structure generated by the model can be used to generate hypotheses about

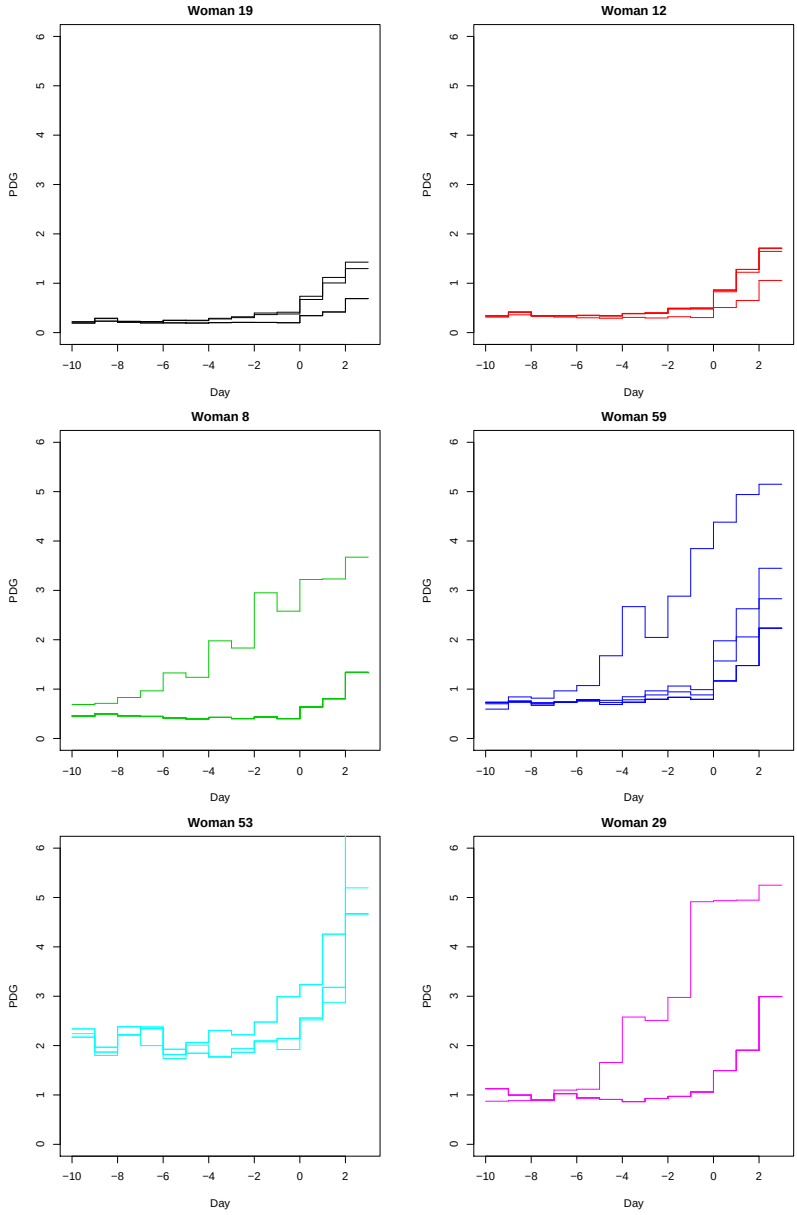


Figure 3.3: Reconstructed profiles for some representative women. Patient 19 was chosen for cluster 1, patient 12 for cluster 2, patient 8 for cluster 3, patient 59 for cluster 4, patient 53 for cluster 5 and patient 29 for cluster 6.

mechanisms underlying hormonal trajectories. The groups generated by the model can also be used as an explanatory variable for outcomes like EPL or infertility, providing insight into the mechanisms behind these negative outcomes.

In so far, our results have been based on piecewise constant splines. However, the degree of smoothness in the functional estimates is likely to play a role in the clustering structure. Therefore, we need to compare our results against those obtained with other basis functions, such as natural cubic splines. We also need to consider a larger sample that includes more fertile women. From a statistical perspective, the main issue with this larger sample size is that fertile women tend to get pregnant earlier in the study, therefore providing a smaller number of curves. This translates into much more uncertainty on the clustering structure.

It would be interesting to generalize the model to include an additive structure that allows flexible incorporation of predictors. In this setting, the nDP would allow clustering of women based on how the predictors vary in their effects across cycles. Another minor extension of the model is to place a prior on the number of degrees of freedom ν, similarly to Gottardo *et al.* (2006). Although this is an appealing direction to explore, some care must be exercised as there might be identifiability issues. Indeed, lighter tails for the observational errors will typically mean a larger number of distinct curves for each woman, which in turn will usually imply a larger number of clusters. Therefore, it might be difficult in practice to separate the effect of the degrees of freedom ν form the concentration parameters α and β

Chapter 4

Dynamic nonparametric linear models

4.1 Introduction

One of the main constraints associated with classical time series analysis is the parametric assumptions involved in the analysis. Even if the evolution process is modeled in a flexible or nonparametric way, observational and evolution noise are typically assumed to follow some parametric distribution. This means that inferences end up being restricted to the moments of the assumed distributions, and changes not captured by those moments are overlooked by the model. Besides, in many applications, the natural measurement object is the distribution itself, which can potentially present skewness and multimodality as part of its features. For example in molecular epidemiology studies, one focus is the profile of changes in the distribution of DNA damage across time. Also, in option pricing, interest lies on the estimation of the risk-neutral distribution underlying the observed process (Melick and Thomas, 1997; Panigirtzoglou and Skiadopoulos, 2004).

Nonparametric methods have proven also useful in the valuation of options and derivatives. Since the seminal papers of Black and Scholes (1973) and Merton (1973), the option pricing literature has concerned itself with relaxing the key simplifying as-

sumptions such as constant volatility, zero transactions costs and a flat yield curve, as well as with using stochastic processes flexible enough to handle returns exhibiting fat-tails and skewness. Those modeling relaxations include working with kernel methods to price options (Ait-Sahalia, 1996; Ait-Sahalia and Duarte, 2003), as well as extracting implied probability densities of the S&P 500 (Panigirtzoglou and Skiadopoulos, 2004). In the analysis of credit rating scores, Huang *et al.* (2004) compared credit rating prediction performance between backpropagation neural network (BNN) and support vector machines (SVM), obtaining around 80% of accuracy for both methods in the context of corporate credit rating analysis. In the context of corporate bond credit rating, Chaveesuk *et al.* (1999) explore three of the most well known supervised neural network paradigms-backpropagation, radial basis function and learning vector quantization-for the task of rating US corporate bonds.

In this paper we develop statistical methods appropriate to estimate and predict densities that evolve in discrete time. We are particularly interested in models where computationally efficient algorithms can be developed. Our models use countably infinite mixtures of Gaussian distributions to represent the unknown density at each time point. These methods can be conceived as an extension of the Dirichlet Process Mixture model (Antoniak, 1974; Escobar and West, 1995) to collections of distributions that evolve in discrete time. It has been shown that, under mild conditions, these infinites mixtures have full support, in the sense of being dense on the space of absolutely continuous distributions (Lo, 1984). Dependence is built into the mixing distribution by allowing the atoms to evolve dynamically as linear state-space models. Indeed, the models we present can also be regarded as an extension of the Gaussian Dynamic Linear Models (DLMs) of West and Harrison (1997), which are Bayesian versions of the popular Kalman Filter. The DLMs represent a very flexible class of models with well-known properties, and have been successfully applied on a number

of different areas, including econometrics, engineering and climatology (Kim, 1994; Pesaran *et al.*, 1995; West, 1995; Lamon *et al.*, 1998; West *et al.*, 1999).

Although physical and economic phenomenon typically occur in continuous time, discrete-time models provide good approximations as long as a fine enough scale is used. In addition, most real life phenomena are actually observed in discrete and equispaced intervals, making discrete-time models a natural tool for empirical analysis. This paper focuses on financial applications, but the class of dynamic Dependent Dirichlet Processes we present in this paper has multiple applications in areas like engineering (dynamic spectra estimation), climatology (modeling rainfall distributions) and biostatistics/epidemiology (genetic epidemiology studies, dynamic random effect models).

Most of the classical literature on the Dirichlet Process focuses on exchangeable samples. However, recent work has started to develop methods for dependent distributions, either by forming convex combinations of independent processes (Müller *et al.*, 2004; Dunson *et al.*, 2007b; 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; DeIorio *et al.*, 2004; Gelfand *et al.*, 2005; Griffin and Steel, 2006b). Particularly relevant for this paper are the works of Griffin and Steel (2006b), who induce dependence through permutations of otherwise fixed atoms; Griffin and Steel (2006a), who construct the mixing distribution at a new time point as a linear combination of the mixing distribution at the previous time plus an innovation; and Tang and Ghosal (2006), who are concerned with estimating the conditional distribution of a general autoregressive process. Our approach, while similar to Griffin and Steel (2006b), allows the atoms to evolve in time, in the same way as state-space model. This formulation provides a more intuitive interpretation in finance applications, simplifies the implementation of the model

and casts it as a direct extension of widely used models for time series analysis.

This chapter is organized as follows: Section 4.2 describes a model that uses location-scale mixtures of normals to construct time dependent processes, proposing a Markov Chain Monte Carlo (MCMC) scheme and some interesting special cases of the general formulation. Section 4.3 considers an alternative formulation that uses location mixtures of DLMs with time varying variances. Then, section 4.4 describes an application of these models to the estimation of risk neutral distributions implied by option prices. Finally, we close in Section 4.5 with a brief discussion of the models and some interesting future research points.

4.2 Location-scale mixtures of time dependent processes

4.2.1 Definition and properties

Recall from chapter 1 that, given a set D, we can replace the baseline distribution underlying Sethuraman's stick-breaking construction with a stochastic process $\{\boldsymbol{\eta}(t) \; \forall \; t \in D\}$ and define

$$K_t(\cdot) = \sum_{l=1}^{\infty} w_l^*(t)\delta_{\boldsymbol{\eta}_l^*(t)}(\cdot) \tag{4.1}$$

where $\boldsymbol{\eta}_l^*(t) \sim \boldsymbol{\eta}(t)$ and $w_l^*(t) = z_l^*(t)\prod_{s=1}^{l-1}(1 - z_s^*(t))$ with $z^*(t) \sim \mathsf{beta}(1, \alpha(t))$ for all $t \in D$. The resulting stochastic process is called a Dependent Dirichlet Process (DDP) (MacEachern, 2000) and defines a distribution on the collection of random distributions on the space D, such that every K_t is marginally a Dirichlet Process. In the sequel, we consider mixtures of Gaussian distributions by a discrete-time DDP. Therefore, we take $D = \mathbb{N}$ and assume our underlying stochastic process to be a general random walk. For computational reasons, we focus on "single p" DDP models

where $z_l^*(t) = z_l^* \sim \mathsf{beta}(1, \alpha)$, independently of t. More specifically, letting y_{it} be the i-th observation ($i = 1, \ldots, n_t$) obtained at time $t = 1, \ldots, T$, our model reduces to:

$$
\begin{aligned}
y_{it} &\sim \int \mathsf{N}\left(\mathbf{F}_{it}'\boldsymbol{\theta}_t, \sigma^2\right) K_t(d\boldsymbol{\theta}_t, d\sigma^2) \qquad K_t = \sum_{l=1}^{\infty} w_l^* \delta_{\left(\theta_{lt}^*, \sigma_l^{2*}\right)} \\
w_l^* &= z_l^* \prod_{s=1}^{l-1}(1 - z_s^*) \qquad\qquad z_l^* \sim \mathsf{beta}(1, \alpha) \\
\boldsymbol{\theta}_{lt}^* | \boldsymbol{\theta}_{l,t-1}^*, \phi_l^* &\sim \mathsf{N}(\mathbf{G}_t \boldsymbol{\theta}_{l,t-1}^*, \sigma_l^{2*} \mathbf{W}_t) \qquad \boldsymbol{\theta}_{0l}^* | \phi_l^* \sim \mathsf{N}(\mathbf{m}_0, \sigma_l^{2*} \mathbf{C}_0) \\
\sigma_l^{2*} &\sim \mathsf{Gam}(s_0, s_0 S_0)
\end{aligned}
\tag{4.2}
$$

where $\mathsf{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2, while $\mathsf{IGam}(a, b)$ denotes an Inverse-Gamma distribution with mean a/b and a degrees of freedom (see appendix A).

Our model assumes that the distribution for any observation y_{it} can be written as a mixture of normal components with means $\mathbf{F}_{it}\boldsymbol{\theta}_{lt}$ and variance $\sigma_l^2 = 1/\phi_l$, for some known matrix $\mathbf{F}_{it}$. The parameters defining the means of these distributions are allowed to move, with the evolution matrix $\mathbf{G}_t$ and the innovation variance $\mathbf{W}_t$ controlling the direction and magnitude of these changes. This formulation is rather general, and by appropriately choosing the structural matrices $\mathbf{F}_t$, $\mathbf{G}_t$ and $\mathbf{W}_t$ our model can easily accommodate patterns like trends, periodicities, dynamic regressions and even autoregressive or moving average models for densities. The variances of the mixture components are assumed to be constant in time, but allowed to be change across component; and the weights associated each distribution are also estimated from the data, and assumed to be constant in time. Therefore, the estimates of the model can be interpreted as kernel density estimates with adaptive bandwidths and time varying kernels.

The dynamic DDP can be interpreted as a multiprocess type II model (West and Harrison, 1997). In this multiprocess interpretation, an infinite number of DLM mod-

els with the same evolutionary structure but a different set of state parameters are available at every point in time. These components represent different evolutionary paths consistent with a common underlying regime described by $\mathbf{G}_t$ and $\mathbf{W}_t$, and where each observation is allocated to one of these components with constant probability. Therefore, predictions using expected posterior means can be interpreted in this model as weighted averages of the predictions obtained under an infinite number of DLMs.

As argued by MacEachern (2000), the model can be reexpressed as a Mixture of Dirichlet processes

$$y_{it} \sim \int \mathsf{N}\left(\mathbf{F}_{it}'\boldsymbol{\theta}_t, \sigma^2\right) K\left(d\boldsymbol{\Theta}, d\sigma^2\right) \qquad\qquad K \sim DP(\alpha K_0)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_T')'$ and K_0 is the joint distribution of $(\boldsymbol{\Theta}, \phi)$ induced by the evolution equations described above, which reduces to a multivariate normal-gamma where

$$\mathbb{E}(\boldsymbol{\theta}_t|\phi) = \left[\prod_{r=1}^{t} \mathbf{G}_{t-r+1}\right] \mathbf{m}_0$$

$$\mathbb{V}(\boldsymbol{\theta}_t|\phi) = \sigma^2 \left\{ \left[\prod_{r=1}^{t} \mathbf{G}_{t-r+1}\right] \mathbf{C}_0 \left\{\prod_{r=1}^{t} \mathbf{G}_{t-r+1}\right]' + \right.$$

$$\left. \sum_{r=1}^{t-1} \left[\prod_{s=1}^{t-r} \mathbf{G}_{t-s+1}\right] \mathbf{W}_r \left[\prod_{s=1}^{t-r} \mathbf{G}_{t-s+1}\right]' + \mathbf{W}_t \right\}$$

$$\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{t+k}, \boldsymbol{\theta}_t|\phi) = \left[\prod_{r=1}^{k} \mathbf{G}_{t+k-r+1}\right] \mathbb{V}(\boldsymbol{\theta}_t|\phi)$$

$$\mathbb{E}(\phi) = \frac{s_0}{s_0 - 1} S_0$$

$$\mathbb{V}(\phi) = \frac{s_0^2}{(s_0-1)^2(s_0-2)} S_0$$

This representation as a DP mixture will be exploited in section 4.2.2 to develop

efficient and simple computational strategies to fit these models. Although the "single p" structure in 4.2 suggests a restrictive model where the same number of components is used to represent every distribution (in principle, at least the largest one needed at any time point), the model is indeed flexible and dense on the space of distributions on D, as argued in MacEachern (2000). Note that a good approximation to a lower number of components can be achieved at any time point by assigning different components similar values of their parameters. Indeed, if $(\boldsymbol{\theta}_{it}, \sigma_i) \approx (\boldsymbol{\theta}_{jt}, \sigma_j)$ then $w_{it}\delta_{(\boldsymbol{\theta}_{it},\sigma_i)} + w_{jt}\delta_{(\boldsymbol{\theta}_{jt},\sigma_j)} \approx (w_{it} + w_{jt})\delta_{(\boldsymbol{\theta}_{it},\sigma_i)}$. Therefore, by having components with similar parameters but that are not allocated to any observation at certain time points, we can approximate variable weights. Therefore, the price to pay for the constant-weight assumption is, in general, a slightly larger number of atoms being used.

Note that our dynamic DDP encompasses a number of other models as limiting cases. On one side by letting $\mathbf{W}_t = \mathbf{0}\ \forall\ t$ we have $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0\ \forall\ t$ and thus $K_t = K_r\ \forall\ t, r$, which is the set up of Escobar and West (1995). On the other hand, by letting $\alpha \to 0$, we revert to the class of parametric DLMs with replicates as discussed in West and Harrison (1997). Note, however, that although increasing the value of $\mathbf{W}_t$ reduces the dependence among distributions, letting $\mathbf{W}_t \to \infty\ \forall\ t$ does not yield independent density estimates at each time point, but an improper distribution for θ_t at all times $t \geq 1$. Indeed, it is known that obtaining independent distributions from a single p DDP is not possible (MacEachern, 2000). Although this is a somewhat unappealing characteristic of the model, it will not be an issue for most practical applications.

It is straightforward to obtain the a priori covariance structure induced on the observations by the process (see appendix G),

$$\mathbb{C}\text{ov}(y_{i',t+k}, y_{i,t}) = \frac{\mathbf{F}'_{t+k}\left[\prod_{s=1}^{k}\mathbf{G}_{t+k-s+1}\right]\mathbb{V}(\boldsymbol{\theta}_t)\mathbf{F}_t}{1+\alpha}\frac{s_0}{s_0-1}S_0 \ \ \forall\ \ i \neq i', s_0 > 1$$

Note the similarities with the replicated, Gaussian DLM model, where the covariance reduces to $\mathbf{F}'_{t+k}\left[\prod_{s=1}^{k}\mathbf{G}_{t+s}\right]\mathbb{C}\text{ov}(\boldsymbol{\theta}_t)\mathbf{F}_t s_0 S_0/(s_0-1)$. Therefore, the covariance between observations at different time points under the discrete time DDP is strictly smaller than under the parametric alternative, which is due to the added uncertainty in the model specification. Also, since this is the induced covariance after marginalizing over the unknown collection $\{K_1,\dots K_T\}$, the replicates at a given time point are not independent unless $\alpha = 0$ (which corresponds to a single DLM model).

4.2.2 Inference

Inferences on the dynamic DDP can be performed using the same computational techniques employed for DP mixtures models (see, for example, Bush and MacEachern (1996); MacEachern and Müller (1998); Neal (2000); Ishwaran and James (2001)). In what follows, we describe an application of the algorithm of MacEachern and Müller (1998). For this purpose, a reparametrization of the model is helpful: let L be the current number of components that have observations allocated to them, n^*_{lt} be the number of observations in time t assigned to group l, $n_l = \sum_t n_{lt}$, $\{\boldsymbol{\Theta}^*_1,\dots,\boldsymbol{\Theta}^*_L\}$ be the current estimated values for those paths and $\xi_{it} = l$ iif $\boldsymbol{\Theta}_{it} = \boldsymbol{\Theta}^*_{lt}$. Given values for the structural parameters $\mathbf{F}_{it}$, $\mathbf{G}_{it}$ and $\mathbf{W}_{it}$ and after initialization of the parameters, an MCMC sampler alternates through the following steps:

1. Generate $\boldsymbol{\Theta}^*_l, \sigma^{*2}_l|\{\mathbf{y}|\xi_{it} = k,\}\cdots$ using the following Forward Filtering / Backward Sampling (FFBS) algorithm (Carter and Kohn, 1994; Fruhwirth-Schnatter, 1994)

(a) Forward filter using the recursions

$$
\begin{aligned}
\mathbf{m}_{lt} &= \begin{cases} \mathbf{a}_{lt} + \mathbf{A}_{lt} e_{lt} & \text{if } n_{lt} > 0 \\ \mathbf{a}_{lt} & \text{if } n_{lt} = 0 \end{cases} \\
\mathbf{C}_{lt} &= \begin{cases} \mathbf{R}_{lt} - \mathbf{A}_{lt}\mathbf{Q}_{lt}\mathbf{A}'_{lt} & \text{if } n_{lt} > 0 \\ \mathbf{R}_{lt} & \text{if } n_{lt} = 0 \end{cases} \\
s_{lt} &= s_{l,t-1} + n_{lt} \\
s_{lt}S_{lt} &= \begin{cases} s_{l,t-1}S_{l,t-1} + \mathbf{e}'_{lt}\mathbf{Q}_{lt}^{-1}\mathbf{e}_{lt} & \text{if } n_{lt} > 0 \\ s_{l,t-1}S_{l,t-1} & \text{if } n_{lt} = 0 \end{cases} \\
\mathbf{A}_{lt} &= \mathbf{R}_{lt}\mathbf{F}^{*}_{lt}\mathbf{Q}_{lt}^{-1} \\
\mathbf{e}_{lt} &= \mathbf{y}_{lt} - \mathbf{f}_{lt} \\
\mathbf{f}_{lt} &= \mathbf{F}^{*}_{lt}\mathbf{a}_{lt} \\
\mathbf{Q}_{lt} &= \mathbf{F}^{*}_{lt}\mathbf{R}_{lt}\mathbf{F}^{*'}_{lt} + \mathbf{I} \\
\mathbf{a}_{lt} &= \mathbf{G}_{lt}\mathbf{m}_{l,t-1} \\
\mathbf{R}_{lt} &= \mathbf{G}_{lt}\mathbf{C}_{l,t-1}\mathbf{G}'_{lt} + \mathbf{W}_{lt}
\end{aligned}
$$

where $\mathbf{y}_{lt}$ is made of all observations assigned to group l at time t, $\mathbf{F}^{*}_{lt}$ is a matrix whose rows are the corresponding $\mathbf{F}_{it}$ vectors and $\mathbf{I}$ is the identity matrix.

(b) Sample $\sigma_l^2|\mathbf{y}_l, \cdots$ from $\mathsf{IGam}(s_{lT}/2, s_{lT}S_{lT}/2)$.

(c) Sample $\boldsymbol{\theta}_{lT}|\sigma_l^2, \mathbf{y}_l, \cdots$ from $\mathsf{N}(\mathbf{m}_{lT}, \mathbf{C}_{lT})$. Then sample $\boldsymbol{\theta}_{lt}|\boldsymbol{\theta}_{l,t+1}\sigma_l^2, \mathbf{y}_l, \cdots$ recursively from $\mathsf{N}(\mathbf{d}_{lt}, \mathbf{D}_{lt})$ where

$$
\begin{aligned}
\mathbf{d}_{lt} &= \mathbf{m}_{lt} + \mathbf{B}_{lt}\left(\boldsymbol{\theta}_{l,t+1} - \mathbf{a}_{l,t+1}\right) \\
\mathbf{D}_{lt} &= \mathbf{C}_{lt} - \mathbf{B}_{lt}\mathbf{R}_{l,t+1}\mathbf{B}'_{lt} \\
\mathbf{B}_{lt} &= \mathbf{C}_{lt}\mathbf{G}_{t+1}R^{-1}_{l,t+1}
\end{aligned}
$$

2. Sample $\xi_{it}|\mathbf{y}, \boldsymbol{\xi}^-, \cdots$ from a multinomial distribution with probabilities:

$$\begin{aligned} q_l &= n_l^- p(y_{it}|\mathbf{y}^-, \boldsymbol{\xi}^-) \\ &= n_l^- \mathsf{T}_{s_{lT}^-}\left(y_{it}|\mathbf{F}_{it}'\mathbf{h}_{lt}^-, S_{lT}^-(1+\mathbf{F}_{it}'\mathbf{H}_{lt}^-\mathbf{F}_{it})\right) \\ q_0 &= \alpha p(y_{it}|S_0) \\ &= \alpha \mathsf{T}_{s_0}\left(y_{it}|\mathbf{F}_{it}'\mathbf{h}_{t0}, S_0(1+\mathbf{F}_{it}'\mathbf{H}_{0t}\mathbf{F}_{it})\right) \end{aligned}$$

where the superscript indicates removal of observation (i,t) from the sample, q_l for $l = 1, \ldots, L^-$ is the probability of allocation observation (i,t) to cluster l, q_0 is the probability of allocating the observation to a new cluster, $\mathbf{h}_{lT} = \mathbf{m}_{lT}$, $\mathbf{H}_{lT} = \mathbf{C}_{lT}$ and

$$\begin{aligned} \mathbf{h}_{lt} &= \mathbf{m}_{lt} + \mathbf{B}_{lt}\left(\mathbf{h}_{l,t+1} - \mathbf{a}_{l,t+1}\right) \\ \mathbf{H}_{lt} &= \mathbf{C}_{lt} - \mathbf{B}_{lt}(\mathbf{H}_{l,t+1} - \mathbf{R}_{l,t+1})\mathbf{B}_{lt}' \end{aligned}$$

Both $\mathbf{h}_{0t}$ and $\mathbf{H}_{0t}$ can be calculated using the same recursions with $n_{t0} = 0 \; \forall \; t$.

The filtering and smoothing relations described above are generalizations of those found in Carter and Kohn (1994), Fruhwirth-Schnatter (1994) and West and Harrison (1997), intended to deal with missing data or multiple observations at any point in time. Note that Step 2 can be computationally expensive since it implies running Forward-Filtering/Backward-Smoothing (FFBS) steps for each observation in the sample. A careful implementation requires at least one and at most two such steps for each observation: one to recalculate the parameters for the group to which the observation currently belongs and possibly another one to calculate those of the group were it is to be assigned.

Typically, the matrices $\mathbf{G}_t$ and $\mathbf{W}_t$ governing the evolution of the system will be unknown. However, since $\mathbf{G}_t$ and $\mathbf{W}_t$ define the moments of the baseline measure, inference on these quantities can be performed as discussed in Escobar and West

(1998). Therefore, inferences on structural parameters like discount factors, periodic components or autoregressive coefficients can be accommodated very easily. Details on some specific examples are discussed in the following sections. The sampler can also be extended to obtain backward-smoothed and/or k-step-ahead density estimates by calculating the corresponding predictive distributions. These predictive distributions, obtained from the Pólya urn scheme, correspond to mixtures of T densities of the form

$$p(y_{t+1}|\mathbf{y}_t, \ldots, \mathbf{y}_1, \cdots) =$$

$$\sum_{l=1}^{L} \frac{n_{l\cdot}}{n+\alpha} \mathsf{T}_{s_{lt}}(y_{t+1}|f_{l,t+1}, Q_{l,t+1}) + \frac{\alpha}{\alpha+n} \mathsf{T}_{s_{lt}}(y_{t+1}|f_{0,t+1}, Q_{0,t+1})$$

where $n_{l\cdot} = \sum_{t-1}^{T} n_{lt}$ and $f_{0,t+1}$, $Q_{0,t+1}$ are calculated from $\mathbf{m}_{lt}$, $\mathbf{C}_{lt}$, s_{lt} and S_{lt} using the same recursions as in step (1.a). Finally, if a Gamma prior is used to model the parameter α governing the precision of the Dirichlet process, the data augmentation scheme described Escobar and West (1995).

4.2.3 An example: Distribution Autoregressive Models (DAR)

Autoregressive (AR) models are one of the most popular tools in finance and econometrics, both because of their flexibility and interpretability. For example, as noted by a number of authors (Box and Jenkins, 1974; West, 1997; Aguilar *et al.*, 1999), high order stationary AR processes can be seen as approximations to Moving Average (MA) processes. Even more, Wold's decomposition ensures that high order AR processes are good approximations to any stationary process.

AR models have been used to model not only the mean structure of the time series, but also to understand other aspects of the distributions. For example, ARCH models (Engle, 1982) use an AR process to model the volatility of the process. The goal of this section is to push the idea of AR processes beyond that of a model for *the*

moments of distributions and get an equivalent formulation for the whole distribution. For simplicity, we start our discussion with the first-order, distribution autoregressive process (DAR(1)), which takes the form

$$y_{it} \sim \mathsf{N}(\mu_{it}, \sigma_{it}^2) \qquad (\mu_{it}, \sigma_{it}^2) \sim K_t$$

$$K_t = \sum_{l=1}^{\infty} w_l^* \delta_{(\mu_{lt}^*, \sigma_l^2)} \qquad \mu_{lt}^* \sim \mathsf{N}(\rho \mu_{l,t-1}^*, \sigma_l^2 U)$$

The name DAR(1) comes from the fact that the stochastic process defining the location of the Gaussian distributions used to represent the unknown density follow an autoregressive process with autocorrelation ρ and variance $\sigma^2 U$. This is a special case of the general model described in section 4.2 where $\mathbf{F}_t = 1$, $\mathbf{G}_t = \rho$ and $\mathbf{W}_t = U$. Therefore, the correlation a priori induced on the observations is

$$\mathbb{C}\mathrm{or}(y_{i,t}, y_{i',t+k}) = \frac{\rho^k}{1+\alpha}$$

This formulation extends the latent AR process models (West and Harrison, 1997) to infinite mixtures. As in the latent AR process, and unlike the typical Gaussian AR(1) process, $\rho = 0$ implies uncorrelated but dependent observations. Indeed, the case $\rho = 0$ generates identifiability issues since it is not possible to separate the evolution noise of the underlying process from the observational noise.

The model is completed by placing priors on ρ, U, μ_0 and α. For computational simplicity a conditionally conjugate distribution for the variance of the autoregressive process is used, $U \sim \mathsf{IGam}(a_U, b_U)$. Also, in order to ensure stationarity, we set

$$\rho \sim \mathsf{N}\left(0, \tau^2\right) \mathbf{1}_{(-1,1)} \qquad \mu_{l0}^* \sim \mathsf{N}\left(0, \frac{\sigma^2 U}{1-\rho^2}\right) \;\; \forall \;\; l$$

where $\mathsf{N}(a,b)\mathbf{1}_\Omega$ denotes the normal distribution with mean a, variance b and restricted to the set Ω. Finally, the DP precision factor α is given a $\mathsf{Gam}(a_\alpha, b_\alpha)$, which is conditionally conjugate.

Implementation of this models is a straightforward extension of that in section 4.2.2. Conditional on ρ, U and α, the model is a discrete-time DDP. On the other hand, conditional on the allocation indicators, the sample paths $\{(\boldsymbol{\mu}_l^*, \sigma_l^{*2})\}_{l=1}^L$ are iid samples from the baseline measure. Therefore, samples from ρ and U can be easily obtained using the following full-conditional distributions.

- The variance of the autoregressive process can be obtained by sampling U from its full conditional distribution,

$$U|\cdots \sim \mathsf{IGam}\left(a_U + \frac{T}{2}, b_U + \sum_l \left[\frac{(1-\rho^2)\mu_{l0}^*}{\sigma_l^{*2}} + \sum_{t=1}^T \frac{\mu_{lt}^* - \mu_{l,t-1}^*}{\sigma_l^{*2}}\right]\right)$$

- The full conditional distribution for the autocorrelation coefficient takes the form

$$p(\rho|\cdots) \propto (1-\rho^2)^{-L/2} \exp\left\{-\frac{1-\rho^2}{2U}\sum_{l=1}^L \frac{\mu_{l0}^{*2}}{\sigma_l^2}\right\}$$
$$\exp\left\{-\frac{1}{2}\left[\frac{b}{U} + \frac{1}{\tau^2}\right]\left[\rho - d\left(\frac{b}{U} + \frac{1}{\tau^2}\right)^{-1}\right]\right\}$$

  where

$$b = \sum_{t=1}^T \sum_{l=1}^L \frac{\mu_{l,t-1}^{*2}}{\sigma_l^{*2}} \qquad d = \sum_{t=1}^T \sum_{l=1}^L \frac{\mu_{l,t}^* \mu_{l,t-1}^*}{\sigma_l^{*2}}$$

  Note that this expression does not correspond to any known distribution. However, we recognize the third term (which happens to contain most of the information provided by the observations) as a normal kernel. Therefore an efficient independent-proposal Metropolis step can be devised to sample from this full-conditional distribution. Given the current value of the autoregression parameter $\rho_{(c)}$ in the previous iteration, propose

$$\rho_{(p)} \sim \mathsf{N}\left(d\left(\frac{b}{U} + \frac{1}{\tau^2}\right)^{-1}, \left(\frac{b}{U} + \frac{1}{\tau^2}\right)^{-1}\right)\mathbf{1}_{(-1,1)}.$$

Then, accept this proposal with probability

$$\min\left\{1, \left(\frac{1-\rho_{(p)}^2}{1-\rho_{(c)}^2}\right)^{-L/2} \exp\left\{-\frac{\rho_{(c)}^2-\rho_{(p)}^2}{2U}\sum_{l=1}^{L}\frac{\mu_{l0}^{*2}}{\sigma_l^{2*}}\right\}\right\}$$

Otherwise retain the previous value $\rho_{(c)}$.

Extending the previous ideas to a DAR(p) is straightforward. The model takes the form

$$y_{it} \sim \mathsf{N}(\mu_{it}, \sigma_{it}^2) \qquad (\mu_{it}, \sigma_{it}^2) \sim K_t$$

$$K_t = \sum_{l=1}^{\infty} w_l^* \delta_{(\mu_{lt}^*, \sigma_l^2)} \qquad \mu_{lt}^* \sim \mathsf{N}\left(\sum_{r=1}^{p} \rho_r \mu_{l,t-r}^*, \sigma_l^2 U\right)$$

In terms of inference, the DAR(p) requires a slight adaptation of the FFBS algorithm described in section 4.2.2 due to the fact that $\mathbf{W}_t$ is a singular matrix. This modification is described in West and Harrison (1997), Chapter 15.3.2 for the Gaussian AR(p) model.

4.3 A model with time-dependent variances

Although the class of models in the previous section is very flexible, they impose some unappealing constrains on the sequence of density estimates. In particular, by assuming that the variance of each component does not change across time we are forcing the variance of the mixture to also remain fixed. Although in some applications this might not be a problem, the fact is that, for finance and econometric applications, this is a serious constrain.

Since allowing for the variances to change simultaneously in time and for each component yields models that are computationally intractable, we compromise by considering in this section models that assume that the variance of each component

is the same for any given time t, but let the variance evolves in time. The resulting model can be interpreted as a Bayesian kernel density method where the kernel bandwidth is adaptive in time instead of in space. Specifically, the model takes the form

$$\begin{aligned}
y_{it} &\sim \int \mathsf{N}\left(\mathbf{F}_{it}\boldsymbol{\theta}_t, \sigma_t^2\right) K_t(d\boldsymbol{\theta}_t) & K_t &= \sum_{l=1}^{\infty} w_l^* \delta_{\boldsymbol{\theta}_{lt}} \\
w_l^* &= z_l^* \prod_{s=1}^{l-1}(1-z_s^*) & z_l^* &\sim \mathsf{beta}(1,\alpha) \\
\boldsymbol{\theta}_{lt}|\boldsymbol{\theta}_{l,t-1}, \sigma_t^2 &\sim \mathsf{N}(\mathbf{G}_t\boldsymbol{\theta}_{l,t-1}, \sigma_t^2\mathbf{W}_t) & \boldsymbol{\theta}_{0l}|\sigma_0^2 &\sim \mathsf{N}(\mathbf{m}_0, \sigma_0^2\mathbf{C}_0) \\
\sigma_t^2 &\sim \frac{\delta\sigma_{t-1}^2}{\zeta_t} & \zeta_t &\sim \mathsf{beta}(\delta n_t, (1-\delta)n_t) \\
\sigma_0^2 &\sim \mathsf{IGam}(s_0, s_0 S_0)
\end{aligned} \tag{4.3}$$

This model is very similar to the one discussed in section 4.2. Note, however, that we are not mixing over the variance. Instead, we make it dependent on time and let it evolve using the stochastic volatility approach of Uhlig (1997).

4.3.1 Inference

The MCMC algorithm for this model is very similar to the one presented in Section 4.2.2. As in the previous case, let L be the current number of components that have observations allocated to them, $n_{lt}^\star$ be the number of observations at time t assigned to group l, $n_l = \sum_t n_{lt}$, $\boldsymbol{\Theta} = \{\Theta_1^\star, ..., \Theta_L^\star\}$, and $\boldsymbol{\Sigma} = \{\sigma_0^2, ..., \sigma_T^2\}$ be the current estimated values for those paths and time-varying variances. Also, $\xi_{it} = l$ iff $\boldsymbol{\theta}_{it} = \boldsymbol{\theta}_{lt}^\star$. Given values for the structural parameters $\mathbf{F}_{it}$, $\mathbf{G}_{it}$ and $\mathbf{W}_{it}$ and after initialization of the parameters, an MCMC sampler alternates through the following steps:

1. Generate $\boldsymbol{\Theta}_l^\star | \{\mathbf{y}|\xi_{it} = l\}$ using the following FFBS algorithm

(a) Forward filtering using the following recursions

$$
\begin{aligned}
\mathbf{m}_{lt} &= \begin{cases} \mathbf{a}_{lt} + \mathbf{A}_{lt} e_{lt} & \text{if } n_{lt} > 0 \\ \mathbf{a}_{lt} & \text{if } n_{lt} = 0 \end{cases} \\
\sigma_t^2 \mathbf{C}_{lt} &= \begin{cases} \mathbf{R}_t - \mathbf{A}_{lt} \mathbf{Q}_{lt} \mathbf{A}'_{lt} & \text{if } n_{lt} > 0 \\ \mathbf{R}_{lt} & \text{if } n_{lt} = 0 \end{cases} \\
\mathbf{A}_{lt} &= \mathbf{R}_{lt} \mathbf{F}_{lt} \mathbf{Q}_{lt}^{-1} \\
\mathbf{e}_{lt} &= \mathbf{y}_{lt} - \mathbf{f}_{lt} \\
\mathbf{f}_{lt} &= \mathbf{F}'_{lt} \mathbf{a}_{lt} \\
\mathbf{Q}_{lt} &= \mathbf{F}'_{lt} \mathbf{R}_{lt} \mathbf{F}_{lt} + \sigma_t^2 \mathbf{I} \\
\mathbf{a}_{lt} &= \mathbf{G}_{lt} \mathbf{m}_{l,t-1} \\
\mathbf{R}_{lt} &= \sigma_{t-1}^2 \mathbf{G}_{lt} \mathbf{C}_{l,t-1} \mathbf{G}'_{lt} + \sigma_t^2 \mathbf{W}_{lt}
\end{aligned}
$$

where $\mathbf{y}_{lt}$ is made of all observations assigned to group l at time t and $\mathbf{F}_{lt}$ is matrix whose rows are the corresponding $\mathbf{F}_{it}$ vectors.

(b) Sample $\boldsymbol{\theta}_{lT}|\mathbf{y}_l$ from $N(\mathbf{m}_{lT}, \sigma_T^2 \mathbf{C}_{lT})$. Then recursively sample $\boldsymbol{\theta}_{lt}|\boldsymbol{\theta}_{l,t+1}, \mathbf{y}_l$ from $N(\mathbf{d}_{lt}, \mathbf{D}_{lt})$ where

$$
\begin{aligned}
\mathbf{d}_{lt} &= \mathbf{B}_{lt} \left(\boldsymbol{\theta}_{l,t+1} - \mathbf{a}_{l,t+1} \right) \\
\mathbf{D}_{lt} &= \sigma_t^2 \mathbf{C}_{lt} - \mathbf{B}_{lt} \mathbf{R}_{l,t+1} \mathbf{B}'_{lt} \\
\mathbf{B}_{lt} &= \sigma_t^2 \mathbf{C}_{lt} \mathbf{G}_{t+1} \mathbf{R}_{l,t+1}^{-1}
\end{aligned}
$$

2. Generate the sequence of variances $\sigma_l^2|\mathbf{y}$ using another FFBS algorithm

(a) Forward filtering using the following recursions

$$\begin{aligned} s_t &= \delta s_{t-1} + n_t + p \\ S_t &= \frac{\delta S_{t-1} + \sum_{i=1}^{n_t} (y_{it} - \mathbf{F}_{it}\theta^{\star}_{\xi_{it},t})^2 + (\theta_t - \theta_{t-1})' \mathbf{W}_t^{-1} (\theta_t - \theta_{t-1})}{\delta s_{t-1} + n_t + p} \end{aligned}$$

(b) Backward sample, starting with $\sigma_T^2 \sim \left(\frac{s_T}{2}, \frac{s_T S_T}{2}\right)$ and then letting

$$\sigma_{t-1}^2 | \sigma_t^2 = \frac{1}{\eta_{t-1} + \frac{\delta}{\sigma_t^2}}$$

for all $0 \leq t < T$ where

$$\eta_{t-1} \sim \mathsf{Gam}\left(\frac{(1-\delta)s_{t-1}}{2}, \frac{S_{t-1}}{2}\right)$$

3. Sample $\xi_{ij} | \mathbf{y}, \xi_{-(ij)}, \boldsymbol{\Sigma}, \cdots$ from a multinomial distribution with probabilities:

$$\begin{aligned} q_l &= n_l^- p(y_{ti} | \mathbf{y}^-, \xi_{(ij)}^- \;\; l = 1, \ldots, L^-) \\ &= n_l^- \mathbf{N}\left(y_{ti} | \mathbf{h}_{lt}, \mathbf{H}_{lt}\right) \\ q_{L+1} &= \alpha p(y_{ti} | S_0) \\ &= \alpha \mathbf{N}\left(y_{ti} | \mathbf{h}_{t0}, \mathbf{H}_{t0}\right)) \end{aligned}$$

As before, $\mathbf{h}_{lT} = \mathbf{m}_{lT}$, $\mathbf{H}_{lT} = \mathbf{C}_{lT}$ and

$$\begin{aligned} \mathbf{h}_{lt} &= \mathbf{B}_{lt}\left(\mathbf{h}_{l,t+1} - \mathbf{a}_{t+1}\right) \\ \mathbf{H}_{lt} &= \mathbf{C}_{lt} - \mathbf{B}_{lt}(\mathbf{H}_{l,t+1} - \mathbf{R}_{l,t+1})\mathbf{B}_{lt}' \end{aligned}$$

4.4 Estimating implied risk-neutral distributions

4.4.1 Option implied risk-neutral distribution (RNPD)

It is common knowledge that market prices of options contain information regarding market expectations. Important information can be extracted from the derivatives

markets and used for several purposes, such as the probabilities of adverse movements in the market, as well as monetary authorities assessing market expectations. This is why cross sections of option prices have been investigated in order to retrieve the implied probability density distribution of the underlying S_t (stock, inflation, currency, interest rates), which represents market expectations. Retrieving the RNPD is a typical example of an *Inverse Problem* (Tikhonov, 1963), and the Bayesian methodology is a way to regularize it through the elicitation of a prior distribution, which acts as a penalization function (Wolpert and Ickstadt, 2004).

In any option pricing model, looking for a suitable and realistic stochastic process to model the underlying stock price is essential. Nonparametric methods using maximum entropy techniques have been successfully used in the case of Lévy processes (Cont and Tankov, 2003). Such models may exhibit stylized features common in financial applications, such as skewness, volatility clustering, jumps, fat-tails (Cont and Tankov, 2003), as well as multimodality of the log-returns $\frac{dS_t}{S_t}$, but are typically awkward to implement. Other approaches (Melick and Thomas, 1997; Rebonato, 2004) have used finite mixtures of parametric distributions to fit the RNPD using number of mixtures as well, but to the cost of overfitting the observed RNPD, which usually leads to poor prediction. The novelty or our estimation relies on making no assumptions other than using the call-put parity equation to generate observations under the RNPD. The no arbitrage condition from equation (4.4) enables us to imply the existence (although not the uniqueness) of a RNPD (Delbaen and Schachermayer, 2006), whose posterior distribution is the focus of interest. Besides, our algorithm allows us to not only determine sequentially in time the optimal number of components, but also the dependence among the RNPD.

In what follows, we focus on *European options*. The holder of a European call option has the right, but not the obligation, to buy an underlying security at a

specified date (*expiration date*) for a contractually specified amount (*strike price*), irrespective of the market value of the security on that date. The underlying securities of options can be stocks, indices such as the *Standard and Poor's 500*, interest rates, etc. At the expiration date T, the value of the option is $(S_T - K, 0)^+$, the maximum of $S_T - K$ and zero. Payoff is at later time T, so under constant discount rate r present value of the call option at time t will be $\exp(-r(T-t))(S_T - K)^+$, where K is the strike price. In what follows, let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$ be a probability space, equipped with a filtration satisfying the *usual hypotheses* (see Protter, 1990, p. 3, for definition). The classical framework of option pricing assumes that a call option C_t whose payoff $(S_T - K, 0)^+$ depends on our underlying S_T at the expiration date T, can be computed via the following integration:

$$\begin{aligned} C_t &= \exp(-r(T-t))\, E_{\mathbb{Q}}\{(S_T - K, 0)^+ \mid \mathcal{F}_t\} \\ &= \exp(-r(T-t)) \int_{\mathcal{S}} (S_T - K, 0)^+ d\mathbb{Q}(S_T \mid \mathcal{F}_t) \end{aligned}$$

where $\mathcal{S}$ is the sample space for the stock price S_T at terminal date T, K is the strike and integration is performed under the risk-neutral measure $\mathbb{Q}$, such that the discounted stock price $\exp(-rt)\, S_t$ is a $\mathbb{Q}$-martingale (Delbaen and Schachermayer, 2006), yielding

$$C_t = \exp(-r(T-t)) \int_{\mathcal{S}} \max(S_T - K, 0) d\mathbb{Q}(S_T \mid \mathcal{F}_t)$$

Similarly, for the put option

$$P_t = \exp(-r(T-t)) \int_{\mathcal{S}} \max(K - S_T, 0) d\mathbb{Q}(S_T \mid \mathcal{F}_t)$$

From the put-call parity price we obtain (Hull, 2005):

$$\begin{aligned} C_t - P_t &= S_t - K \exp(-r(T-t)) \\ \Rightarrow S_t &= C_t - P_t + K \exp(-r(T-t)) \end{aligned} \tag{4.4}$$

Given the call and put prices for each trade (which are typically available as part of market data), equation (4.4) can be used to obtain observations from the risk neutral probability distribution of S_t across different strikes K. These observations can then be used to estimate the RNPD of the underlying S_t at time each time t in a nonparametric fashion, for example through a discrete time DDP. Using a nonparametric method in this setup is attractive because it assumes the least amount of assumptions regarding the governing dynamics of the probabilistic structure of S_t.

4.4.2 RNPD in the S&P500 INDEX

In what follows, we concentrate on options of the S&P500 index with three-month maturity times traded between January 4th 1993 and March 17th 1994, for a total of $n = 4385$ trades spread over $T = 293$ days, with sample sizes in any specific day varying between 2 and 26. The data was constructed by Yacine Ait-Sahalia and has been used in the empirical study of Duffie *et al.* (2000). A plot of the prices inferred from the put-call parity (4.4) using the Libor as the interest rate (Panigirtzoglou and Skiadopoulos, 2004) is presented in Figure 4.1. Note that the distributions on any specific day may be highly skewed and may have very heavy tails. Means (plotted in orange) vary wildly, specially during the summer of 1993 when fewer trades occur and extreme values are highly influential. Even more, the kernel density estimates in Figure 4.2 show that the risk neutral distributions change dramatically across time and might actually be multimodal.

We use the DAR(1) model described in section 4.2.3 to model the unknown distribution of S_t. In order to justify a zero mean DAR(1) process, the data has been recentered using the global mean. The goal is to obtain smoothed estimates for the mean and volatility of the process, as well as for the distribution itself, along with a predictive model that can be used to infer the distribution at subsequent dates.

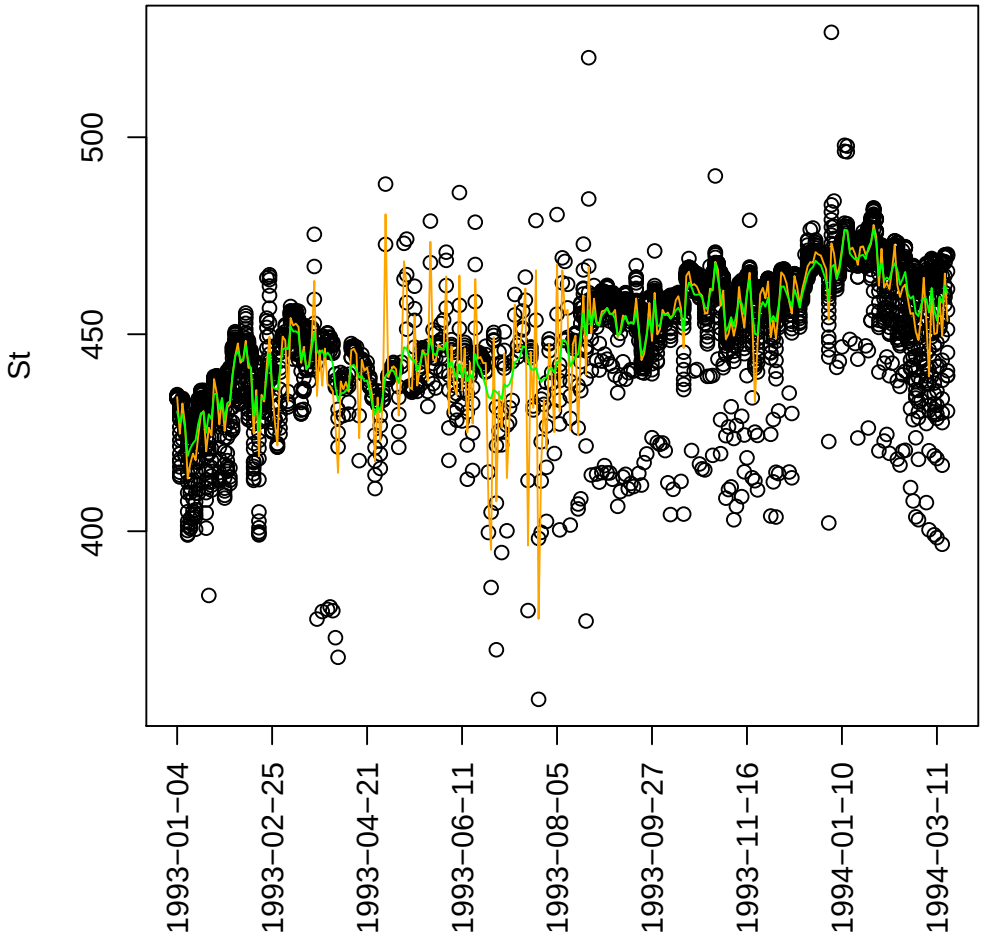


Figure 4.1: S&P500 prices induced by the call-put non-arbitrage condition. Dots correspond to the raw data, the orange line corresponds to the empirical mean of the observations at the corresponding time point and the green line to the smoothed mean under the DAR(1) model.

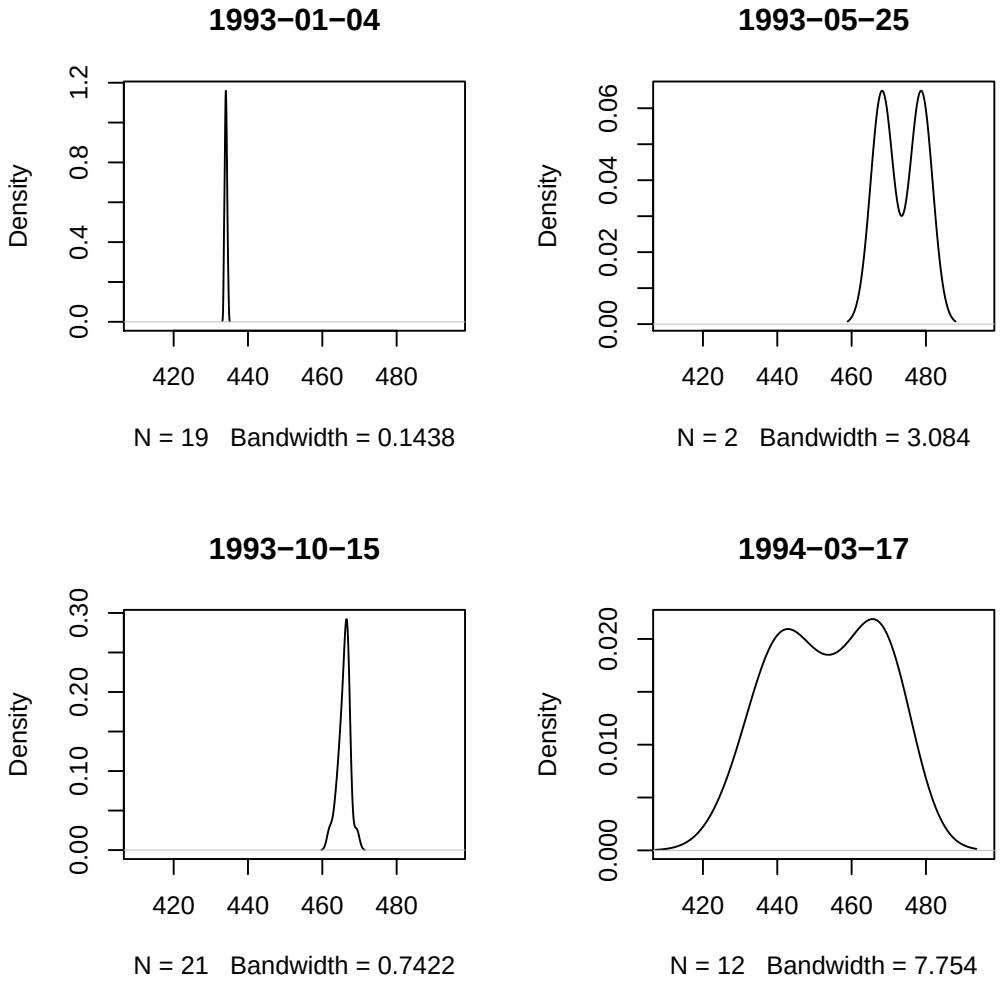


Figure 4.2: Kernel density estimates of S&P500 prices on Jan 4, 1993; May 25, 1993; Oct 15, 1993 and Mar 17, 1994. The number of observations N and the bandwidth estimated through cross-validation are shown bellow each plot.

Hyperparameters have been chosen as $\tau^2 = 1$, $s_0 = 1$ $S_0 = 40$ (about one sixth of the empirical observational variance on the whole sample) and $a_U = b_U = 1$. In order to increase the flexibility of the model, a hyperprior on the concentration parameter of the Dirichlet Process, $\alpha \sim \mathsf{Gam}(1, 1)$ was used. All results presented here are based on 25,000 samples of the posterior distribution obtained after a burn-in period of 5,000 iterations, and they seem to be robust to moderate changes in the prior distribution.

Table 4.1 shows posterior estimates for the concentration and baseline measure parameters in the model. As expected, the autocorrelation in the process is rather high. The evolution variance is also relatively large, being about 12 times larger than the average observational variance. Roughly speaking, the model uses between 7 and 10 mixture components to model the collection of 293 distributions, but three of those components capture around 90% of the trades (50%, 25% and 15% respectively), with the rest of the components with higher observational variance explaining the very extreme observed values. This observation has a potential explanation in the field of behavioral finance. It has been argued that individuals tend to over-weight small probability outcomes related to significant losses, which is therefore consistent with high-implied volatility on deep out of the money derivatives (Rasiel, 2003). Indeed, the observational variance of every component of the mixture decreases as the number of data points allocated to that group increases, which happens in the middle around the current market price. On the contrary, with extreme observations which might constitute a group (cluster) of their own, there is a much higher uncertainty and thus observational variance regarding the distribution. This analysis makes the implied volatility smile nothing but the consequence of human behavior, which is to choose a direction of where the underlying might go, and then choose the probability distribution, explaining the fact that the RNPD is multimodal. This last feature

Table 4.1: Posterior mean, median and symmetric 95% probability interval for some parameters in the DAR(1) model fitted to the S&P500 data

	Mean	Median	90% Prob Int
α	0.92	0.87	(0.36,1.76)
ρ	0.974	0.974	(0.962,0.985)
U	13.12	13.05	(11.23,15.38)

of the RNPD also shows the proportion of different market participants (number of mixtures) and their bullish or bearish views regarding the future on the S&P 500.

If the implied volatility is calculated using the original Black & Scholes (which assumes a single lognormal with constant volatility), a volatility smile would follow, where the implied volatility is an decreasing/increasing function of the strike price. This is so because the higher probability of tail observations vis-a-vis the lognormal distribution will imply higher market prices for those deep out-of-the-money calls. Anything leading to fat tails vis-a-vis the lognormal will deliver the volatility smile, and mixture distributions are one such thing. However, our methodology gives the nice interpretation that fat tails are often a consequence of the market being divided into several broad groups, the bulls and the bears (together with their nuances), who make their bets accordingly in their private valuation of options. Figure 4.3 shows smoothed density estimates $p(y_t|\mathbf{y}_1, \ldots, \mathbf{y}_T)$ and one-step-ahead predicted densities $p(y_t|\mathbf{y}_1, \ldots, \mathbf{y}_{t-1})$ for the last 8 days in the series. Note that the model seems to provide both a good fit to the observed data and sensible predictions of future observations, even for extreme observations. Both density estimates are effective in capturing the main characteristics of the data, namely high skewness and multimodality. However, the modes show up more sharply in the smoothed densities because of the additional noise being convoluted during the prediction.

Another interesting feature of the model is its robustness to small samples with extreme values, which is clearly hinted by the behavior of the estimated mean during

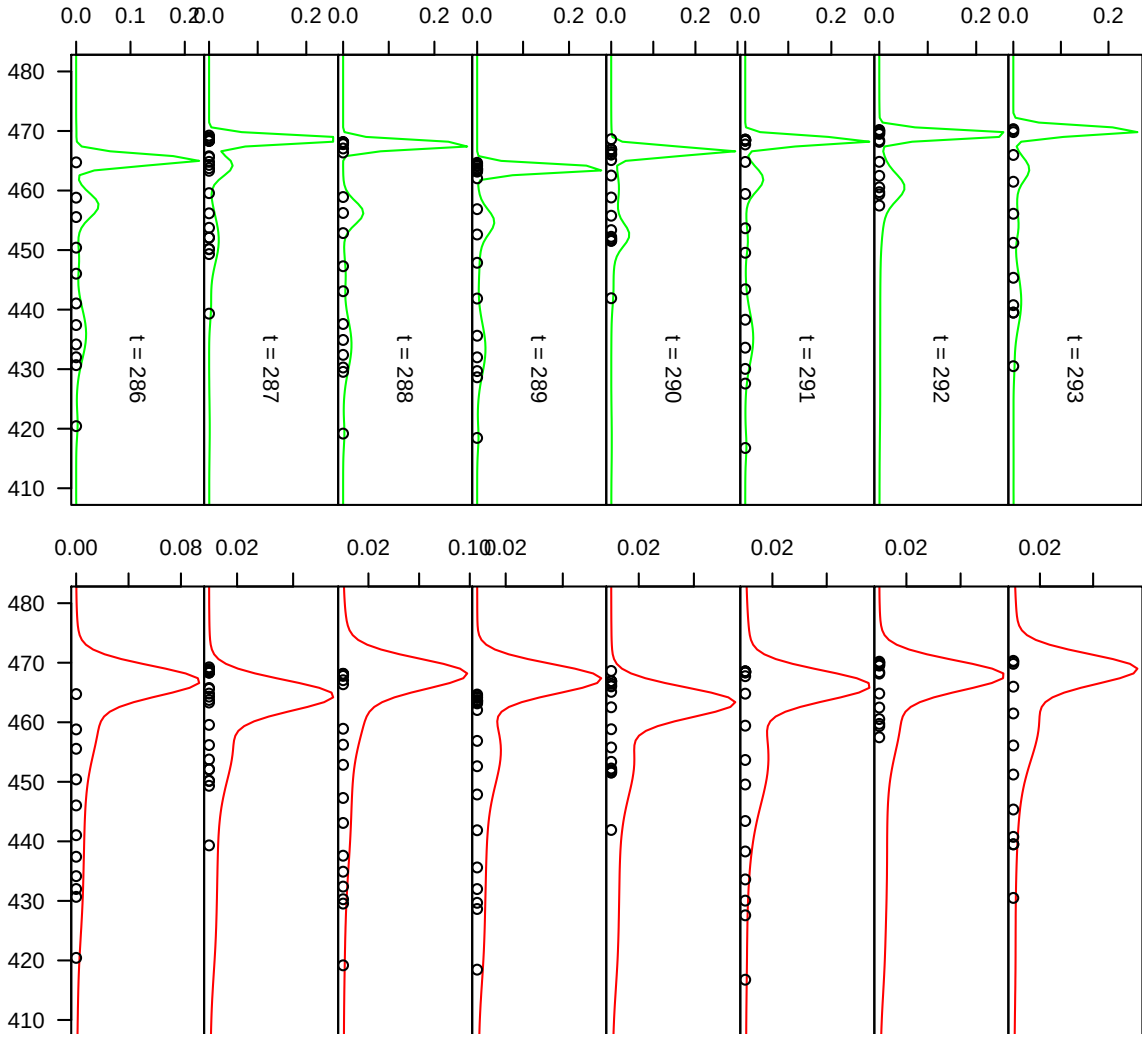


Figure 4.3: Smoothed (green) and one-step-ahead predicted densities (red) between March 8, 1994 ($t = 286$) and March 17, 1994 ($t = 293$) obtained from the DAR(1) model. The dots correspond to the actual observations.

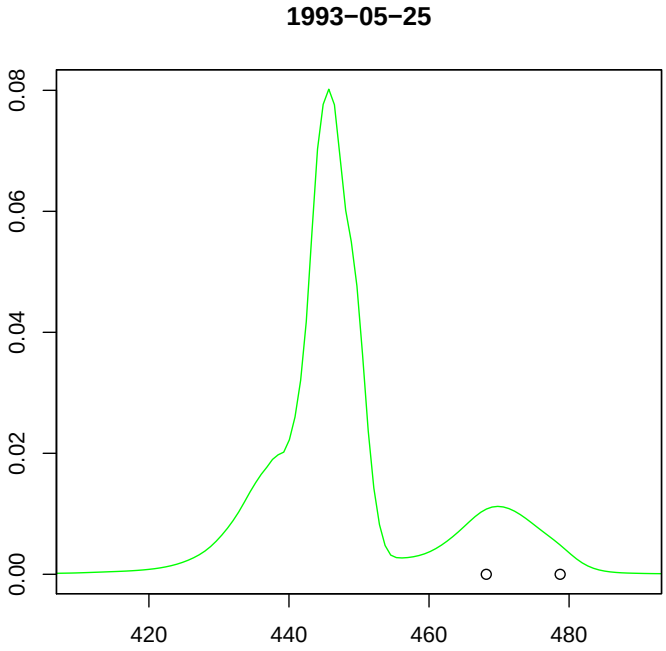


Figure 4.4: Smoothed density estimate for May 25, 1993 obtained from the DAR(1) model.

the summer of 1993 (green line in Figure 4.1). For example, consider the market during May 25, 1993. In this date, only two trades occurred, and the raw density estimate was presented in the upper right panel of Figure 4.2 is highly skewed towards relatively high prices. Most parametric analysis would associate this with a large shift in the location of the distribution. The smoothed density on this date resulting from the DAR(1) model can be seen in figure 4.4. Since there is a bulk of information contained in the behavior of the market in adjacent days, the two observed market valuations are attributed by the model not to the mainstream investors (represented by the higher weight component), but to their more risk-prone counterparts. This is an interesting example of the regularization properties inherent to this class of models.

The model can also be used to obtain robust smoothed estimates for the mean and volatility underlying the process. The time varying aspect of the first and second moments can be of interest for financial institutions interested in determining and measuring the exposure in their portfolios to financial risk, and more precisely to market risk. By using the dynamic DDP model, the risk measures can then be computed under distributions not only exhibiting fat tails, volatility clustering, varying mean returns, multimodality, but estimation uncertainty as well. This last feature can therefore take into account model risk due to the use of nonparametric methods (Cont, 2006).

Figure 4.5 shows a plot of the interquartile range (IQR) associated with the smoothed densities generated by the model. In normal distributions, $IQR/1.349$ is a consistent estimator of the standard deviation of the process. For more general distributions, the IQR provides a robust alternative to the variance that can be used to compare the volatility at different points in time. From Figure 4.5 we can observe two high volatility periods: one in late June 1993 and another one by the end of the recorded series in March 1994. It is interesting to notice that the DDP extends Bollerslev (1986) to a nonparametric method exhibiting not only mean shifts in the distribution but also volatility clustering, while working directly in the price level scale.

4.5 Discussion

In this paper we discuss a class of models for dynamic density estimation in discrete time that allows us to borrow information across adjacent observations and obtain robust inferences, not only for the distribution itself, but also for other quantities of interest like the mean or the quantiles of the process.

The example in this paper focuses on distribution autoregressive models to esti-

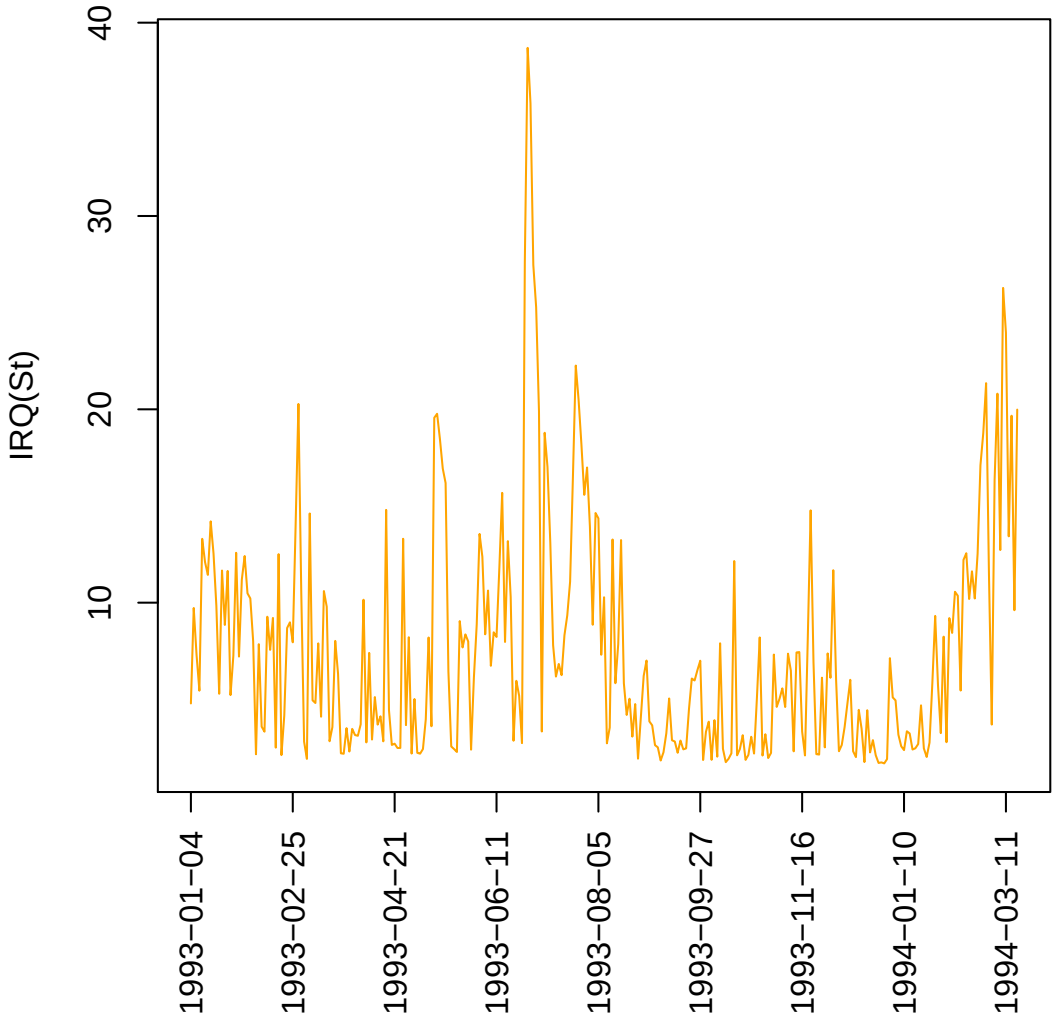


Figure 4.5: Estimated volatility (interquantile range) in the S&P500 series.

mate implied risk-neutral distributions in options markets, which is in itself a novel contribution. It allows us to drop cumbersome parametric assumptions that are clearly not justified by the data but are part of the current state-of-the-art in the field. In spite of this, our formulation is not constrained to DAR models and is indeed much more general: trends, periodicities and even covariates can be easily included into the model. Also, applications extend not only to other areas in finance (stochastic volatility models, risk management, credit analysis and FX options) but also to other fields like epidemiology, climatology and engineering, some of which are the focus of current research. However, despite their generality, it is important to emphasize that the models described in this paper induce dependence in the distributions themselves and assumes that observations are exchangeable within every time point given that distribution. Therefore, our model is not suitable for the analysis of longitudinal studies where the same experimental unit is followed at different times points.

One of the key characteristics of these models is the simplicity of their implementation. By using common weights to define the collection of distributions, computational tools for standard Dirichlet Process can be employed. The samplers we have described in this chapter use the Pólya urn representations to marginalize over the collection of unknown distributions and sample the paths of the different components in the mixture. However, there is an ample literature on MCMC techniques for Dirichlet processes that can be exploited to obtain alternative exact or approximate samplers. Some attractive options that can be readily implemented are truncation approximations or variational methods, with the latter being specially appealing for very large sample samples.

Another enormous advantage is its interpretability as a multiprocess type II model. The constant weight assumption allows us to think in term of alternative

evolution paths for the process receiving different weights depending of the behavior of the process. These paths can have natural meaning in the context of a specific application, as in the option RNPD example. Also, the estimates of the parameters underlying $\mathbf{G}_t$ of $\mathbf{W}_t$ (if there are any) can be interpreted as averages over observed paths

There are two main drawbacks with our model formulation. As discussed by MacEachern (2000) and Griffin and Steel (2006b), discrete time DDPs do not posses the intuitive asymptotic behavior expected as $\mathbf{W}_t \to \infty$. Indeed, it is impossible to generate a collection of independent distributions within this framework. However, this is hardly a limitation (at least in the context of financial application) since in most cases the existence of dependence is not in question and it is rarely the case that dependence tests are required.

Identifiability issues also arise with single-observation time series. Indeed, it is rare in finance and econometric applications to have replicates at any given time. As noted by Griffin and Steel (2006b), it is hard for constant weight constructions to differentiate between variability in the baseline measure and multimodality unless replicates are available. The intuition in the context of discrete-time DDPs is straightforward: are observed changes in the process due to a unimodal process with fairly large evolution variance or to an almost constant multimodal distribution? We argue that the order in which new components are added as the number of observations grows contains valuable information, and therefore strong prior distributions enforcing small values of $\mathbf{W}_t$ will essentially solve the identifiability problem. This type of constrains, which might seem awkward in a spatial context like Gelfand *et al.* (2005), can be introduced in discrete time DDPs more naturally. This specific topic is a work in progress.

Chapter 5

Nonparametric functional data analysis through Bayesian density estimation

The last fifteen years have seen a revolution in the amount and quality of data being collected in empirical research. 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 $j = 1, \ldots, J$ indexes an experimental condition and $i = 1, \ldots, n_j$ 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 $f_1, \ldots, f_J : \mathbb{R}^p \to \mathbb{R}^q$ 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 di-

vided 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 variables, 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 (2007) 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) develop 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 set of conditional distributions $p_1(\mathbf{y}|\mathbf{x}), \ldots, p_J(\mathbf{y}|\mathbf{x})$, 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 $p_1(\mathbf{y}, \mathbf{x}), \ldots, p_J(\mathbf{y}, \mathbf{x})$ that uses dependent Dirichlet processes mixtures (MacEachern, 2000; DeIorio *et al.*, 2004; Gelfand *et al.*, 2005). This method is conceptually related to the double kernel method of Fan *et al.* (1996) and 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 heteroscedasticity 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 described in chapter 2 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; Heard *et al.*, 2006), functional clustering is used to identify coregulated genes, which are typically assumed to be members of a common transcription pathway.

This chapter is organized as follows: Section 5.1 reviews the original model presented by Müller *et al.* (1996) for nonparametric regression through Dirichlet process mixtures. In section 5.2 we introduce our method for multiple curves and discuss its properties. In section 5.3 we give conditions for posterior consistency of these models, providing theoretical support for our approach. Finally, section 5.5 illustrates the approach through application to temperature profile data in the North Atlantic, while section 5.7 contains a brief discussion.

5.1 Single curve nonparametric regression

Consider the following application of the model for multivariate density estimation described in (1.2), which we will call the MEW model:

$$\begin{aligned}
\mathbf{z}_i = (\mathbf{y}_i, \mathbf{x}_i) &\sim \mathsf{N}_{p+q}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) \qquad i = 1, \ldots, n \\
(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i) &\sim H \qquad\qquad H \sim \mathsf{DP}(\alpha H_0) \\
H_0 = \mathsf{NIW}_{p+q}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0) &\qquad \alpha \sim \mathsf{Gam}(a_\alpha, b_\alpha) \\
\boldsymbol{\theta}_0 &\sim \mathsf{N}_{p+q}(\boldsymbol{\theta}_{00}, \mathbf{D}_{00}) \qquad \boldsymbol{\Sigma}_0 \sim \mathsf{W}_{p+q}(\gamma, \boldsymbol{\Sigma}_{00}) \\
\kappa_0 &\sim \mathsf{Gam}(a_\kappa, b_\kappa),
\end{aligned} \tag{5.1}$$

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

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

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 (1.3), a posterior estimate for the conditional density can be obtained as

$$g^n(\mathbf{y}|\mathbf{x}) = \int \frac{\phi_q(\mathbf{y}|\boldsymbol{\theta}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\theta}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy})\phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})}{\int \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx}) H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n)} H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n). \tag{5.2}$$

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 (5.2) by calculating the conditional expectation,

$$f^n(\mathbf{x}) = \mathbb{E}(\mathbf{y}|\mathbf{x}, \mathbf{z}^n) = \int \frac{(\boldsymbol{\theta}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\theta}_x))\phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})}{\int \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx}) H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n)} H_0^n(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}^n). \tag{5.3}$$

For any fixed $\mathbf{x}$, the conditional distribution in (5.2) is a locally weighted mixture of normals, with the conditional expectation in (5.3) 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 5.3 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 $\alpha \to 0$, the prior on H becomes a single point mass with probability one, and the model in 5.1 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).

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.

5.2 Hierarchical nonparametric models for curves

Section 5.1 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.

Consider now the problem of inferring multiple curves $f_1, \ldots, f_J$, using the data $\mathbf{z}_1^{n_1}, \ldots, \mathbf{z}_J^{n_J}$, where $\mathbf{z}_j^{n_j} = (\mathbf{z}_{1j}, \ldots, \mathbf{z}_{j,n_j})$ is the set of n_j observations obtained under experimental condition j. The model described in Section 5.1 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) 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 5.1.

As an illustration, consider a model for functional clustering using mixtures of nested Dirichlet Processes. Recall from section 1.3.1 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 $\{H_1, \ldots, H_J\}$ used to estimate the joint probability distributions $\{p_1(\mathbf{y}, \mathbf{x}), \ldots, p_J(\mathbf{y}, \mathbf{x})\}$, 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 5.1, where

$$
\begin{aligned}
\mathbf{z}_{ij} = (\mathbf{y}_{ij}, \mathbf{x}_{ij}) \sim \mathsf{N}_{p+q}(\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_{ij}) \qquad & i = 1, \ldots, n_j \; ; \; j = 1, \ldots, J \\
(\boldsymbol{\theta}_{ij}, \boldsymbol{\Sigma}_{ij}) \sim H_j \qquad & \mathcal{H} = \{H_1, \ldots, H_J\} \sim \mathsf{nDP}(\alpha, \beta, H_0) \\
H_0 = \mathsf{NIW}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0) \qquad & \kappa_0 \sim \mathsf{Gam}(a_\kappa, b_\kappa) \\
\boldsymbol{\theta}_0 \sim \mathsf{N}(\boldsymbol{\theta}_{00}, \mathbf{D}_{00}) \qquad & \boldsymbol{\Sigma}_{00} \sim \mathsf{W}(\gamma, \boldsymbol{\Sigma}_{00}) \\
\alpha \sim \mathsf{Gam}(a_\alpha, b_\alpha) \qquad & \beta \sim \mathsf{Gam}(a_\beta, b_\beta).
\end{aligned}
\tag{5.4}
$$

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

$$\begin{aligned} f_j^{n_+}(\mathbf{x}) &= \mathbb{E}_{H_j^{n_+}}(\mathbf{y}|\mathbf{x}, \mathbf{z}_1^{n_1}, \ldots, \mathbf{z}_J^{n_J}) \\ &= \int \frac{(\boldsymbol{\theta}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\theta}_x))\phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx})}{\int \phi_p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\Sigma}_{xx}) H_{0j}^{n_+}(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}_1^{n_1}, \ldots, \mathbf{z}_j^{n_J})} H_{0j}^{n_+}(d\boldsymbol{\theta}, d\boldsymbol{\Sigma}|\mathbf{z}_1^{n_1}, \ldots, \mathbf{z}_J^{n_J}). \end{aligned} \tag{5.5}$$

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

Although an explicit form is not available for the estimated regression function $f_j^{n_+}(\mathbf{x})$ or the estimated density $g_j^{n_+}(\mathbf{y}|\mathbf{x})$, 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 positive 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, 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 $\alpha \to 0$, the model assumes a single cluster of curves (i.e. all the samples arise from the same underlying function), while $\alpha \to \infty$ 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, in controlling the number of distinct $(\boldsymbol{\theta}_{lk}^*, \boldsymbol{\Sigma}_{lk}^*)$, the parameter β 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.

5.3 Posterior consistency

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 $j = 1, \ldots, J$ is as follows: 1) Covariates $\mathbf{x}_{1j}, \mathbf{x}_{2j}, \ldots$ 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 $g_j^n(\mathbf{y},\mathbf{x})$ and $g_j^0(\mathbf{y},\mathbf{x})$ are absolutely continuous, the result follows from the definition of weak consistency. □

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

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

Therefore, from Bayes' rule,

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

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

Corollary 2. *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 2 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 5.1, Corollary 2 can be made more specific.

Corollary 3. *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., $f^{n+}(\mathbf{x}) \to f^0(\mathbf{x})$ for every $\mathbf{x} \in D_{\mathbf{x}}$.

Proof. This is a consequence of Theorem 1 and Corollary 2. □

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 functions 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 $\int \int |g^\epsilon(\mathbf{x}) - g^*(\mathbf{x})| d\mathbf{x} < \epsilon''$. □

The results for the MEW model can be extended to the nested Dirichlet Process. In chapter 2 we provided 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 3. *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 $n_1, \ldots, n_J$ 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 $\{1, \ldots, J\}$. 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 size 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 in Theorem 1.

□

Corollary 4. *Each functional estimate arising from the model in 5.4 is consistent on a class that is dense on the integrable functions on a compact set* $D_{\mathbf{x}}$.

General consistency results for the class of dependent Dirichlet processes is 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.

5.4 Computational implementation

We implement the nDP model using the two-level truncation algorithm described in section 2. 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 $j = 1, \ldots, J$ from a multinomial distribution with probabilities

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

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

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

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

$$(u_k^* | \cdots) \sim \mathsf{beta}\left(1 + m_k, \alpha + \sum_{s=k+1}^{K} m_s\right), \quad k = 1, \ldots, K-1, \quad u_K^* = 1,$$

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

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

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

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

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) $\mathsf{NIW}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0)$.

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

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

where

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

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

$$(\boldsymbol{\Sigma}_0|\cdots) \sim \mathsf{W}\left(\gamma + c\nu_0, \gamma\boldsymbol{\Sigma}_0 0^{-1} + \nu_0\bar{\mathbf{D}}\right)$$

where c is the number of non-empty components.

8. Sample the mean precision parameter κ_0 from

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

9. Sample the concentration parameters α and β from

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

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

5.5 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 5.2 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 section 5.4. Hyperparameters were set according to the empirical distribution of the data, with $\boldsymbol{\theta}_{00}$ equal to the overall sample mean and $\boldsymbol{\Sigma}_{00}$ equal to the sample covariance matrix. For the other parameters associated with the baseline measure, we chose $\mathbf{D} = \boldsymbol{\Sigma}/100$, $a_\kappa = 1$ 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 we pick $a_\alpha = b_\alpha = a_\beta = b_\beta = 3$.

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 dendogram 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 5.1 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

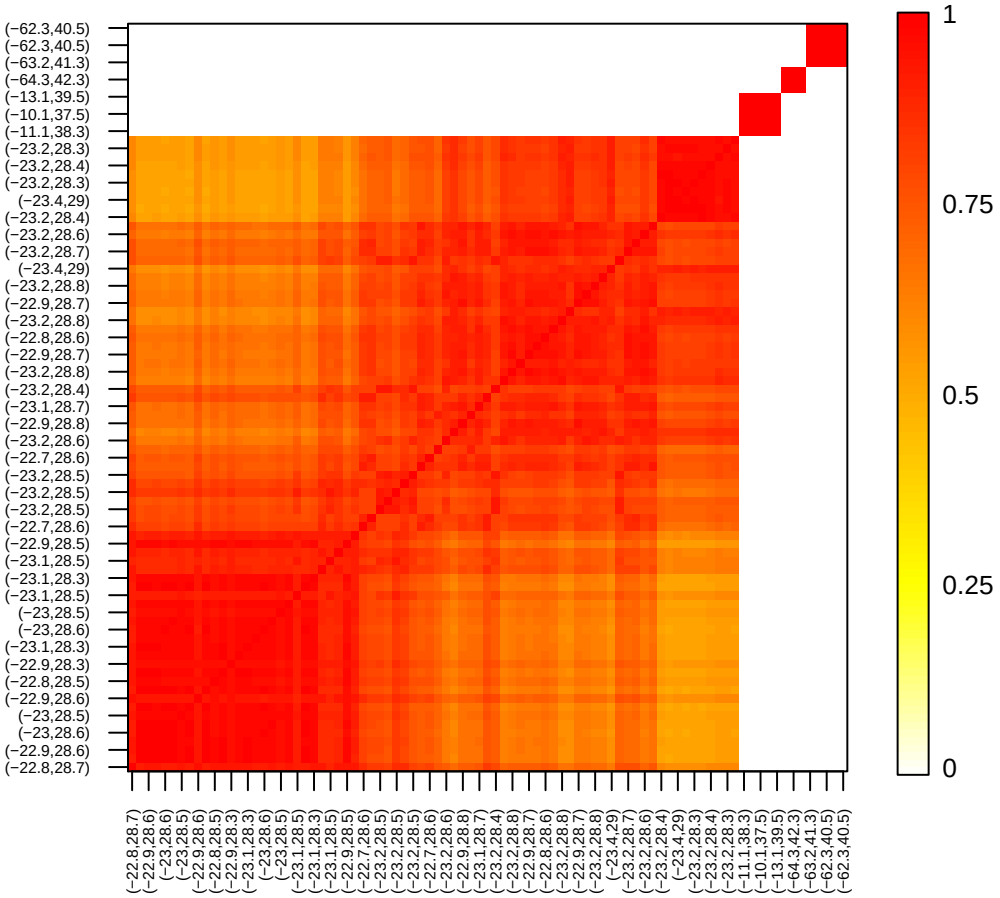


Figure 5.1: Heatmap with the probabilities of pairwise joint classification in the CTD data. Pixel (i, j) represents the posterior probability of locations i and j being clustered together.

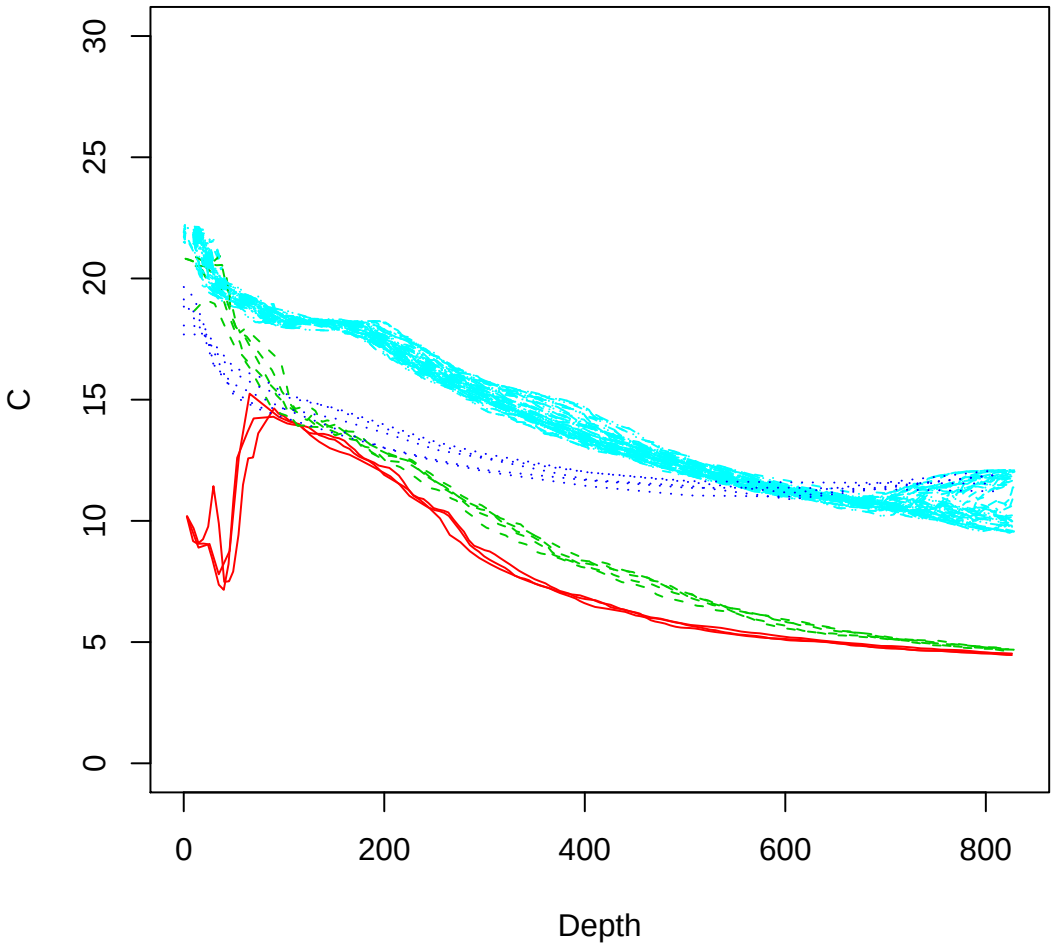


Figure 5.2: Raw profiles collected in the North Atlantic between June 15 and June 22, 1986. Colors indicate cluster membership.

observations each. Figures 5.2 and 5.3 show the raw curves and the location where the data 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 5.4 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.

The only unappealing feature of our functional estimates is the bump in the

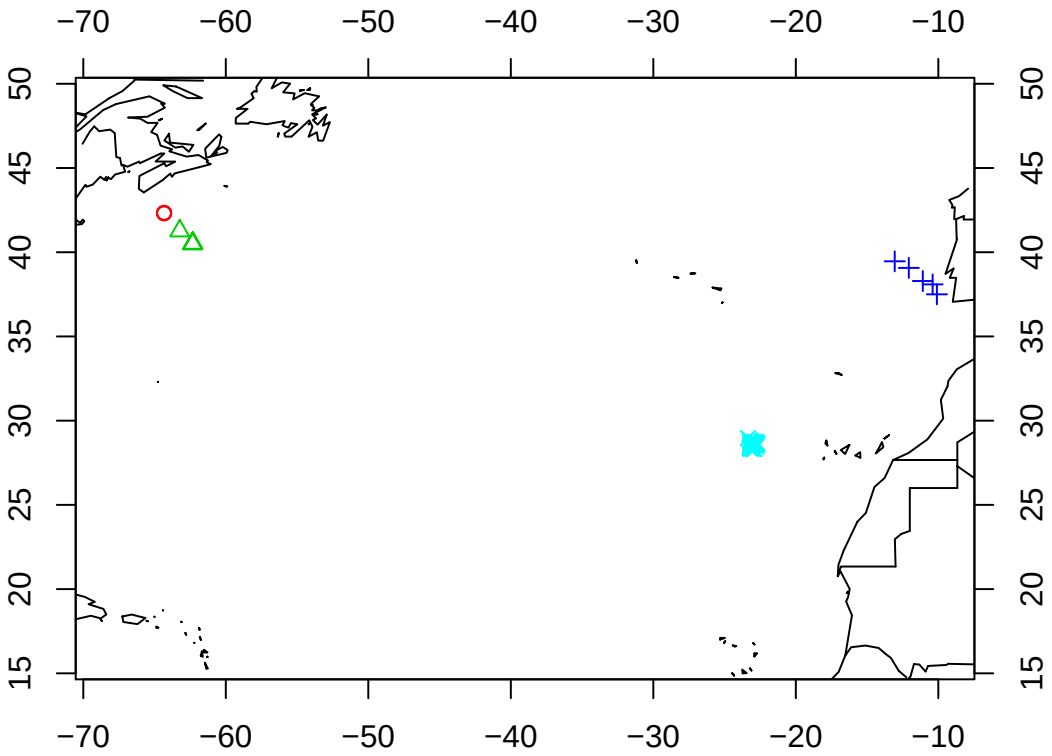


Figure 5.3: Geographic locations where the CTD data were collected. Colors indicate cluster membership.

dark blue curve appearing around 700m. This bump is due to the sparseness in the data off the Portuguese coast (in these locations, observations below the 300 m mark where collected only every 100 m). As was discussed in Section 5.3, 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.

5.6 A short study of racial differences in pregnancy outcomes

Understanding differences in health care outcomes across different ethnic groups is important, not only from an equality perspective, but also in terms of public policy and treatment design. Although race is a social construct, it can be helpful as a proxy for genetic, social and/or environmental factors that are hard or impossible to observe.

Of particular interest is how the relationship between gestational age of the fetus and its weight at delivery varies across ethnic groups, and especially the concept of small-for-gestational-age (SGA). Traditionally, SGA indicates that the baby is in the lower 10^{th} percentile for birth weight compared with a national reference for babies born at the same gestational age, which ignores factors such as race. However, as an individual with very small parents is more likely to be classified as SGA, differences in growth rates and body size among ethnic groups can lead to substantial misclassification. Since SGA babies are subject to a greater risk of infant mortality and short and long term morbidity, misclassification errors typically lead to unnecessary procedures and higher costs.

In this section, we develop a nonparametric regression model that allows us to:

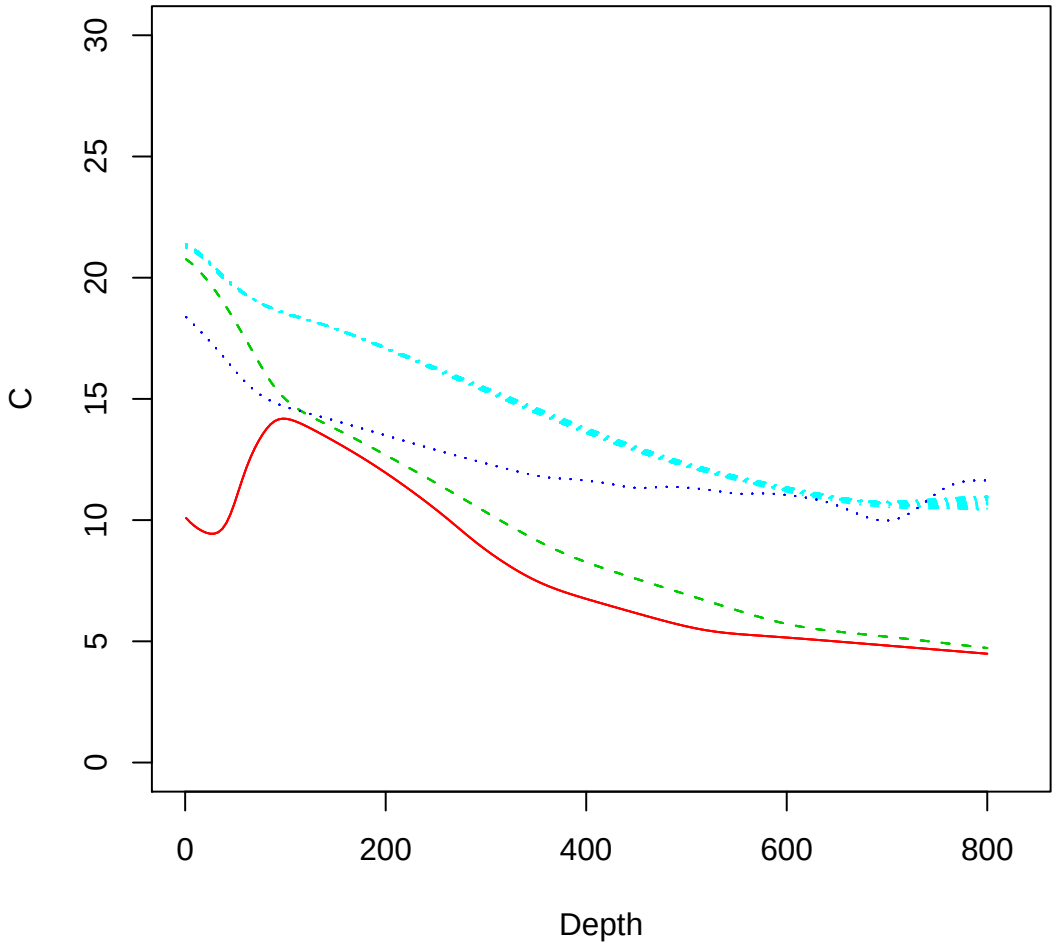


Figure 5.4: Fitted CTD curves obtained after model averaging. There are actually 87 distinct curves represented in the plot but, due to the tight cluster membership, most are undistinguishable. Colors indicate cluster membership.

a) obtain estimates of the joint distribution of gestational age and birth weight; b) perform mean and quantile regression of birth weight and gestational age, generating separate curves for each race while borrowing information across them; and c) test for evidence on the differences between races.

Our data set consists of a small subset from the National Collaborative Perinatal Project, conducted by NIH's National Institute of Neurological Diseases and Stroke between 1959 and 1974. It contains information about 1007 Caucasian births and 1187 African American births (adjudicated based on mother's race). Multiple births were excluded, but we were not restricted to first born children. Although the data set is old, it is very detailed, containing information on gestational age at the daily level. Birth weight of the infant is reported in kilograms, while gestational age (in weeks) is measured between the last menstrual period (LMP) and delivery.

We used a hierarchical Dirichlet process to model the collection of joint distributions of gestational age and birth weight for different races. The Pólya urn MCMC sampler described in Teh *et al.* (2006) was used to fit model. Posterior quantiles were obtained using the truncation method described in Kottas and Gelfand (2002). All results are based on 40,000 iterations, obtained after a burn-in period of 5,000 observations. As with our oceanographic example, the baseline probability measure is taken as $\mathsf{NIW}(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0)$. We fixed $\nu_0 = 3$ and took $\kappa_0 \sim \mathsf{Gam}(1, 100)$ and $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ to be the empirical mean and variance of the data. Finally, for the concentration parameters we took $\alpha \sim \mathsf{Gam}(1, 1)$ and $\beta \sim \mathsf{Gam}(1, 1)$.

Figures 5.5 and 5.6 show the raw data points together with the mean and 10% quantile regression curves for Caucasians and African-American populations. While the mean curve is useful to determine the weight for a normal child, the quantile curve identifies babies who are especially at risk. As expected, both types of curves are strictly increasing up to 42 weeks of gestation. However, for very long gestational

ages, the curves are first decreasing and then increasing again. This inconsistency might be due to the inaccuracies related to the use of the LMP as the sole indicator for gestation, especially for women with very irregular menstrual cycles. Since during the time of the study it was uncommon to induce delivery, this long gestational periods could also signal problems with the pregnancy.

Confidence bands, depicted as dashed lines, tend to be wider at the extremes (where less data is available). For early gestation times, African-American babies tend to be slightly larger than Caucasian babies. However, after 35 weeks (32 for the quantile curve), the relationship is inverted.

In order to explore the differences in growth rates in more detail, we present in Figures 5.7 and 5.8 the estimated posterior probabilities of the mean and quantile regression curves for African- Americans differing by 75 grams or more. These plots reveal mild evidence of no biologically significant difference between races below 37 weeks (the dip around 33 weeks being due to increased uncertainty about the shape of the function). On the other hand, they also provide very strong evidence that African-American children born between 37 and 43 weeks of gestations tend to be smaller than Caucasian babies. For gestations longer than 43 weeks we observe again an unexpected change, which might again be related to the use of LMP as an indicator of conception.

Our model provides another way to look for differences between the conditional distributions of birth weight given gestation. Since the set of basis functions are common to all curves, we can calculate the posterior probability that future draws from the conditional distribution of each race come from the same mixture component, which is shown in Figure 5.9. The pattern is similar to the one observed in the two figures above, which was to be expected. However, it is remarkable that the probability of common component draw is over 0.25 for the whole range of interest even

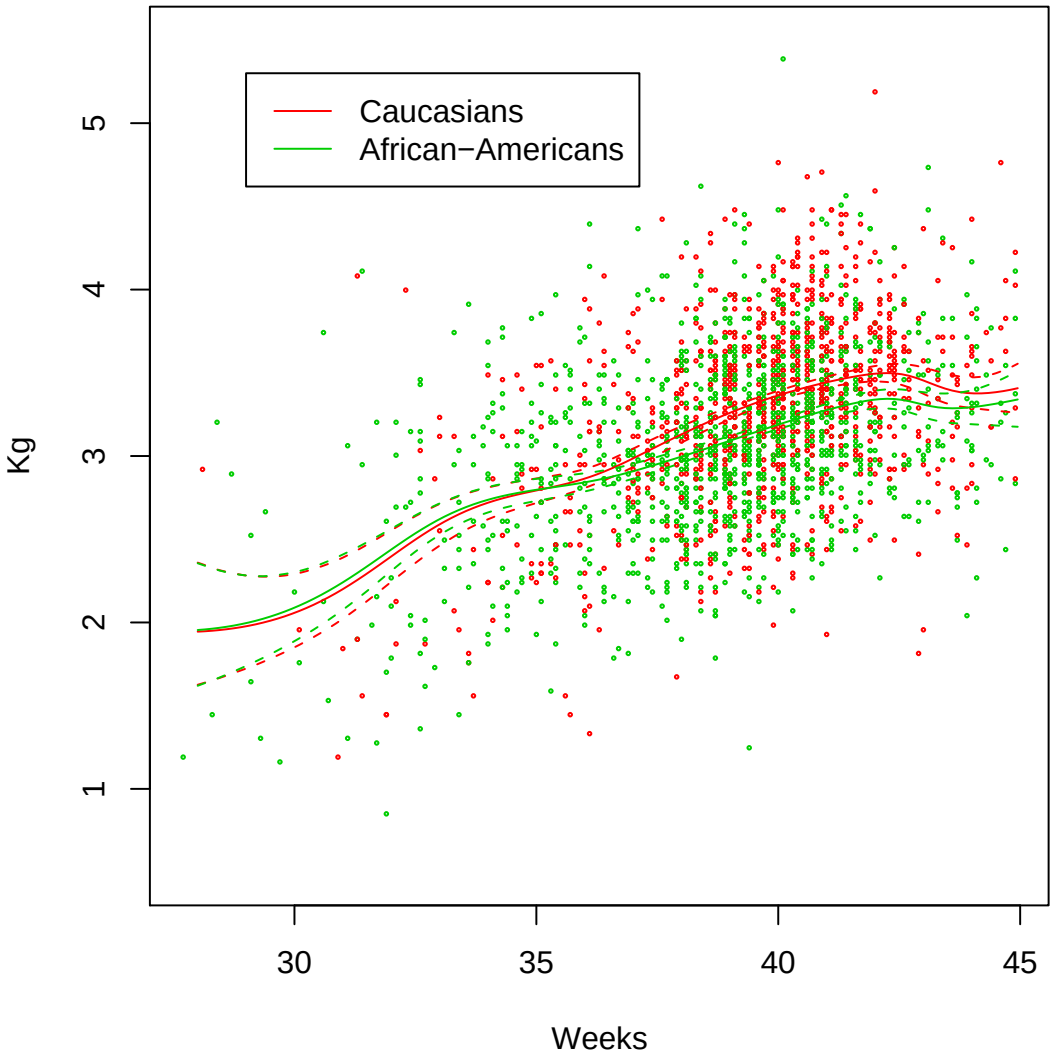


Figure 5.5: Estimated mean regression curves relating birth weight and gestational age in African-American and Caucasian populations. Dashed lines represent point-wise probability bands.

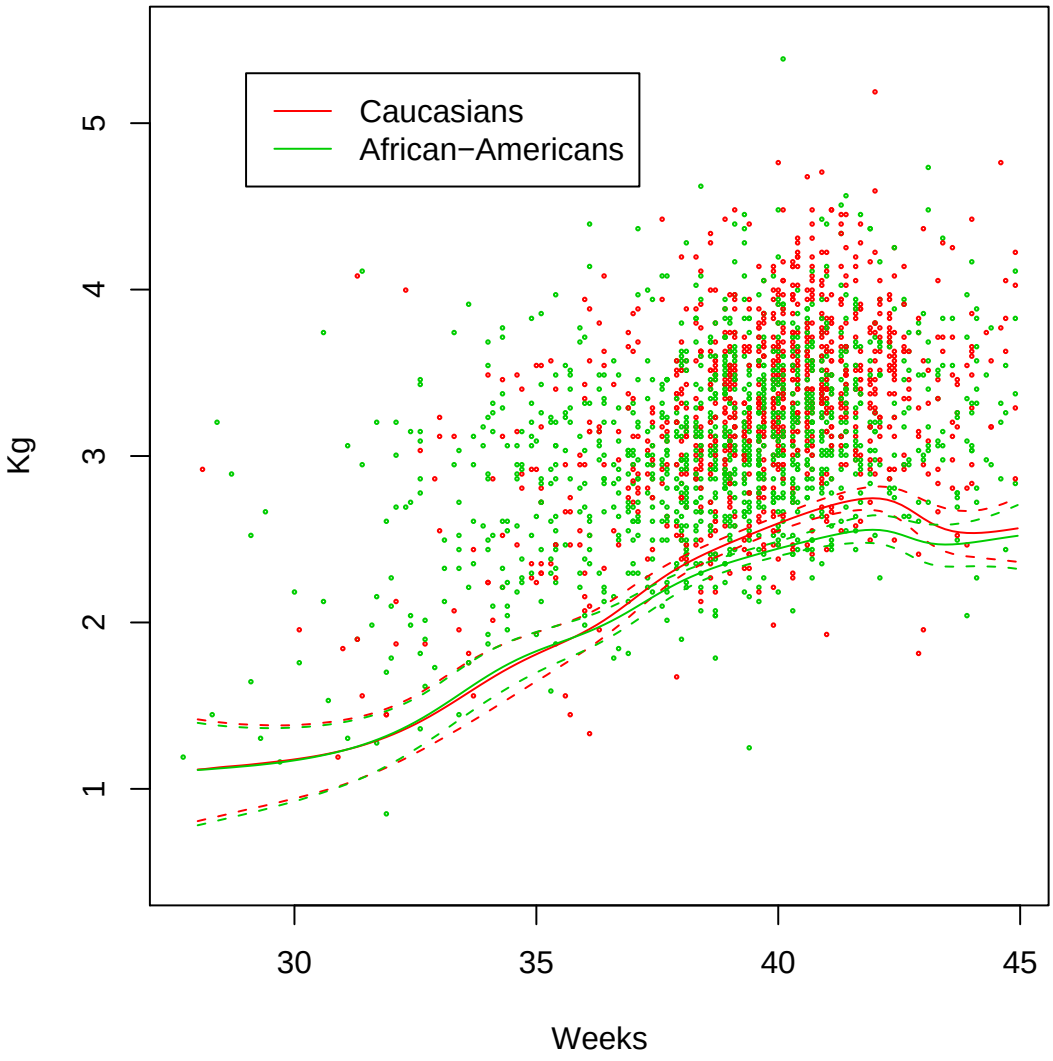


Figure 5.6: Ten percent quantile regression curves relating birth weight and gestational age in African-American and Caucasian populations. Dashed lines represent pointwise probability bands.

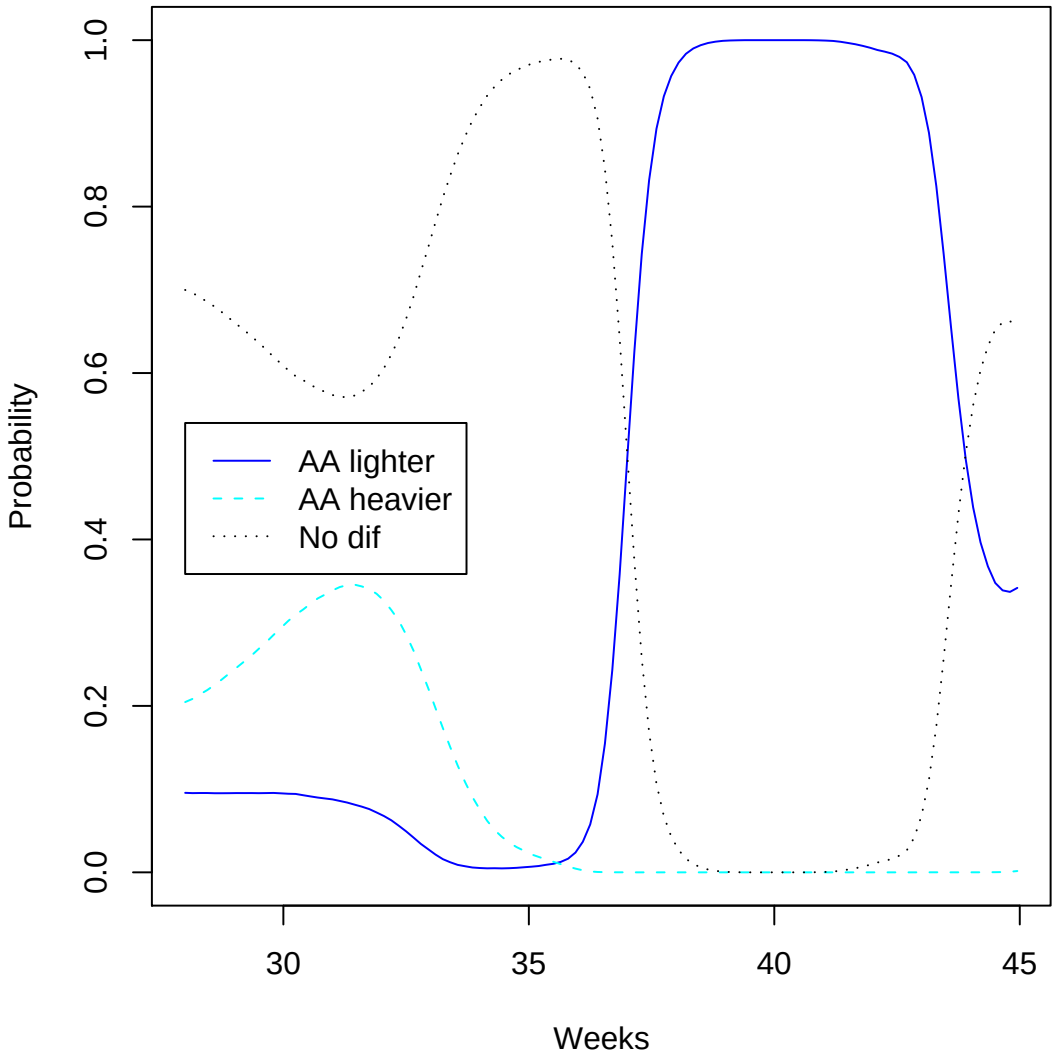


Figure 5.7: Probability that the birth weight of the average African-American (AA) child differs from the birth weight of the average Caucasian by more than 75 grams, as a function of gestational age.

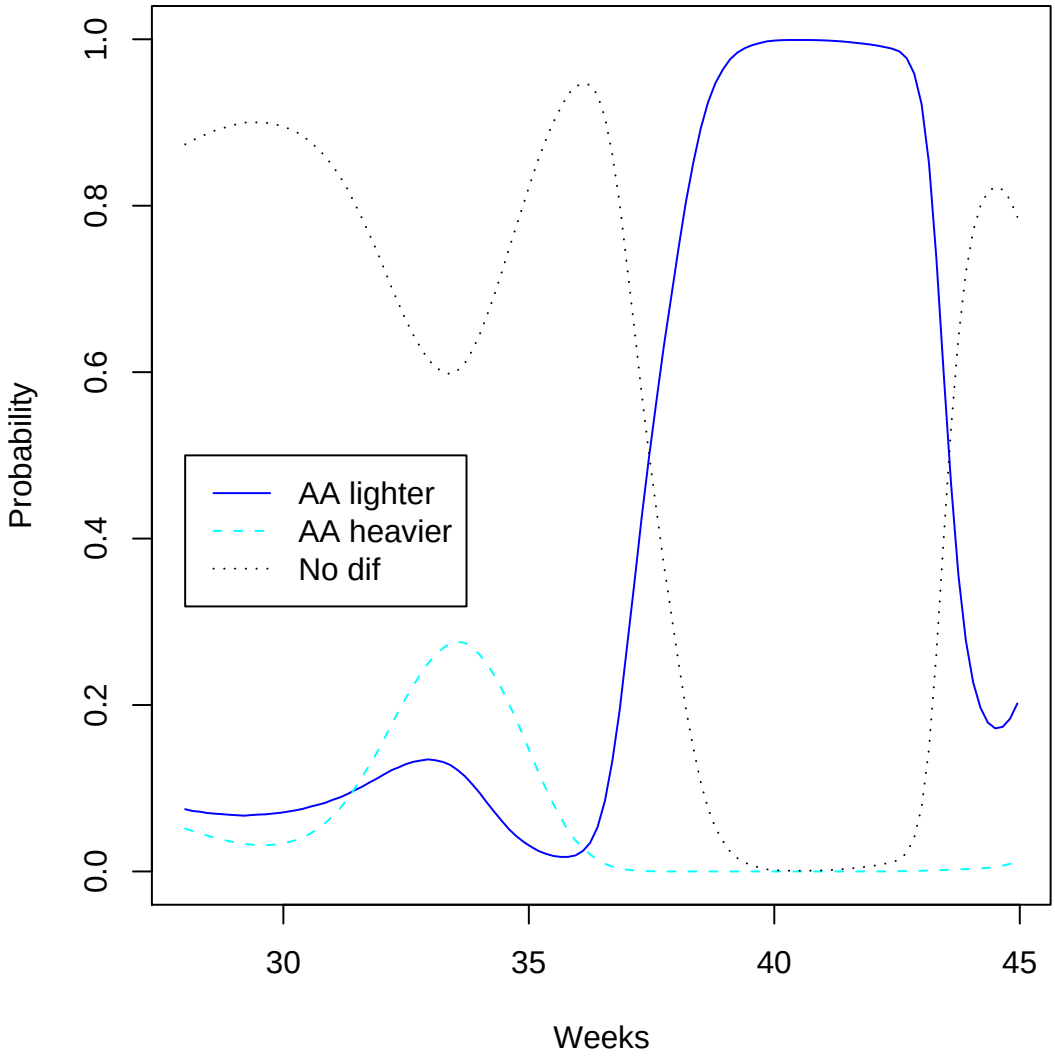


Figure 5.8: Probability that the birth weight of the 10%-quantile African-American (AA) child differs from that of the 10%-quantile Caucasian by more than 75 grams, as a function of gestational age.

when the probability of a large difference is almost one between 38 and 42 weeks. This probably due to very heavy tails in the conditional distributions between these two time points.

5.7 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 a very flexible function estimate. 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 and the hierarchical Dirichlet process as priors on the collection of mixing distributions that define the joint density of outcomes and predictors. The nDP 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.

The HDP model seems to produce interesting, if somewhat disconcerting, results. In particular, it is unclear whether the differences between races are due to differential exposure to risk factors across groups or to problems with the measurement of

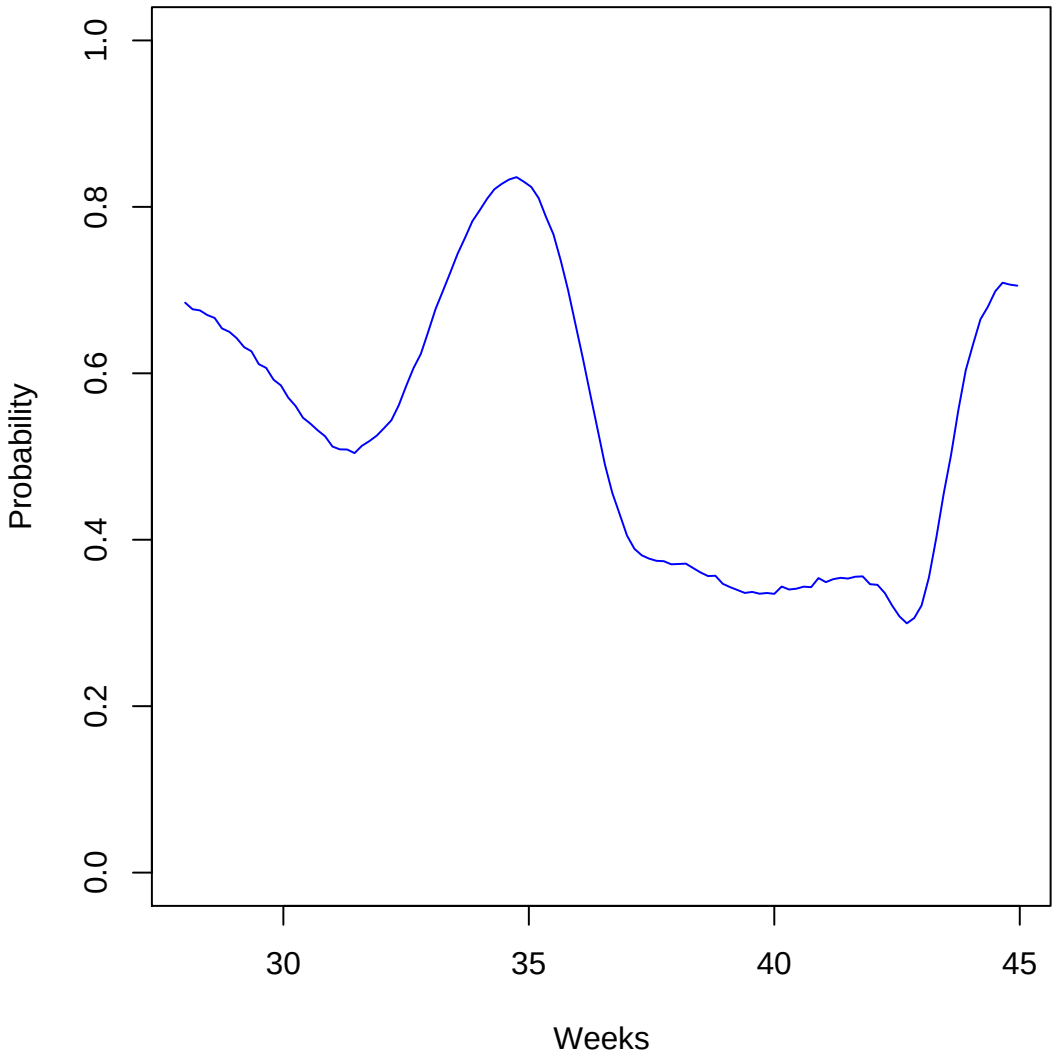


Figure 5.9: Probability that the conditional distributions for both races are represented using the same mixture component, as a function of gestational age.

gestational time. Since the data was collected between the mid fifties and the mid seventies (when ultrasound technology was not available), LMP was used as the only indicator to establish pregnancy and gestational age. This can be highly inaccurate, especially for women with irregular menstrual cycles, leading to inconsistencies in the data (note, for example, that pregnancies very rarely extend over 40 weeks because delivery is induced at or before this point). Also, it would be important to distinguish between male and female births in order to get a clear picture. Current and future work includes the analysis of more recent and accurate pediatric data sets using the methods described in this chapter.

Chapter 6

Spatial functional data analysis through spatially varying mixtures of normals

In chapter 5 we discussed approaches that allow us to use models for *collections of dependent distributions* to induce models for *collections of curves*, with a particular emphasis on problems where the curves are assumed exchangeable a priori. In this chapter we extend this methodology to account for spatial and/or temporal dependence among curves. The goal is to selectively borrow information, generating a prior structure that enforces curves closer in space to have a more similar shape.

Most of the literature on spatial functional data analysis is based on representations of the regression function as a linear combination of basis functions, such as splines or wavelets. The coefficients of the basis expansion are then modeled in one of two ways: as some parametric or nonparametric function of space (Fahrmeir and Lang, 2001; Fahrmeir *et al.*, 2004) or as a realization from a (possible multivariate) Gaussian process (Assuncao, 2003; Banerjee *et al.*, 2004). The main practical issue with these approaches is the selection of the basis functions, which can dramatically affect bias, smoothness and efficiency.

In this chapter, we elaborate on a completely different approach based on density estimation techniques that use dependent Dirichlet processes (MacEachern, 2000).

As in chapter 5, we assume that the joint distribution of outcomes and predictors at any given location can be well represented as a countable mixture of normals. The value of desired function for any value of the predictor can be recovered from this joint distribution by computing the corresponding conditional expectation, which takes the form of a predictor-dependent mixture of linear terms. Then, we induce space/time dependence in the collection of distributions by building dependence in the mean and variances of the normal components, while keeping the weights of the components fixed. The resulting model defines a stochastic process on the space of curves with an index space $D \subset \mathbb{R}^d$ that allows us not only to estimate the function value at unobserved locations, but also to interpolate/predict functional forms at unknown locations.

Spatially dependent Dirichlet processes were originally proposed by Gelfand *et al.* (2005), who use a Gaussian baseline process to generate a global surface-selection mechanism. Duan *et al.* (2007) extended these ideas to allow local surface selection, and Gelfand *et al.* (2007) consider the relationship between finite and infinite versions of the process. In practical applications, these models have been restricted to nonparametric analogs of the Gaussian process, which, for the purpose of our application, means that only the locations of the normal components are allowed to vary in space. This is clearly too restrictive for the purpose of functional data analysis because it means that the slope of each linear piece in the function is the same. One important contribution of this paper is to construct a nonparametric process on the space of positive-definite matrices that allows us to create a model that is rich enough to capture complicated functional forms.

This chapter is organized as follows: Section 6.1 describes the model and reviews some of its properties. Section 6.2 develops a MCMC sampler based on a truncated version of the process. In section 6.3 we demonstrate the advantages of the model

using a simulation study. Finally, section 6.4 provides a short discussion and some future directions for our work.

6.1 Model specification

Our approach to nonparametric regression relies on a flexible multivariate stochastic process

$$\{\mathbf{z}(\mathbf{s}) = (\mathbf{x}(\mathbf{s}), \mathbf{y}(\mathbf{s})) : \mathbf{s} \in D \subset \mathbb{R}^d\}$$

that determines the joint distribution of both regressors $\mathbf{x} \in \mathbb{R}^p$ and outcomes $\mathbf{y} \in \mathbb{R}^q$ at every possible location $\mathbf{s} \in D$. In the sequel, we assume that a collection of observations $\{\mathbf{z}_j(\mathbf{s}_i) = (\mathbf{y}_j(\mathbf{s}_i), \mathbf{x}_j(\mathbf{s}_i))\}_{j=1}^{n(\mathbf{s}_i)}$ has been collected at each site $\mathbf{s}_i$, for $i = 1, \ldots, m$.

We model the stochastic process $\mathbf{z}(\mathbf{s})$ in a hierarchical fashion. At the highest level of the hierarchy, we assume that the observations at a given location $\mathbf{s}$ come from an (infinite) mixture of normals,

$$\mathbf{z}_j(\mathbf{s}) \sim \int \phi_{p+q}(\mathbf{z}_j(\mathbf{s})|\boldsymbol{\mu}(\mathbf{s}), \boldsymbol{\Sigma}(\mathbf{s}))G_s(\boldsymbol{\mu}(\mathbf{s}), \boldsymbol{\Sigma}(\mathbf{s}))$$

$$G_s(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{(\mu_k^*(\mathbf{s}), \Sigma_k^*(\mathbf{s}))}(\cdot)$$

where $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

In the single-location case, this type of scale-mixture of normals provide a model that is dense in the space of absolutely continuous distributions (Lo, 1984). Functional estimates recovered from this model are covariate-weighted mixtures of linear functions, taking the form,

$$\begin{aligned} f_{\mathbf{s}}^m(\mathbf{x}) &= \mathbb{E}(\mathbf{y}|\mathbf{x}, \mathbf{z}^m) \\ &= \int \frac{(\boldsymbol{\mu}_y(\mathbf{s}) + \boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})(\mathbf{x} - \boldsymbol{\mu}_x(\mathbf{s})))\phi_p(\mathbf{x}|\boldsymbol{\mu}_x(\mathbf{s}), \boldsymbol{\Sigma}_{xx}(\mathbf{s}))}{\int \phi_p(\mathbf{x}|\boldsymbol{\mu}_x(\mathbf{s}), \boldsymbol{\Sigma}_{xx}(\mathbf{s}))G_{\mathbf{s}0}^m(d\boldsymbol{\mu}(\mathbf{s}), d\boldsymbol{\Sigma}(\mathbf{s})|\mathbf{z}^m)} G_{\mathbf{s}0}^m(d\boldsymbol{\mu}(\mathbf{s}), d\boldsymbol{\Sigma}(\mathbf{s})|\mathbf{z}^m) \end{aligned}$$

where $\mathbf{z}^m = \{\{\mathbf{z}_j(\mathbf{s}_i)\}_{j=1}^{n(\mathbf{s}_i)}\}_{i=1}^m$ is the collection of all observed values, $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the density of a p-variate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ and $G_{\mathbf{s}0}^m$ is the posterior expected mixing distribution at location $\mathbf{s}$. By allowing the distribution of the parameters of the mixture components to depend on space, we allow the regression function to adapt.

We assume that the collection of spatially-changing mixing distributions $\mathcal{G} = \{G_{\mathbf{s}} : \mathbf{s} \in D\}$ is unknown and assign $\mathcal{G}$ a dependent Dependent process, where $w_k = v_k \prod_{l<k}(1 - v_l)$, $v_k \sim \mathsf{beta}(1, \alpha)$ and the atoms are iid samples following

$$\mu_{kl}^*(\mathbf{s}) \sim \mathsf{GP}(\mu_{0l}, \tau_l^2, \rho_l^{\mu}(\cdot, \cdot|\boldsymbol{\eta}_l^{\mu_l})) \;\; l = 1, \ldots, p+q$$

$$\boldsymbol{\Sigma}_k^{*-1}(\mathbf{s}) = \mathbf{U}\mathbf{B}_k(\mathbf{s})\mathbf{B}_k'(\mathbf{s})\mathbf{U}'$$

$$[\mathbf{B}_k(\mathbf{s})]_{ij} \sim \mathsf{GP}(0, 1, \rho_j^b(\cdot, \cdot|\boldsymbol{\eta}_j^b)) \;\; j = 1, \ldots, p+q, \;\; i = 1, \ldots, \nu_b$$

where $\mathsf{GP}(\mu, \tau^2, \gamma)$ denotes a Gaussian process with mean μ, variance τ^2 and correlation function γ, ρ_l^{μ} and ρ_j^b are correlation functions known up to parameters $\boldsymbol{\eta}_l^{\mu}$ and $\boldsymbol{\eta}_j^b$. In our applications below, we assume exponential correlation functions with unknown range parameter, but other choices are straightforward. In order to construct spatially varying precision matrices, we borrow the idea for spatial Wishart processes from Gelfand *et al.* (2004). Since $[\mathbf{B}(\mathbf{s})]_{ij} \sim \mathsf{N}(0, 1)$ marginally for $j = 1, \ldots, p+q$ and $i = 1, \ldots, \nu_b$, $\boldsymbol{\Sigma}_k^{*-1}(\mathbf{s})|\mathbf{U}$ follows a Wishart distribution with parameters ν_b and $\mathbf{U}\mathbf{U}'$ for each location $\mathbf{s} \in D$. Therefore, our model is marginally equivalent to the single curve in Müller *et al.* (1996).

Our model is completed by specifying prior distributions for the hyperparameters in the model. For the location of the baseline processes we choose

$$\boldsymbol{\mu}_0 \sim \mathsf{N}(\boldsymbol{\mu}_{00}, \boldsymbol{\Omega})$$

$$\mathbf{U}\mathbf{U}' \sim \mathsf{W}(\gamma, \mathbf{S}_{00})$$

where $\mathbf{U}$ is a lower diagonal matrix with positive diagonal entries. For the prior variances of the location parameters we set $\tau_l^2 \sim \mathsf{IGam}(a_\tau, b_\tau)$. Finally, for the parameters

governing the spatial dependence in the process,

$$\boldsymbol{\eta}_l^{\mu} \sim p_l^{\mu}(\boldsymbol{\eta}_l^{\mu})$$

$$\boldsymbol{\eta}_l^{b} \sim p_l^{b}(\boldsymbol{\eta}_l^{b})$$

for some suitable distributions p_j^w and p_j^w. For the examples below , we use Gamma priors on the range of the correlation functions.

It is interesting to consider the spatial covariance structure induced by our model. For the single component model, obtained by letting $\alpha \to 0$, we obtain

$$[\mathbf{y}(\mathbf{s})|\mathbf{x}(\mathbf{s}) = \mathbf{x}_0] \sim \mathsf{N}\left(\boldsymbol{\mu}_y(\mathbf{s}) + \boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})(\mathbf{x}_0 - \boldsymbol{\mu}_x(\mathbf{s})), \boldsymbol{\Sigma}_{yy}(\mathbf{s}) - \boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})\boldsymbol{\Sigma}_{xy}(\mathbf{s})\right)$$

The spatial covariance between responses $\mathbf{y}(\mathbf{s})$ and $\mathbf{y}(\mathbf{s}')$ collected at locations $\mathbf{s}$ and $\mathbf{s}'$ conditionally on predictor values $\mathbf{x}(\mathbf{s}) = \mathbf{x}_0$ and $\mathbf{x}(\mathbf{s}') = \mathbf{x}_0'$ is

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}(\mathbf{y}(\mathbf{s}), \mathbf{y}(\mathbf{s}')|\mathbf{x}(\mathbf{s}) = \mathbf{x}_0, \mathbf{x}(\mathbf{s}') = \mathbf{x}_0') = \\
&\quad = \mathbb{E}(\mathbb{C}\mathrm{ov}(\mathbf{y}(\mathbf{s}), \mathbf{y}(\mathbf{s}')|\mathbf{x}(\mathbf{s}) = \mathbf{x}_0, \mathbf{x}(\mathbf{s}') = \mathbf{x}_0', \boldsymbol{\mu}(\mathbf{s}), \boldsymbol{\mu}(\mathbf{s}'), \boldsymbol{\Sigma}(\mathbf{s}), \boldsymbol{\Sigma}(\mathbf{s}'))) + \\
&\qquad \mathbb{C}\mathrm{ov}(\mathbb{E}(\mathbf{y}(\mathbf{s})|\mathbf{x}(\mathbf{s}) = \mathbf{x}_0, \boldsymbol{\mu}(\mathbf{s}), \boldsymbol{\Sigma}(\mathbf{s})), \mathbb{E}(\mathbf{y}(\mathbf{s}')|\mathbf{x}(\mathbf{s}') = \mathbf{x}_0', \boldsymbol{\mu}(\mathbf{s}'), \boldsymbol{\Sigma}(\mathbf{s}'))) \\
&\quad = \mathbb{C}\mathrm{ov}(\boldsymbol{\mu}_y(\mathbf{s}) + \boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})(\mathbf{x}_0 - \boldsymbol{\mu}_x(\mathbf{s})), \boldsymbol{\mu}_y(\mathbf{s}') + \boldsymbol{\Sigma}_{yx}(\mathbf{s}')\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s}')(\mathbf{x}_0' - \boldsymbol{\mu}_x(\mathbf{s}'))) \\
&\quad = \mathbb{C}\mathrm{ov}(\boldsymbol{\mu}_y(\mathbf{s}), \boldsymbol{\mu}_y(\mathbf{s}')) + \mathbb{C}\mathrm{ov}(\boldsymbol{\mu}_y(\mathbf{s}), \boldsymbol{\Sigma}_{yx}(\mathbf{s}')\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s}')(\mathbf{x}_0' - \boldsymbol{\mu}_x(\mathbf{s}'))) + \\
&\qquad \mathbb{C}\mathrm{ov}(\boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})(\mathbf{x}_0 - \boldsymbol{\mu}_x(\mathbf{s})), \boldsymbol{\mu}_y(\mathbf{s}')) + \\
&\qquad \mathbb{C}\mathrm{ov}(\boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})(\mathbf{x}_0 - \boldsymbol{\mu}_x(\mathbf{s})), \boldsymbol{\Sigma}_{yx}(\mathbf{s}')\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s}')(\mathbf{x}_0' - \boldsymbol{\mu}_x(\mathbf{s}'))) \\
&\quad = \mathbb{C}\mathrm{ov}(\boldsymbol{\mu}_y(\mathbf{s}), \boldsymbol{\mu}_y(\mathbf{s}')) - \mathbb{C}\mathrm{ov}(\boldsymbol{\mu}_y(\mathbf{s}), \boldsymbol{\mu}_x(\mathbf{s}'))\mathbb{E}(\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s}')\boldsymbol{\Sigma}_{xy}(\mathbf{s}')) - \\
&\qquad \mathbb{E}(\boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s}))\mathbb{C}\mathrm{ov}((\boldsymbol{\mu}_x(\mathbf{s}), \boldsymbol{\mu}_y(\mathbf{s}')) + \\
&\qquad \mathbb{C}\mathrm{ov}(\boldsymbol{\Sigma}_{yx}(\mathbf{s})\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s})(\mathbf{x}_0 - \boldsymbol{\mu}_x(\mathbf{s})), \boldsymbol{\Sigma}_{yx}(\mathbf{s}')\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{s}')(\mathbf{x}_0' - \boldsymbol{\mu}_x(\mathbf{s}')))
\end{aligned}$$

Although a closed form is not available even in this simple setting, it is clear that the process is not separable and not stationary (even if the underlying processes are), either in space or covariates. Also, terms making up the expression have a

straightforward interpretation. The covariance between the mean values of $\mathbf{y}(\mathbf{s})$ and $\mathbf{y}(\mathbf{s}')$ (which corresponds to the first term in the sum) is adjusted by the dependence between the mean values of $\mathbf{y}(\mathbf{s})$ and $\mathbf{x}(\mathbf{s}')$ (represented by the second term) and $\mathbf{x}(\mathbf{s})$ and $\mathbf{y}(\mathbf{s}')$ (given by the third term), as well as by the actual values of $\mathbf{x}_0$ and $\mathbf{x}_0'$. For the infinite mixture model, the expression is even more complicated because it includes the weights of the components. However, it is again clear that the functional process is nonstationary and nonseparable.

6.2 Inference

We apply an MCMC algorithm that truncates the stick-breaking process to a predetermined level K and uses a finite-mixture sampler. The validity of this type of scheme as a reasonable approximation to the infinite mixture model has been argued in Gelfand *et al.* (2007) and Petrone *et al.* (2006). Alternatively, the algorithm presented here can be easily used as the basis for a retrospective sampler Roberts and Papaspiliopoulos (2007). After choosing starting values for each parameter in the model, the algorithm proceeds by updating blocks of parameters one at a time through the following steps:

1. Sample the component probabilities w_k by generating the stick-breaking ratios

$$(v_k|\cdots) \sim \mathsf{beta}\left(1+n_k, \alpha+\sum_{r=k+1}^{K} n_r\right) \quad k=1,\ldots,v_{K-1}$$

   where $n_k = \sum_{i=1}^{m} n_k(\mathbf{s}_i)$ is the total number of observations assigned to component k. From these ratios, the weights are reconstructed as $w_k = v_k \prod_{r=1}^{l-1}(1-v_r)$ where $v_K = 1$.

2. The group indicators are sampled from a discrete distribution such that

$$\mathbb{P}(\xi_j(\mathbf{s}_i)=k|\mathbf{z},\ldots) \propto w_k\ \phi_{p+q}(\mathbf{z}_j(\mathbf{s}_i)|\boldsymbol{\mu}_k^*(\mathbf{s}_i), \boldsymbol{\Sigma}_k^*(\mathbf{s}_i)) \qquad k=1,\ldots,K$$

3. Sample the mean vectors $\boldsymbol{\mu}_k^* = (\boldsymbol{\mu}_k^*(s_1), \ldots, \boldsymbol{\mu}_k^*(s_m))'$. The conditional prior given $\boldsymbol{\mu}_0$, $\{\tau_l^2\}_{l=1}^p$ and $\{\rho_l^\mu\}_{l=1}^p$ can be written as $\boldsymbol{\mu}_k^* \sim \mathsf{N}(\mathbf{d}, \mathbf{D})$, where $\mathbf{d} = \mathbf{1}_m \otimes \boldsymbol{\mu}_0$, $\mathbf{1}_m$ is a column vector of length m, $A \otimes B$ denotes the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{D} = \sum_{l=1}^{p+q} \tau_l^2 \mathbf{Q}_l \otimes \mathbf{T}_l$, $\mathbf{Q}_l$ is a $m \times m$ matrix such that $[\mathbf{Q}_l]_{rt} = \rho_l^\mu(\mathbf{s}_r, \mathbf{s}_t|\boldsymbol{\eta}_l^\mu)$, and $\mathbf{T}_j = \mathbf{e}_j\mathbf{e}_j'$, where $\mathbf{e}_j$ is a vector of length $p+q$ such that

$$[\mathbf{e}_j]_l = \begin{cases} 1 & \text{if } j = l \\ 0 & \text{otherwise} \end{cases}$$

Through sufficiency, the likelihood reduces to

$$p(\bar{\mathbf{z}}_k|\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*) \propto \exp\left\{-\frac{1}{2}(\bar{\mathbf{z}}_k - \boldsymbol{\mu}_k^*)'\boldsymbol{\Sigma}_k^{*-1}(\bar{\mathbf{z}}_k - \boldsymbol{\mu}_k^*)\right\}$$

where $\bar{\mathbf{z}}_k = (\bar{\mathbf{z}}_k(s_1), \ldots, \bar{\mathbf{z}}_k(s_m))$, $\bar{\mathbf{z}}_k(\mathbf{s}_i) = \sum_{\{j:\zeta_j(\mathbf{s}_i)=k\}} \mathbf{z}_j(\mathbf{s}_i)/n_k(\mathbf{s}_i)$ is the average of all observations collected at site $\mathbf{s}_k$, $n_k(\mathbf{s}_i)$ is the number of observations in site $\mathbf{s}_i$ assigned to component k, and

$$\boldsymbol{\Sigma}_k^* = \text{Bldiag}\{\boldsymbol{\Sigma}(s_1)/n_k(\mathbf{s}_1), \ldots, \boldsymbol{\Sigma}(s_m)/n_k(\mathbf{s}_m)\}.$$

Therefore, the posterior reduces to

$$(\boldsymbol{\mu}_k^*|\mathbf{z}, \ldots) \sim \mathsf{N}\left([\mathbf{D}^{-1} + \boldsymbol{\Sigma}_k^{*-1}]^{-1}[\mathbf{D}^{-1}\mathbf{d} + \boldsymbol{\Sigma}_k^{*-1}\bar{\mathbf{z}}_k], [\mathbf{D}^{-1} + \boldsymbol{\Sigma}_k^{*-1}]^{-1}\right)$$

for $k = 1, \ldots, K$.

4. Sample each of the entries $[\mathbf{B}_k(\mathbf{s}_i)]_{rt}$ separately for every location using a random-walk Metropolis Hastings algorithm with proposal distribution

$$[\mathbf{B}_k(\mathbf{s}_i)]_{rt}^{(p)} \sim \mathsf{N}\left([\mathbf{B}_k(\mathbf{s}_i)]_{rt}^{(c)}, \kappa^2\right)$$

where $[\mathbf{B}_k(\mathbf{s}_i)]_{rt}^{(c)}$ and $[\mathbf{B}_k(\mathbf{s}_i)]_{rt}^{(p)}$ correspond to the current and proposed value of the (r,t) entry of matrix $\mathbf{B}(\mathbf{s}_i)$ and κ^2 is a tuning parameter controlling the

variance of the proposal around the current value. The likelihood term is

$$p(\mathbf{z}|\boldsymbol{\mu}_k, \mathbf{B}_k, \mathbf{U}) \propto \left[\prod_{i=1}^{m} |\mathbf{B}_k(\mathbf{s}_i)\mathbf{B}_k'(\mathbf{s}_i)|^{1/2}\right] \times$$

$$\exp\left\{-\frac{1}{2}\sum_{\{j:\xi_j(\mathbf{s}_i)=k\}} (\mathbf{z}_j(\mathbf{s}_i) - \boldsymbol{\mu}_k^*(\mathbf{s}_i))'\mathbf{U}\mathbf{B}_k(\mathbf{s}_i)\mathbf{B}_k'(\mathbf{s}_i)\mathbf{U}'(\mathbf{z}_j(\mathbf{s}_i) - \boldsymbol{\mu}_k^*(\mathbf{s}_i))\right\}$$

while the prior is given by

$$[\mathbf{B}_k(\mathbf{s}_i)]_{rt} \mid [\mathbf{B}_k(\mathbf{s}_{(-i)})]_{rt} \sim$$

$$\mathsf{N}\left(\boldsymbol{\Psi}_{i,-i}^r[\boldsymbol{\Psi}_{-i,-i}^r]^{-1}[\mathbf{B}_k(\mathbf{s}_{(-i)})]_{rt}, \boldsymbol{\Psi}_{i,i}^r - \boldsymbol{\Psi}_{i,-i}^r[\boldsymbol{\Psi}_{-i,-i}^r]^{-1})\boldsymbol{\Psi}_{-i,i}^r\right)$$

where $-i$ subscript indicated a vector with the i-th component (or location) removed and $\boldsymbol{\Psi}_{ut}^r = \rho_r(\mathbf{s}_u, \mathbf{s}_t|\boldsymbol{\eta}_r^\mu)$ is partitioned in four blocks that separate the terms corresponding to the i-th location from the others.

5. Sample each entry of $\mathbf{U}$ using another random-walk Metropolis step. Due to the sign constrains implicit in the construction of $\mathbf{U}$, proposals for its diagonal elements are made from a log-normal distribution, while off-diagonal entries are proposed from a normal distribution. The posterior distribution is proportional to

$$p(\mathbf{U}|\cdots) \propto |\mathbf{U}|^{\nu-p-1+\sum_{i=1}^m n(\mathbf{s}_i)} \exp\left\{-\frac{1}{2}\operatorname{tr} S_{00}^{-1}\mathbf{U}\mathbf{U}'\right\}$$

$$\exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\sum_{j}^{n(\mathbf{s}_i)} \mathbf{e}_j(\mathbf{s}_i)'\mathbf{U}\mathbf{B}_{\xi_j(\mathbf{s}_i)}(\mathbf{s}_i)\mathbf{B}_{\xi_j'(\mathbf{s}_i)}(\mathbf{s}_i)\mathbf{U}'\mathbf{e}_j(\mathbf{s}_i)\right\}$$

where $\mathbf{e}_j(\mathbf{s}_i) = \mathbf{z}_j(\mathbf{s}_i) - \boldsymbol{\mu}_{\xi_j(\mathbf{s}_i)}^*(\mathbf{s}_i)$.

6. Sample the concentration parameter α from

$$(\alpha|\cdots) \sim \mathsf{Gam}\left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1}\log(1-v_k)\right)$$

7. For $l = 1, \ldots, p+q$, sample τ_l^2, the prior variances of the mean parameters from

$$\tau_l^2|\mathbf{z}, \cdots \sim$$

$$\mathsf{IGam}\left(a_\tau + \frac{(p+q)K^*}{2}, b_\tau + \frac{1}{2}\sum_{\{k:n_k>0\}}(\boldsymbol{\mu}_{kl}^* - \mu_{0l}\mathbf{1}_m)'\mathbf{Q}_l^{-1}(\boldsymbol{\mu}_{kl}^* - \mu_{0l}\mathbf{1}_m)\right)$$

where $K^* = \#\{k : n_k > 0\}$ is the number of components with observations assigned to them, $\boldsymbol{\mu}_{kl}^* = (\mu_{kl}^*(\mathbf{s}_1), \ldots, \mu_{kl}^*(\mathbf{s}_m))$ is a subvector of the entries corresponding to the l-th coordinate on the k-th component and $\mathbf{1}_m$ is a vector of ones of length m.

8. Sample the spatial correlation parameters for the mean, $\boldsymbol{\eta}_l^\mu$, using a random walk metropolis-algorithm. The exact form of the proposals depends on the specific correlation function and the prior chosen, but the posterior distribution is given by,

$$p(\boldsymbol{\eta}_l^\mu|\cdots) \propto$$

$$|\mathbf{Q}_l|^{-K^*/2}\exp\left\{-\frac{1}{2\tau_l^2}\sum_{\{k:n_k>0\}}(\boldsymbol{\mu}_{kl}^* - \mu_{0l}\mathbf{1}_m)'\mathbf{Q}_l^{-1}(\boldsymbol{\mu}_{kl}^* - \mu_{0l}\mathbf{1}_m)\right\}p_l^\mu(\boldsymbol{\eta}_i^\mu)$$

9. Similarly, sample the spatial correlation parameters for the mean, $\boldsymbol{\eta}_l^b$, using a random walk metropolis-algorithm.

6.3 A simulation example

We present in this section a simulation example designed to demonstrate the versatility of the model. We randomly picked 20 locations on the unit square, shown in black and labeled 1 to 20 in Figure 6.1. We sampled values of the function at 50 random points for each location. For each location $\mathbf{s}$, the $x(\mathbf{s})$ coordinates were picked

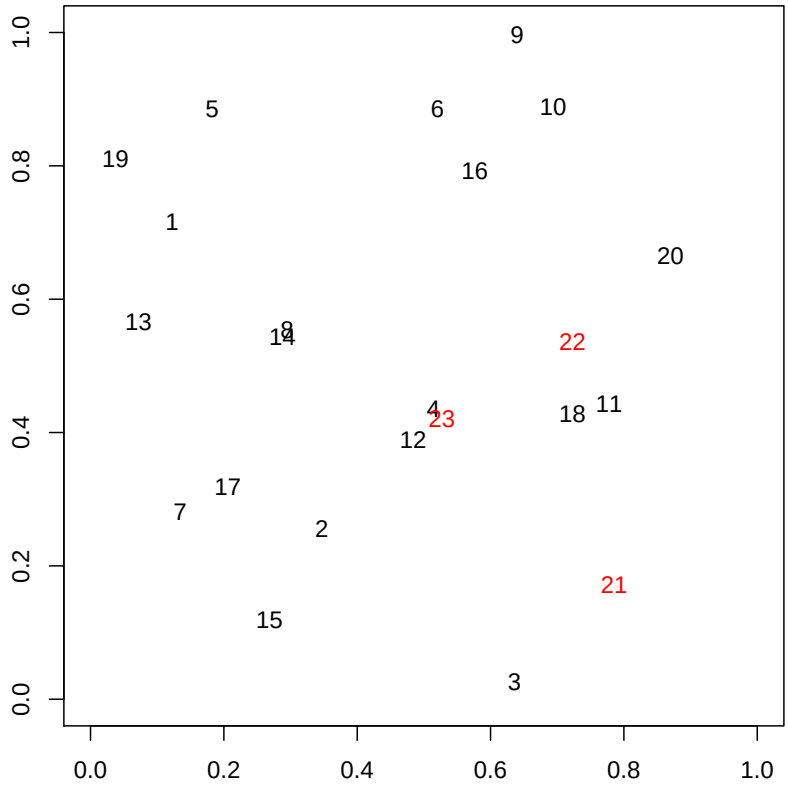


Figure 6.1: Locations for the functional data analysis simulation exercise.

uniformly at random in the interval $[0, 25]$, while the corresponding y coordinates where sampled from the conditional distribution $y(\mathbf{s})|x(\mathbf{s}) \sim \mathsf{N}(m_{\mathbf{s}}(x), 0.04^2)$, where

$$m_{\mathbf{s}}(x) = \begin{cases} a_1(\mathbf{s}) & 0 \leq x < c_1(\mathbf{s}) \\ \left[\frac{c_2(\mathbf{s})-x}{c_2(\mathbf{s})-c_1(\mathbf{s})}\right] a_1(\mathbf{s}) + \left[1 - \frac{c_2(\mathbf{s})-x}{c_2(\mathbf{s})-c_1(\mathbf{s})}\right] \left(a_2(\mathbf{s}) + \frac{a_3(\mathbf{s})}{1+\exp\{-b_1(\mathbf{s})\}}\right) & c_1(\mathbf{s}) \leq x < c_2(\mathbf{s}) \\ a_2(\mathbf{s}) + \frac{a_3(\mathbf{s})}{1+\exp\{-b_1(\mathbf{s})-b_2(\mathbf{s})(x-c_2(\mathbf{s}))\}} & x \geq c_2(\mathbf{s}) \end{cases}$$

The value of the coefficients $(a_1(\mathbf{s}), a_2(\mathbf{s}), a_3(\mathbf{s}), b_1(\mathbf{s}), b_2(\mathbf{s}), c_1(\mathbf{s}), c_2(\mathbf{s}))$ were ran-

domly sampled from independent spatial processes,

$$a_1(\mathbf{s}) \sim \mathsf{GP}(2.00, 0.005, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

$$a_2(\mathbf{s}) \sim \mathsf{GP}(0.20, 0.005, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

$$a_3(\mathbf{s}) \sim \mathsf{GP}(1.65, 0.006, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

$$b_1(\mathbf{s}) \sim \mathsf{GP}(2.0, 0.006, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

$$b_2(\mathbf{s}) \sim \mathsf{GP}(0.4, 0.006, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

$$c_1(\mathbf{s}) \sim \mathsf{GP}(5.0, 0.5, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

$$c_2(\mathbf{s}) \sim \mathsf{GP}(12.0, 0.5, \gamma(\mathbf{s}, \mathbf{s}'|0.2))$$

where $\gamma(\mathbf{s}, \mathbf{s}'|\lambda)$ is a Gaussian correlation matrix with spatial range 3λ. The resulting true curves can be seen in Figure 6.2. In addition, we also obtained the coefficients at 3 additional randomly chosen locations (labeled 21 to 23 in Figure 6.1), but no samples where obtained from them. The curves at these additional locations (depicted in red in Figure 6.2) will serve to investigate the predictive power of the model.

We used the algorithm of section 6.2 to fit the model. As mentioned above, we use exponential correlation functions with unknown range, which is given an exponential prior distribution with expectation 1. The location $\boldsymbol{\mu}_{00}$ was set equal to the sample mean, while $\boldsymbol{\Sigma}_{00}$ was set equal to the sample variance. For τ^2 an inverse gamma with 2 degrees of freedom and expectation 1 is employed. As in previous implementations, α was given a $\mathsf{Gam}(3, 3)$ prior.

All results are based on 75,000 iterations, obtained after a burn-in period of 25,000 samples. In order to initialize the sampler, the covariate space was divided in five sections of equal size, and observations within each block were allocated to the same component of the mixture. Other parameters were initialized accordingly.

Figure 6.3 shows the curves fitted by our model at 6 of the 20 observed locations, along with the observed points and the true function. We can see that the model

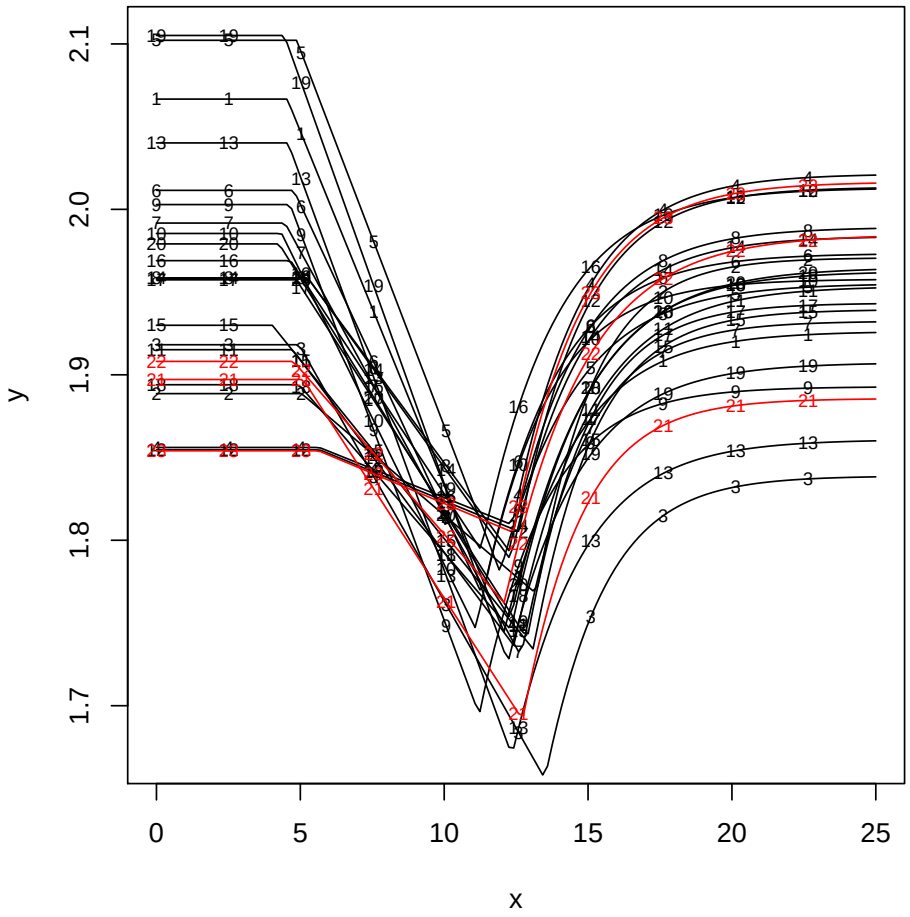


Figure 6.2: True curves used in the functional data analysis simulation exercise.

does a reasonable job at reconstructing the profiles, even for a relatively small sample size like ours. We can see some unexpected bumps in the functional estimates, but these tend to happen in regions that correspond to covariate gaps.

Figure 6.4 shows the predicted curves at the 3 unobserved locations, along with the true curves and pointwise probability bands. Again, the model seems to do a remarkable job at predicting these functional forms, specially at location 23. However, probability bands are relatively wide (specially for location 21). Both of these results are, at least in part, due to fact that location 23 has two very close neighbors (locations 4 and 12), while location 21 is relative isolated from the rest of the observations.

6.4 Discussion

This chapter describes a spatial extension of the density estimation approach to nonparametric regression proposed in chapter 5. We have shown that the idea is computationally feasible, and that the method is capable of fitting and predicting complicated functions using little or no prior information about their shape. However, we note that gaps in the covariate space can yield estimates and predictions that look bumpy due to the way information is borrowed across components and the use of spatially fixed weights. Also, probability bands constructed around the function tend to be wide, compared with those obtained from a parametric fit using the true functional form.

Currently, we have two applications in mind for this model. In the first one, we aim at understanding the spatial relationship between real-estate prices and some covariates of interest such as property size. In the second, we plan to revisit the CDT data analyzed in chapter 5 and construct a model that allows us to incorporate spatial location information in flexible ways.

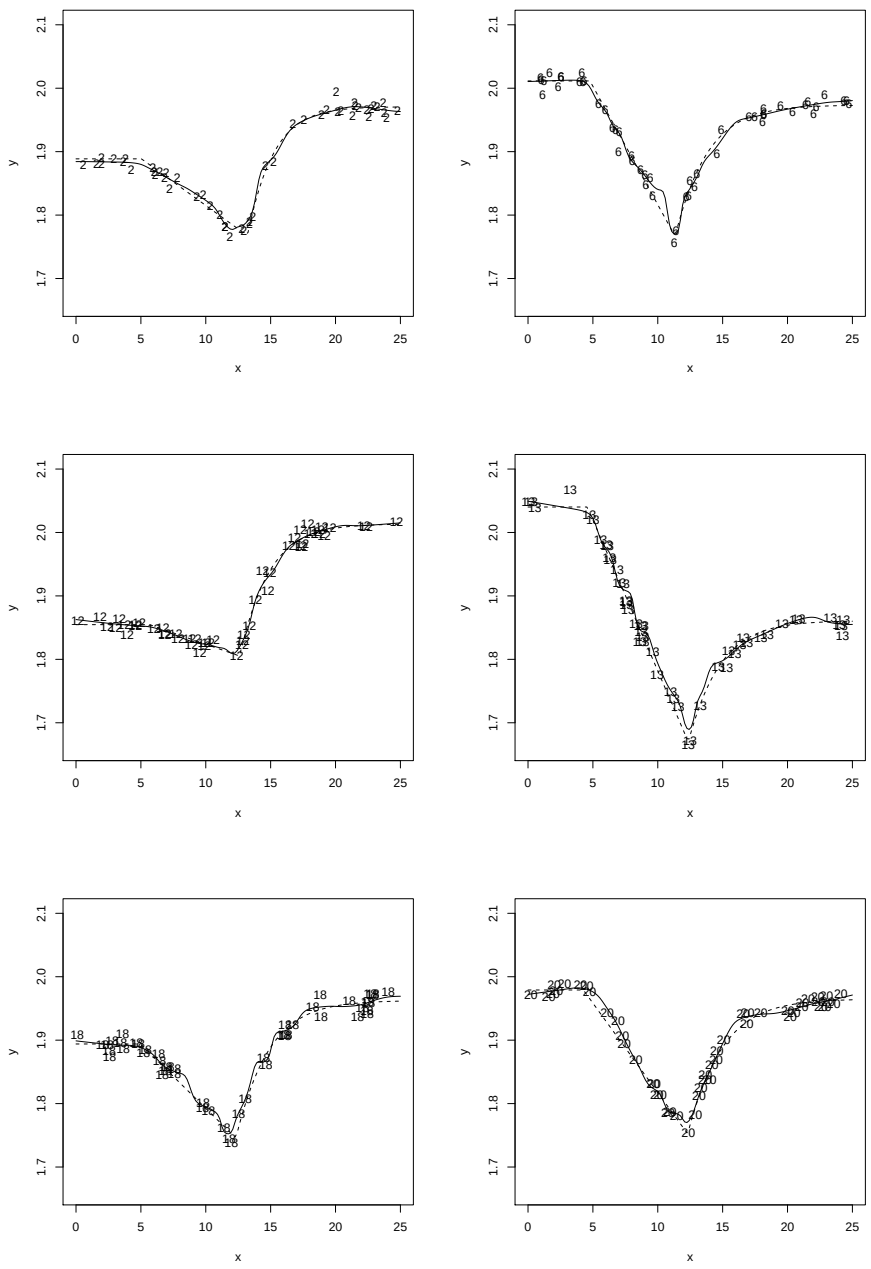


Figure 6.3: Reconstructed profiles from the simulation example at some chosen locations. Dashed lines correspond to the true shapes. Numbers correspond to the actual observations.

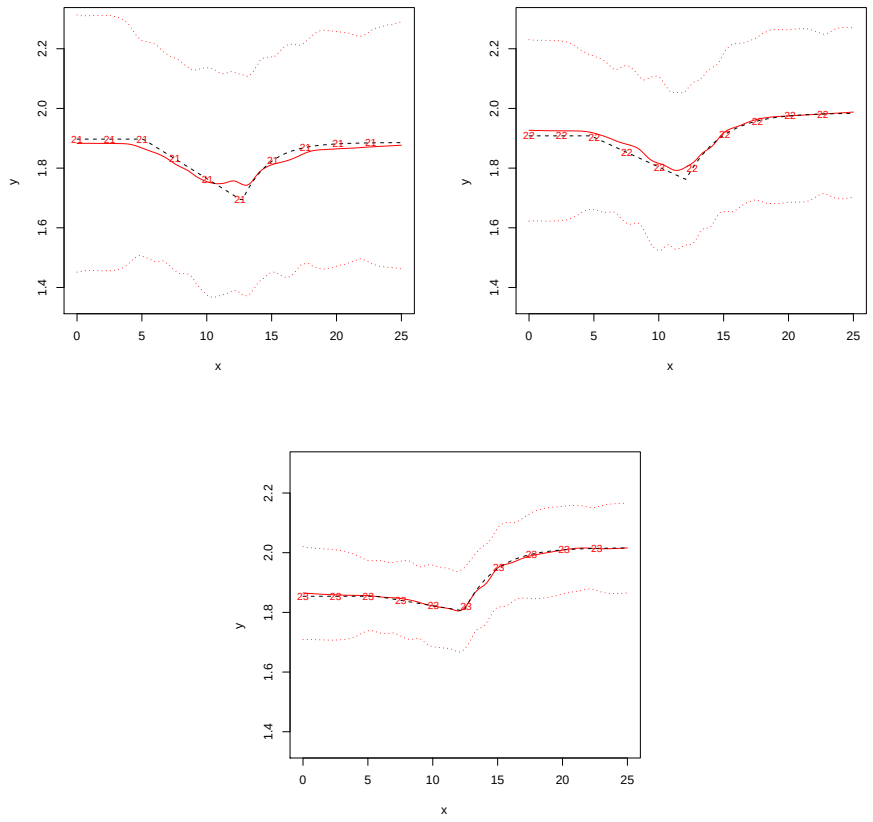


Figure 6.4: Predicted curves at the unobserved locations (solid lines). The true curves are shown as dashed lines, while pointwise probability bands are shown as dotted lines.

Chapter 7

Latent stick breaking processes

7.1 Introduction

In previous chapters we have discussed models for collections of distributions where dependence is introduced in the atoms of the stick-breaking construction. For example, the nested Dirichlet process in chapter 2 used atoms that were in turn samples from a Dirichlet process, inducing a partition of the collection of distributions into clusters. On the other hand, the dynamic nonparametric models in chapter 4 build dependence in the distributions across time by using vector autoregressive models in the atoms of the stick-breaking process.

In contrast, this chapter focuses on developing prior distributions on stochastic processes on an index space $D \subset \mathbb{R}^d$ with rich common marginal distributions. We no longer focus on building different distributions for each possible value of the index space, but instead construct a stochastic process where observations at different locations are dependent but have a common (albeit unknown) marginal distribution. We call the resulting construction a latent stick-breaking process (LaSBP). As the name indicates, we rely on a stick-breaking construction (Ongaro and Cattaneo, 2004; Ishwaran and James, 2001) to represent the unknown marginal distribution, while

introducing an underlying (latent) Gaussian process on D to drive the selection of the atoms at each location $\mathbf{s} \in D$. In order to avoid assuming that the observations have a discrete distribution, we use the LaSBP to model the parameters of a continuous distribution, resulting on a LaSBP mixture (LaSBPM). This construction is, to the best of our knowledge, unique, and provides an excellent alternative to generate flexible temporal, spatial or spatio-temporal models for discrete or continuous data.

Due to the discrete nature of the marginal distribution, the LaSBP also induces a segmentation of the index space D based on flexible covariance structures. This is especially interesting because segmentation models typically assume a Markovian structure in order to simplify computation. Since the selection process in the LaSBP inherits the characteristics of the underlying Gaussian process, our models can easily accommodate a much richer variety of correlation structures, including stationary CAR-like processes of varying degrees (through correlation functions with compact support) and non-stationary processes. Prediction and interpolation is performed through model averaging over the random partition structures, leading to smooth predictions.

This chapter is divided as follows: Sections 7.2 and 7.3 defines the latent stick-breaking process and discusses some of its properties. Section 7.4 describes a MCMC algorithm for posterior computation. In section 7.5 we discuss a simple application to a stochastic volatility model. Finally, in section 7.6 we present a short discussion and future directions.

7.2 Latent stick-breaking process

Our goal is to construct a stochastic process $\{\theta(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ on $(\mathbf{\Theta}, \mathcal{B})$ such that $\theta(\mathbf{s}) \sim G$ for all $\mathbf{s} \in D$ and some unknown G, but $\mathbb{C}\text{ov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) \neq 0$. This is

a nonparametric analog to the Gaussian process. To do this, consider the triplet

$$\left\{\{z(\mathbf{s}) : \mathbf{s} \in D\}, \{v_l\}_{l=1}^{L}, \{\theta_l^*\}_{l=1}^{L}\right\} \tag{7.1}$$

The *latent* Gaussian process $z(\mathbf{s})$ is such that $z(\mathbf{s}) \sim \mathsf{N}(0,1)$ for all $\mathbf{s} \in D$ and $\mathbb{C}\mathrm{or}(z(\mathbf{s}), z(\mathbf{s}')) = \gamma(\mathbf{s}, \mathbf{s}')$. The sequences of *stick breaking ratios* $\{v_l\}_{l=1}^{L}$ is such that $v_l \sim \mathsf{beta}(a_l, b_l)$ for $l < L$ and $v_L = 1$. The sequence of *atoms* $\{\theta_l^*\}_{l=1}^{L}$ is constructed by imposing an order constraint on a sample from a baseline measure H on $(\mathbf{\Theta}, \mathcal{B})$ using the following mechanism: sample $\theta_1^* \sim H$ and for $l > 1$ draw $\theta_l \sim H_l$, where H_l is defined as the restriction of H to the set $S_l = \{\theta : \theta > \theta_{l-1}\}$, i.e., $H_l(B) = H(B \cap S_l)/H(S_l)$ for any measurable set $B \in \mathcal{B}$.

We can use the sequences $\{v_l\}_{l=1}^{L}$ and $\{\theta_l^*\}_{l=1}^{L}$ to define a random distribution

$$G(\cdot) = \sum_{l=1}^{L} w_l \delta_{\theta_l^*}(\cdot)$$

where the probability weights are defined as $w_l = v_l \prod_{k<l}(1-v_k)$ and satisfy $\sum_l w_l = 1$ almost surely. The random distribution G will be the unknown marginal distribution of the process. Note that, taking $L = \infty$, $a_l = 1$, $b_l = b$ and removing the order constraints on $\{\theta_l^*\}_{l=1}^{L}$, the random G follows a Dirichlet process with baseline measure bH. We introduce the order constraint to link the ordering in the underlying Gaussian with the ordering of the atoms $\{\theta_l^*\}_{l=1}^{L}$, which is necessary to ensure identifiability and obtain sensible interpolations and predictions.

Now, for any finite set of locations $\mathbf{s}_1, \ldots \mathbf{s}_n \in D$, let

$$\begin{aligned}\mathbb{P}\left(\theta(\mathbf{s}_1) = \theta_{l_1}^*, \ldots, \theta(\mathbf{s}_n) = \theta_{l_n}^*\right) = \\ \mathbb{P}\Big(\Phi(z(\mathbf{s}_1)) \in [\pi_{l_1-1}, \pi_{l_1}), \ldots, \Phi(z(\mathbf{s}_n)) \in [\pi_{l_n-1}, \pi_{l_n})\Big)\end{aligned} \tag{7.2}$$

where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function and $\pi_l = 1 - \prod_{k \le l}(1 - v_k)$ is the proportion of the unit stick assigned to the first l atoms,

with $\pi_0 = 0$. Clearly, the set of joint distributions in (7.2) satisfies Kolmogorov's consistency conditions and therefore the process is valid. We call $\theta(\mathbf{s})$ a Latent Stick Breaking Process (LaSBP) and denote it as $\theta(\mathbf{s}) \sim \mathsf{LaSBP}_L(\{a_l\}_{l=1}^L, \{b_l\}_{l=1}^L, H, \gamma)$.

In order to gain some intuition into this construction, we show in Figure 7.1 three random realizations of the $\theta(\mathbf{s})$ surface on $D = [0,1]^2$ associated with three different marginal distributions G sharing the same underlying latent process $z(\mathbf{s})$. For this simulation, we took $L = 100$ and $a_l = 1$, $b_l = \alpha$, for three different values of α. Locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ were chosen on an evenly spaced 60×60 grid. The resulting surfaces are tiled, segmenting the space according to the level of the underlying Gaussian process. The parameter α influences the roughness of the segmentation (essentially, the number of different levels that the random surface takes), the baseline measure H controls the actual level of the tiles, and the correlation function γ influences the relative size of the different tiles.

Figure 7.1 also demonstrates the importance of the order constraint in the definition of the process. We use it to link the behavior of the atoms of the stick-breaking construction to the behavior of the latent process. In particular, we want to enforce the internal consistency condition, $z(\mathbf{s}) > z(\mathbf{s}') \Rightarrow \theta(\mathbf{s}) \geq \theta(\mathbf{s}')$. Note that predictions on the value $\theta(\mathbf{s}')$ for an unobserved location $\mathbf{s}'$ depends on the predictions for a $z(\mathbf{s}')$, in particular

$$\begin{aligned}\mathbb{P}(\theta(\mathbf{s}') = \theta_l^* | \theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n), \{z(\mathbf{s}_i)\}_{i=1}^n, \{\theta_l\}_{l=1}^L, \{v_l\}_{l=1}^L) \\ = \mathbb{P}\left(\Phi(z(\mathbf{s}')) \in [\pi_{l-1}, \pi_l) | z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)\right)\end{aligned}$$

Without the order constraint forcing adjacent atoms to have similar values this prediction might yield unreasonable small or large values. This issue can also be approached as an identifiability problem. For any given set of locations, there are multiple possible latent surfaces that can fit the process, each one associated with a different ordering of the atoms. By using the order constrain, we are assigning higher

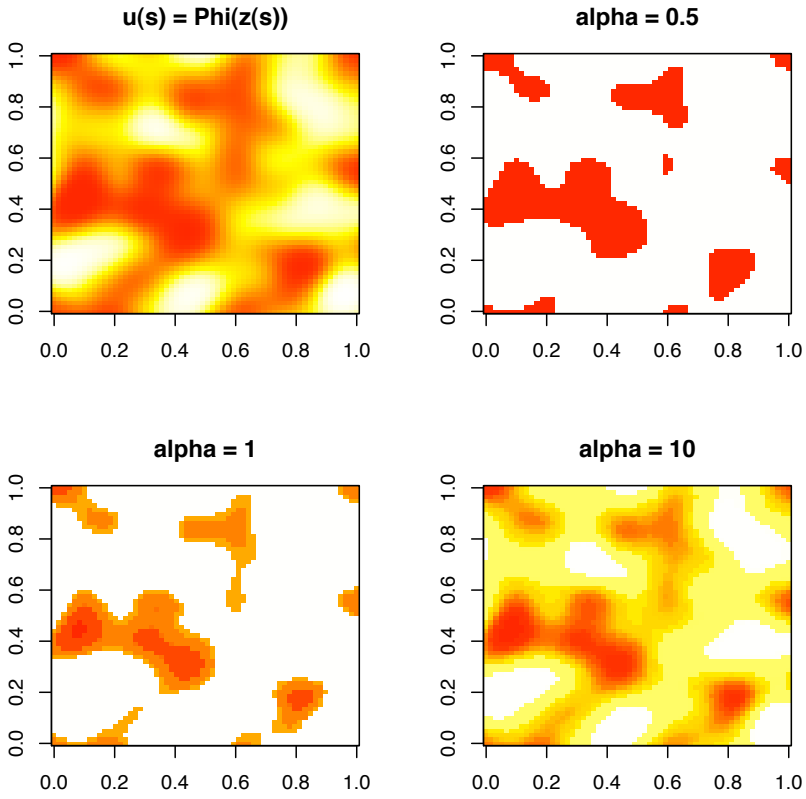


Figure 7.1: Realizations from a LaSBP process on $[0, 1]^2$ with a standard Gaussian baseline measure. We illustrate the effect of different concentrations while keeping the underlying Gaussian process (shown in the upper left panel) constant to simplify interpretation.

posterior density to smoother underlying surface $z(\mathbf{s})$.

7.3 Properties

In what follows, let $\xi(\mathbf{s}) = l$ iff $z(\mathbf{s}) \in [\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_l))$ be a latent process indicating the membership of locations to components of the stick-breaking process and $\beta_l(\mathbf{s}) = \mathbb{P}(\xi(\mathbf{s}) = l)$ represent the marginal probability that a realization of the process at location $\mathbf{s}$ is assigned to component l. Similarly, let $\beta_{lk}(\mathbf{s}, \mathbf{s}') = \mathbb{P}(\xi(\mathbf{s}) = l, \xi(\mathbf{s}') = k)$ be the joint probability of a realization of the process at locations $\mathbf{s}$ and $\mathbf{s}'$ taking values θ_l^* and θ_k^* respectively. Note that these probabilities are random *apriori* since $\pi_1, \ldots, \pi_L$ are random. By definition, $\beta_l(\mathbf{s}) = w_l$ and

$$\beta_{lk}(\mathbf{s}, \mathbf{s}') = \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_l), \Phi^{-1}(\pi_k)) - \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_k)) - \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_l), \Phi^{-1}(\pi_{k-1})) + \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_{k-1}))$$

where $\mathbf{\Phi}_r(\cdot, \cdot)$ denotes the cumulative distribution of the standard bivariate normal with correlation r.

First, consider the expectation of the LaSBP process $\theta(\mathbf{s})$. If ψ is a measurable function on $(\mathbf{\Theta}, \mathcal{B})$, then

$$\mathbb{E}(\psi(\theta(\mathbf{s}))) = \sum_{l=1}^{L} \mathbb{E}(w_l) E(\psi(\boldsymbol{\theta}_l^*)) = \sum_{l=1}^{L} \frac{a_l}{a_l + b_l} \left[\prod_{k<l} \frac{b_k}{a_k + b_k} \right] E(\psi(\boldsymbol{\theta}_l^*)) > \mathbb{E}_H(\psi(\boldsymbol{\theta}^*))$$

Due to the order constrain, some care must be exercised when interpreting the baseline measure in our model. Although H still plays a role controlling the average level of the process, it is not true in general that such level is independent of the choice of the sequences a_l and b_l, as with typical stick-breaking processes. Also, the average level of the process will be strictly larger than the mean of the baseline measure H.

The covariance of the LaSBP $\theta(\mathbf{s})$ process is given by

$$\mathbb{C}\text{ov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) = \sum_{l=1}^{L} \sum_{k=1}^{L} \left[\mathbb{E}(\beta_{lk}(\mathbf{s}, \mathbf{s}')) \mathbb{E}(\theta_l^* \theta_k^*) - \mathbb{E}(\beta_l(\mathbf{s}) \beta_k(\mathbf{s}')) \mathbb{E}(\theta_l^*) \mathbb{E}(\theta_k^*) \right]. \quad (7.3)$$

Although we do not have a closed form expression for the covariance function, it is clear from (7.3) that the process on $\theta(\mathbf{s})$ is isotropic, stationary or non-stationary if $z(\mathbf{s})$ also has such a characteristic.

Realizations from the LaSBP $\theta(\mathbf{s})$ are *discontinuous* almost surely. This is clear from Figure 7.1, and is a consequence of the stick-breaking construction. However, note that

$$\beta_{lk}(\mathbf{s}, \mathbf{s}') \to \beta_l(\mathbf{s})\beta_k(\mathbf{s}') \quad \text{as} \quad \gamma(\mathbf{s}, \mathbf{s}') \to 0$$

$$\beta_{lk}(\mathbf{s}, \mathbf{s}') \to \begin{cases} \beta_l(\mathbf{s}) & l = k \\ 0 & \text{otherwise} \end{cases} \quad \text{as} \quad \gamma(\mathbf{s}, \mathbf{s}') \to 1$$

In particular, this implies that $\mathbb{P}(\theta(\mathbf{s}) = \theta(\mathbf{s}')) \to 1$ as $\mathbf{s}' \to \mathbf{s}$ almost surely. Therefore the selection mechanism (which assigns atoms to locations) is continuous even if the resulting realizations are not. This result also shows that, as the range of the Gaussian process goes to zero, samples at different locations become iid observations from G. On the other hand, as the range grows, the LaSBP reduces to a parametric model with a prior H on the unknown parameter.

It is important to note that $\theta(\mathbf{s})$ and $\theta(\mathbf{s}')$ are never independent. Indeed, even if $\gamma(\mathbf{s}, \mathbf{s}') = 0$, implying that $\mathbb{P}(\theta(\mathbf{s}) = \theta_l^*, \theta(\mathbf{s}') = \theta_{l'}^*) = \mathbb{P}(\theta(\mathbf{s}) = \theta_l^*)\mathbb{P}(\theta(\mathbf{s}') = \theta_{l'}^*)$, $\mathbb{C}\text{ov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) \neq 0$. Indeed, if $\gamma(\mathbf{s}, \mathbf{s}') = 0$ then from 7.3

$$\mathbb{C}\text{ov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) = \sum_{l=1}^{L}\sum_{k=1}^{L} \mathbb{E}(\beta_l(\mathbf{s})\beta_k(\mathbf{s}'))\mathbb{C}\text{ov}(\theta_l^*\theta_k^*) > 0$$

since $\mathbb{C}\text{ov}(\theta_l^*\theta_k^*) = \mathbb{V}(\theta_l^*) > 0$ if $l = k$ and the order constraint implies $\mathbb{C}\text{ov}(\theta_l^*\theta_k^*) \geq 0$ for $l \neq k$.

Finally, it is also interesting to formally consider the effect of the stick-breaking distribution on the surface. Assume for simplicity that $a_l = a$ and $b_l = b$ for all l, which we will do in most practical applications. Note that as $\frac{a}{a+b} \to 1$ we have

$w_1 \to 1$ and the model degenerates again into a parametric model with a prior H on the unknown parameter. Therefore, if the true surface is indeed close to flat, identifiability issues could arise between the spatial range and the parameters of the stick-breaking distributions.

We also note that, due to the order constraint on the atoms, the typical choice $a = 1$ and $b = \alpha$ can induce severe artifacts in the model. It is well known that the weights for a regular Dirichlet process are stochastically ordered. Although this is not an issue when the atoms are exchangeable, under the order constrain, this implies that the components with smaller parameters receive higher weight.

7.4 Computation

For inference purposes, we develop an MCMC algorithm. We assume $a_l = a$, $b_l = b$, $L < \infty$ and use a blocked sampler (Ishwaran and James, 2001). If $N = \infty$, a retrospective sampler (Roberts and Papaspiliopoulos, 2007) can be easily obtained as an extension. In order to facilitate computation, the latent process, $z(\mathbf{s})$ and $\xi(\mathbf{s})$ are sampled explicitly. Using the familiar bracket notation, the joint distribution is given in this case by

$$\prod_{i=1}^{n}[y(\mathbf{s}_i)|\boldsymbol{\theta}^*, \xi(\mathbf{s}_i)] \times \prod_{i=1}^{n}[\xi(\mathbf{s}_i)|z(\mathbf{s}_i), \mathbf{v}] \times [z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)|\lambda] \times \prod_{l=1}^{\infty}[v_i|\alpha] \times \prod_{l=1}^{\infty}[\boldsymbol{\theta}_l] \times [\alpha] \times [\lambda]$$

After setting up initial values for all parameters in the model, the algorithm proceeds by sequentially updating the parameters according to the following steps:

1. Jointly update the latent processes $\xi(\mathbf{s})$ and $z(\mathbf{s})$, one location at a time, by first sampling $\xi(\mathbf{s}_i)$ from a discrete distribution such that

   $$\mathbb{P}(\xi(s_i) = l) \propto p(y(\mathbf{s}_i)|\theta_l) \times$$
   $$\mathbb{P}(z(\mathbf{s}_i) \in [\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_l))|z(\mathbf{s}_1), \ldots, z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i+1}), \ldots, z(\mathbf{s}_n))$$

where the prior probability of component l can be obtained by a univariate integration, and then sampling $z(\mathbf{s}_i)$ from the restricted univariate normal distribution defined by

$$[z(\mathbf{s}_i)|z(\mathbf{s}_1),\ldots,z(\mathbf{s}_{i-1}),z(\mathbf{s}_{i+1}),\ldots,z(\mathbf{s}_n)]\mathbf{1}_{\Omega_l}$$

where $\Omega_l = \{z(\mathbf{s}_i) : z(\mathbf{s}_i) \in [\Phi^{-1}(\pi_{\xi(\mathbf{s}_i)-1}), \Phi^{-1}(\pi_{\xi(\mathbf{s}_i)}))\}$. Note that, if $\mathbf{s}_i = \mathbf{s}_j$ for some j, then the prior probability for observation i being assigned to component $\xi(\mathbf{s}_j)$ is one and therefore $\xi(\mathbf{s}_i) = \xi(\mathbf{s}_j)$ and $z(\mathbf{s}_i) = z(\mathbf{s}_j)$, as expected.

2. Sample the stick-breaking ratios one at a time from their full conditional

$$[v_l|\mathbf{v}_{(-l)}] \propto v_l^{a-1}(1-v_l)^{b-1}\mathbf{1}_{A_l}$$

where $A_l = \left\{v_l : q_l^l \leq v_l < q_l^u\right\}$ and

$$q_l^l = \max_{\{i:\xi(\mathbf{s}_i)\geq l\}}\left\{1 - \frac{1-\Phi(z(\mathbf{s}_i))}{\prod_{k\leq\xi(\mathbf{s}_i),k\neq l}(1-v_k)}\right\}$$

$$q_l^u = \min_{\{i:\xi(\mathbf{s}_i)\geq l+1\}}\left\{1 - \frac{1-\Phi(z(\mathbf{s}_i))}{\prod_{k\leq\xi(\mathbf{s}_i)-1,k\neq l}(1-v_k)}\right\}$$

Note that v_l depends on $\mathbf{v}_{(-l)}$ only through the constraint on the support. Also, v_l depends on $z(\mathbf{s})$ only through $z(\mathbf{s}_i) : \xi(\mathbf{s}_i) \geq l$. Therefore, for $l > l^* = \max_i\{\xi(\mathbf{s}_i)\}$, v_l is conditionally independent from $\mathbf{v}_{(-l)}$.

3. Sample the atoms $\{\theta_l\}$. The full conditional is given by

$$p(\theta_l|\cdots) \propto h(\theta_l)\prod_{\{i:\xi(\mathbf{s}_i)=l\}} p(\mathbf{y}(\mathbf{s}_i)|\theta_l)$$

where h is the density associated with the baseline measure H. If H is conjugate with the likelihood p, this is straightforward.

4. The prior parameters on the stick-breaking ratios a and b are jointly sampled using a random-walk Metropolis algorithm. Since both parameters need to be positive, new values $(a^{(p)}, b^{(p)})$ are proposed from a log-normal density centered on the logarithm of the of the current value $(a^{(c)}, b^{(c)})$. Specifically,

$$(\log a^{(p)}, \log b^{(p)}) \sim \mathsf{N}\left(\begin{pmatrix}\log a^{(c)} \\ \log b^{(c)}\end{pmatrix}, \boldsymbol{\Sigma}\right)$$

with $\boldsymbol{\Sigma}$ being a tuning parameter. The proposed values are accepted with probability $\min\{1, \delta\}$ where

$$\delta = \left(\frac{\Gamma\left(a^{(p)} + b^{(p)}\right)\Gamma\left(a^{(c)}\right)\Gamma\left(b^{(c)}\right)}{\Gamma\left(a^{(c)} + b^{(c)}\right)\Gamma\left(a^{(p)}\right)\Gamma\left(b^{(p)}\right)}\right)^{L-1} \times$$

$$\left(\prod_{l=1}^{L-1} v_l\right)^{a^{(p)}-a^{(c)}} \left(\prod_{l=1}^{L-1}(1-v_l)\right)^{b^{(p)}-b^{(c)}} \frac{p(a^{(p)}, b^{(p)})}{p(a^{(c)}, b^{(c)})} \frac{a^{(p)}b^{(p)}}{a^{(c)}b^{(c)}}$$

and $p(a, b)$ is the prior distribution for the parameters. In our example we take a and b to have a priori independent gamma distributions centered around 1.

5. The parameters of the underlying Gaussian process can be sampled conditional on the current realization $z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)$ using a random-walk Metropolis-Hastings algorithm, as is customary with other spatial models.

6. Since the conditional moves in step 1 can raise some concerns about mixing rates, we suggest to additionally sample the whole Gaussian process jointly. This can be easily done by noting that the joint full conditional is given by

$$[z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n) | \cdots] \cdots \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Psi})\, \mathbf{1}_\Omega$$

where $[\boldsymbol{\Psi}]_{ij} = \gamma(\mathbf{s}_i, \mathbf{s}_j)$, $\Omega = \cap_{i=1}^n \Omega_i$ and $\Omega_i = \{z(\mathbf{s}_i) : \Phi^{-1}(\pi_{\xi(\mathbf{s}_i)-1}) \leq z(\mathbf{s}_i) < \Phi^{-1}(\pi_{\xi(\mathbf{s}_i)})\}$ as before. Due to the form of the restrictions, we can sample from this truncated distribution by iterative conditioning.

7.4.1 A short note on efficient computational implementation

A naive implementation of steps 1 and 6 can be computational expensive, requiring $o(n)$ expensive matrix inversions. However, it is possible to implement both of these sampling steps using a single additional Cholezky decomposition of $\mathbf{\Psi}$.

First, note that matrix $\mathbf{\Psi}$ and its inverse $\mathbf{\Delta} = \mathbf{\Psi}^{-1}$ need to be calculated at every iteration of step 5 in order to evaluate the Metropolis ratio. Now, let $\mathbf{\Delta}^i_{--}$ be the $(n-1) \times (n-1)$ matrix obtained by removing the i-th row and the i-th column from $\mathbf{\Delta}$. Similarly, let $\mathbf{\Delta}^i_{++} = [\mathbf{\Delta}]_{ii}$ be the i-th diagonal entry of $\mathbf{\Delta}$, and $\mathbf{\Delta}^i_{+-}$ be the i-th row of $\mathbf{\Delta}$ with i-th entry removed. It is easy to show using well known results for partitioned matrices that $z(\mathbf{s}_i)|z(\mathbf{s}_1), \ldots, z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i+1}), \ldots, z(\mathbf{s}_n)$ has mean $-\mathbf{\Delta}^i_{+-}\mathbf{z}^i/\mathbf{\Delta}^i_{++}$ and variance $\mathbf{\Delta}^i_{++}$, where $\mathbf{z}^i = z(\mathbf{s}_1), \ldots, z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i+1}), \ldots, z(\mathbf{s}_n)$. Therefore, no additional matrix inversions are required to complete step 1.

For step 6, consider first the very easy problem of generating an unconstrained sample $\mathbf{z}$ from a multivariate normal distribution $\mathsf{N}(\mathbf{0}, \mathbf{U}\mathbf{U}')$, where $\mathbf{U}$ is a positive definite lower triangular matrix. The usual procedure is to use a change of variables, generating first a vector $\mathbf{x}$ of n iid observations $x_i \sim \mathsf{N}(0,1)$ and then letting $\mathbf{z} = \mathbf{U}\mathbf{x}$. This way, the sequence of conditional distributions $\{p(z_i|z_1, \ldots, z_{i-1})\}$ is never explicitly obtained. It is possible to produce a similar scheme that uses a sequence of (no longer independent) draws $x_1, \ldots x_n$ and a transformation to generate a vector $\mathbf{z}$ of dependent random variables subject to constraints like the ones in our problem. Our approach is summarized in the following proposition

Proposition 4. *Let* $\mathbf{U}$ *be a lower triangular matrix with strictly positive diagonal entries and* $\Omega_i = [c_i, d_i]$ *with* $c_i, d_i \in \mathbb{R}$ *for all* $i = 1, \ldots, n$. *The following set of steps,*

1. *Sequentially sample* $x_i \sim \mathsf{N}(0,1)$ *restricted to the set* $T_i(\Omega_i) = [T_i(c_i), T_i(d_i)]$,

where

$$T_i(y) = \frac{y - \sum_{j<i} U_{ji} x_j}{U_{ii}}$$

2. *Let* $z_i = T_i^{-1}(x_i)$

generates a sample $\mathbf{z} \sim \mathsf{N}(\mathbf{0}, \mathbf{U}\mathbf{U}')$ *subject to the restriction* $\mathbf{z} \in \Omega = \cap_{i=1}^{n} \Omega_i$

Proof. Straightforward using a change of variables. □

7.5 Stochastic volatility and option pricing

In this section we consider an application of the LaSBP to the modeling of stock-market returns. The data set under consideration consists of the weekly returns of the S&P500 index covering the ten-year period between April 21, 1997 and April 9, 2007, for a total of 520 observations. Figure 7.2 shows the evolution of these returns. The series does not exhibit any long term trend and different levels of volatility can be clearly seen. Indeed, it is well known that financial time series typically exhibit heavy tails and periods of low/high volatility tend to cluster together. Also, two slightly different regimes are apparent. Before May 2003, periods of high-volatility are relatively frequent. After May 2003, we can appreciate longer low-volatility periods.

Similarly to other approaches in the literature, we model $r(t)$, the return of the S&P500 at time t, as following a normal distribution with constant mean but time-varying variance. The time-varying variance is then assumed to follow a LaSBP with an inverse-gamma baseline measure, which allows us to simplify computation while providing a flexible model. To demonstrate the methodology we use an exponential correlation function for the latent process, but implementations using more general correlation functions (like Mattern, power exponential or some non-stationary family)

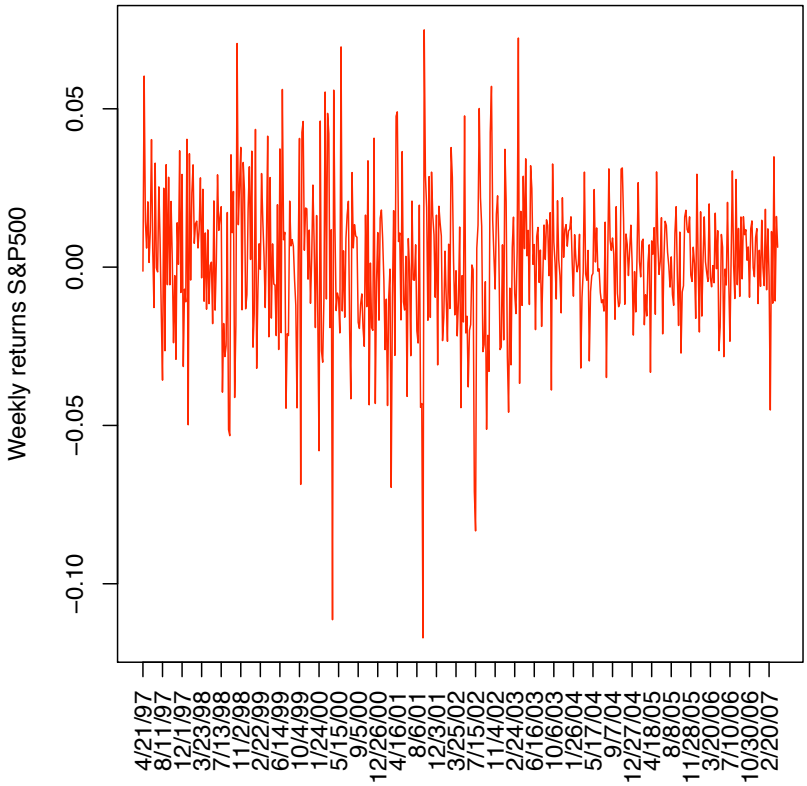


Figure 7.2: Weekly returns on the S&P500 index between April 21, 1997 and April 9, 2007.

is just as straightforward. Also, we take $a_l = 1$ and $b_l = \alpha$ for all l. We noted above that this choice places stochastically larger probabilities on smaller values of volatility, and therefore was not the most adequate choice in general. However, a prior with this structure is adequate in this application because we see low volatility as the norm and high volatility as sporadic shocks to the system. Specifically, we use the following model specification:

$$
\begin{aligned}
r(t) &\sim \mathsf{N}(\mu, \sigma^2(t)) & \sigma^2(t) &\sim \mathsf{LaSBP}_L(1, \alpha, H, \gamma) \\
\gamma(t, t') &= \exp\left\{-\frac{|t - t'|}{\lambda}\right\} & H &= \mathsf{IGam}(\nu, \sigma_0^2) \\
\mu &\sim \mathsf{N}(\mu_0, \tau^2) & \alpha &\sim \mathsf{Gam}(a_\alpha, b_\alpha) \\
\lambda &\sim \mathsf{Gam}(a_\lambda, b_\lambda) & \sigma_0^2 &\sim \mathsf{Gam}(\delta, \sigma_{00}^2)
\end{aligned}
$$

This model results in a marginal distribution that is a scale mixture of Gaussian distributions, yielding a rich model on the class of unimodal distributions that can accommodate fat tails. Also, the structure of the LaSBP naturally induces volatility clustering, while potentially allowing for richer dependence structures than Markov-switching models.

Since the historical annual volatility of the S&P500 is traditionally estimated to be in the range of 12-15%, we take $\tau^2 = 0.15^2/52$ and $\sigma_{00}^2 = 0.12^2/52$. The prior mean for the mean return is taken to be $\mu_0 = 0$, and the degree of freedom for the baseline measure are chosen as $\nu = 2$ and $a_\sigma = 1$. Following standard practice, we picked $a_\alpha = b_\alpha = 1$ so that the prior for the precision parameter α is centered around 1 and has variance 1. The prior for the spatial correlation is also set as $a_\lambda = b_\lambda = 1$, implying that the a-priori range of the correlation process is approximately 3 weeks. A small sensitivity analysis was conducted by changing the hyperpriors on σ_0^2, α and λ, and results seemed mostly robust to this prior selection. All results are based on 50,000 iterations obtained after a burn-in period of 5,000 samples. We resampled the

latent process every 10 iterations to improve mixing.

The model identifies between 2 and 11 volatility regimes, with models having between 3 and 6 components receiving the most weight (estimated posterior probabilities 0.27, 0.33, 0.20 and 0.10 respectively). The mean of the precision α is estimated to be 0.700 (median 0.650, symmetric 90% credible interval (0.260 ; 1.312)). The posterior distribution of the correlation parameter λ is has a mean of 110.7186 (90% credible interval (34.05,241.53)), showing strong evidence for long range dependence in the sample. The width of the credible interval also suggests that the parameters of the latent Gaussian process might be hard to estimate as the sample contains little information about them.

The posterior mean weekly return is estimated to be 0.00194 (credible interval (0.00050 ; 0.00336)), a reasonable value given historical evidence. Figure 7.3 shows the volatility path estimated by the LaSBP model, along with the sample standard deviation and the volatility path obtained from the "standard" stochastic volatility described in Jacquier *et al.* (1994). Although both models provide similar volatility estimates, there are some interesting differences. For example, the low volatility period between October 2003 and December 2006 is estimated by us to be a period with essentially constant volatility, while the regular volatility model tends to fluctuate a lot. Nonetheless, both models adapt very fast to the raise in volatility in early 2007. Also the three high volatility peaks in 2000, 2001 and 2002 are more pronounced under our model.

Predicted volatilities for the 40 weeks following April 9th are shown in figure 7.4. Since the last weeks of our sample happen to belong to a period with below-average volatility, predictions start below the standard deviation of the marginal distribution (which coincides with the historical volatility). However, as time elapses, the volatility tends to converge towards the sample mean. The slow converge is consistent with

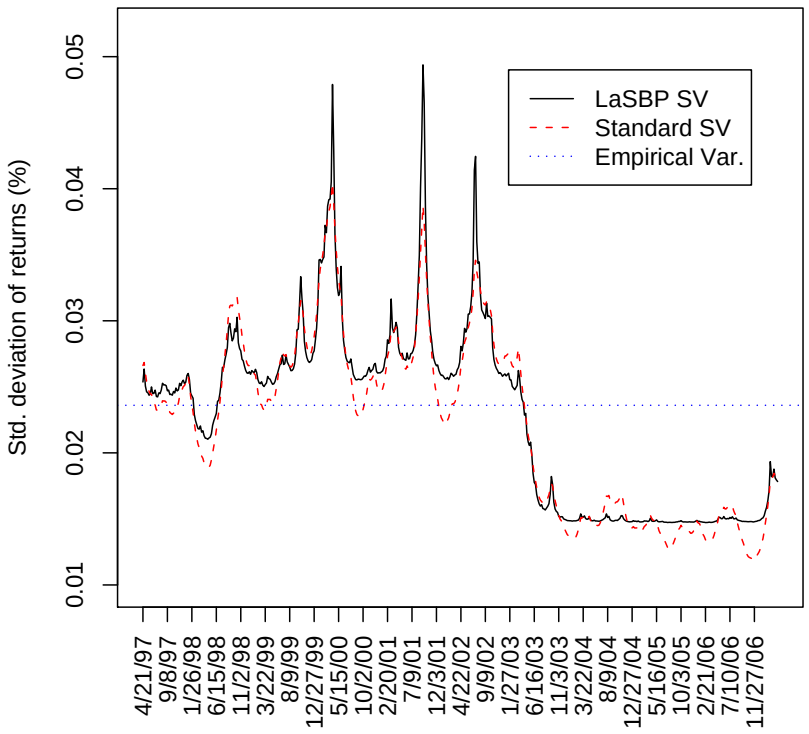


Figure 7.3: Smoothed weekly volatilities for the S&P500 using the LaSBP stochastic volatility model and the standard stochastic volatility model. The flat blue line corresponds to the empirical standard deviation of the sample.

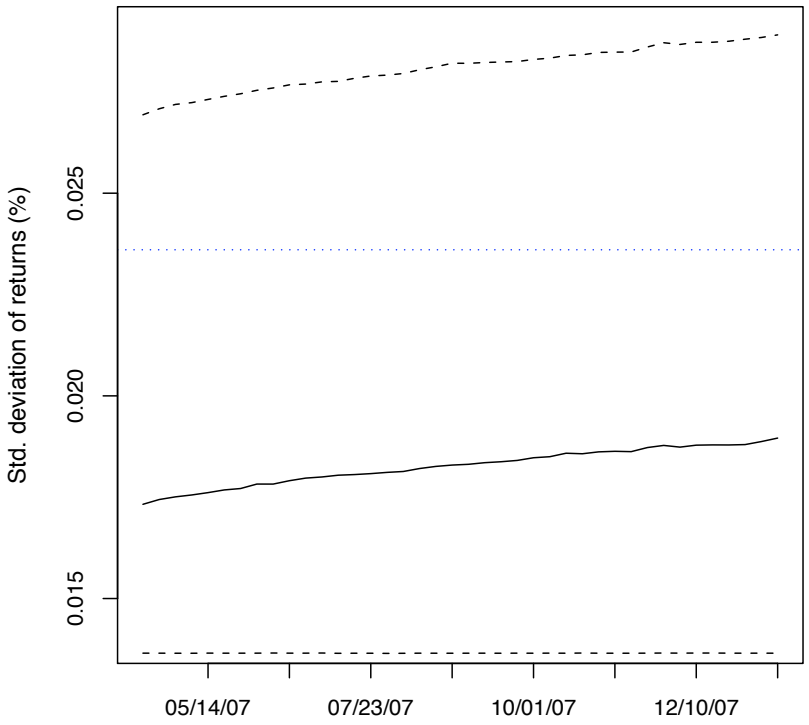


Figure 7.4: Predicted weekly volatilities for the S&P500. Dashed lines corresponds to 90% pointwise credible bands and the dotted blue line to the empirical standard deviation of the sample

the large estimates for the temporal dependence parameter λ. The width of the credible bands reflects considerable uncertainty in the prediction, and are consistent with those obtained from the regular stochastic volatility model (not shown).

One important application of stochastic volatility models is option pricing. European call options are contracts that give the buyer the right to acquire a security at some specific time in the future for a pre-specified price, called the *strike price*. Similarly, put options give the right to sell a security at a pre-specified price. The price a risk-neutral investor is willing to pay for a call option equals the net present

value of its expected payoff,

$$C(t) = \exp\{-q(T-t)\}\mathbb{E}\left\{S(T) - X\right\}$$

where $C(t)$ is the price at time t of an American-style call option with expiration date T and strike price X, q is the (constant) risk-free interest rate and $S(T)$ is the price of the security at expiration. The famous Black & Scholes (B&S) formula for option valuation is obtained by calculating this expectation under the assumption that the returns follow a constant Gaussian distribution.

In order to estimate the price of options under our model, we use the price on April 9, 2007 ($1453.85) and simulations of future returns to estimate the future price of the underlying security. Figure 7.5 shows the theoretical price of in-the-money call options with a strike price of 1460 expiring between one and forty weeks from April 9, 2007, under our SV LaSBP model, the standard stochastic volatility model, and the Black & Scholes formula. The risk-free interest rate is assumed to be equal to the LIBOR (London Inter-Bank Offered Rate) for a one year deposit in U.S. Dollars during March, as reported by Fannie Mae, yielding $q = 5.2009\%$. B&S calculations assume that the mean and variance of the returns equals the historical average over the ten-year period.

For options with early expiration dates the classical SV and the LaSBP SV models yield very similar prices, which are lower than the prices predicted by the Black & Scholes formula. However, as expiration dates increase, the prices generated by both models diverge. While the prices predicted by the standard SV model remain lower than those predicted by Black & Scholes, prices from the LaSBP SV model increase fast, so that for expiration over 20 weeks, they are higher than the BS prices. This is due to the ability of the LaSBP to capture and reproduce heavy tails in the marginal distribution of the process.

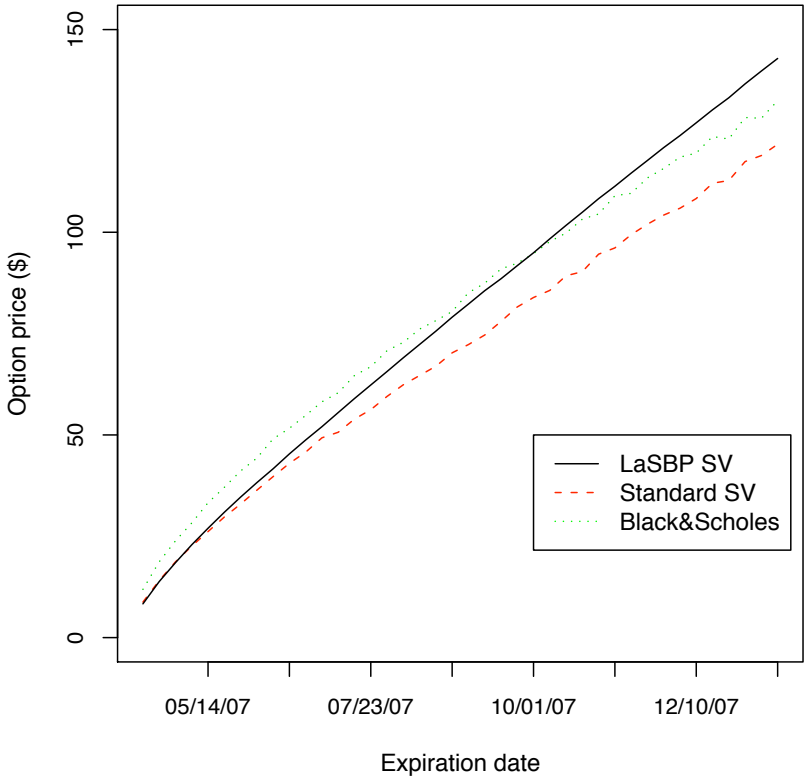


Figure 7.5: Call prices for different maturities.

7.6 Discussion

We have developed and illustrated a novel mechanism to generate stochastic processes with random marginal distributions. This is done by representing the common marginal distribution using a stick-breaking construction and inducing dependence in the selection mechanism. To the best of our knowledge, this is a completely novel approach that shows promise in multiple applications requiring simultaneous estimation of the marginal distribution and the dependence structure.

It is important to emphasize that, in general, it is not appropriate to use a $\mathsf{beta}(1, b)$ distribution as the prior for the stick-breaking ratios. It is well known that this type of prior gives stochastically greater weights to components with smaller indexes. Due to order constraint involved in our construction, this means that such a prior gives larger weights to smaller parameter values. In our stochastic volatility application, this is exactly the type of behavior that we want to encourage, but in other applications this characteristic can be a serious limitation. Therefore, we suggest in general to use the two parameter $\mathsf{beta}(a, b)$ prior and let the data decide.

In principle, direct generalizations to multivariate atoms are not possible because of the order constrain. An alternative is to build the dependence in the selection mechanisms by using multivariate Gaussian processes as latent processes. This could be used in the stochastic volatility example to construct a model that allows both the mean and the variance of the Gaussian distribution to evolve in time. Another interesting addition to this model would be a nonstationary covariance function for the underlying process in order to deal with the varying frequency exhibited by the time series, as well as different alternatives to the Gaussian kernels (for example, uniform or T kernels).

Chapter 8

Concluding remarks and future directions

This dissertation presented two years of work on Bayesian nonparametric density estimation and regression. We feel that this is a balanced combination of theoretical results and interesting applied problems. However, our major contribution has been the development of innovative models with a wide range of applications in public health, earth sciences and finance. Our main theme has been the development of methods for inference on problems where the outcome variable is an infinite dimensional object.

Bayesian nonparametric methods have taken a second wind in the last five years and we expect that they will provide ample opportunities for both applied and theoretical research in the future. We have already mentioned some of the directions we plan to explore after turning the different chapters into publishable papers. We obtained some consistency results in the context of the nested Dirichlet process, but much work remains to be done. In general consistency of priors on collections of distributions like the dependent Dirichlet process is an open problem that we intend to explore. Our dynamic density estimation model provides a setup for improved dynamic portfolio design using expected shortfalls and value-at-risk as objective functions instead of the variance. The latent stick-breaking process offers an interesting

alternative to construct complicated, non-Gaussian spatio-temporal models.

There are also some methodological issues related to summarizing the posterior distributions on high dimensional spaces that have been tackled only partially in the literature and deserve more attention. One of them is how to convey uncertainty about the function estimates provided by nonparametric models. For regression functions, we have provided pointwise probability intervals for our functional estimates. This is not satisfactory in general. For example, pointwise confidence bands around a density function are not densities (i.e., they do not integrate to 1). Besides, pointwise intervals provide local and not global information about the behavior of the estimates. We believe that a decision theory approach is the right path to follow, but it not clear how to pick adequate loss functions that allow for easy computation. Another interesting issue is how to summarize the results from the posterior distribution over clusters. Some advances have been recently made in this topic, but additional research is still needed, particularly on how to choose adequate loss functions and how to summarize uncertainty.

We would also like to investigate the possibility of integrating some of these nonparametric models into more complicated hierarchical specifications. For example, we interested in using functions (including distribution functions) as predictors rather than outcome variables. When basis expansions are used to represent unknown functions, the coefficients of the expansion can be used as predictors. Similarly, the moments of distribution are often employed as predictor variables in applications. However, this is not always a satisfactory solution. For example, the first few moments of a distribution, although easy to interpret, might not capture the relevant characteristics that help explain the dependence between outcome and predictor.

Appendix A

Notation

This appendix establishes the notation and parameterizations we used in the dissertation.

A.1 Dirichlet distribution

We say that $(\theta_1, \ldots, \theta_k) \sim \mathsf{Dir}(\gamma_1, \ldots, \gamma_k)$ if the density satisfies

$$p(\theta_1, \ldots, \theta_k) \propto \prod_{i=1}^{k} \theta_i^{\gamma_i - 1}$$

A.2 Normal-inverse-Wishart distribution

We say that $(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \sim \mathsf{NIW}_p(\boldsymbol{\theta}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0)$ if the joint density can be written as

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

A.3 Gamma distribution

We denote $\tau \sim \mathsf{Gam}(a, b)$ if

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

A.4 Wishart distribution

We write $\mathbf{S} \sim \mathsf{W}_p(\gamma, \mathbf{S}_0)$ if

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

Appendix B

Correlation in the nDP

We start by calculating the correlation between distributions. In the first place,

$$\begin{aligned}\mathbb{E}(G_j(B)G_k(B)) &= \mathbb{E}(G_j(B)G_k(B)|G_j = G_k)\mathbb{P}(G_j = G_k) + \\ &\quad \mathbb{E}(G_j(B)G_k(B)|G_j \neq G_k)\mathbb{P}(G_j \neq G_k) \\ &= \mathbb{E}(G_j^2(B))\frac{1}{\alpha+1} + \mathbb{E}(G_j(B))\mathbb{E}(G_k(B))\frac{\alpha}{\alpha+1} \\ &= \frac{H(B)(1-H(B))}{(\alpha+1)(\beta+1)} + H^2(B)\end{aligned}$$

Finally

$$\begin{aligned}\mathbb{C}\text{ov}(G_j(B), G_k(B)) &= \frac{H(B)(1-H(B))}{(\alpha+1)(\beta+1)} + H^2(B) - H^2(B) \\ &= \frac{H(B)(1-H(B))}{(\alpha+1)(\beta+1)}\end{aligned}$$

and

$$\mathbb{C}\text{or}(G_j(B), G_k(B)) = \frac{\mathbb{C}\text{ov}(G_j(B), G_k(B))}{\sqrt{\mathbb{V}(G_j(B)\mathbb{V}(G_k(B)}} = \frac{1}{\alpha+1}$$

For the correlation between samples of the nDP, note that for the nDP and if

$j = j'$ then

$$\begin{aligned}\mathbb{C}\text{ov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) &= \mathbb{C}\text{ov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j} | \boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j} = \boldsymbol{\theta}^*_{lk}) \mathbb{P}(\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j} = \boldsymbol{\theta}^*_{lk}) + \\ &\quad \mathbb{C}\text{ov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j} | \boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j}) \mathbb{P}(\boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j}) \\ &= \frac{1}{1+\beta} \mathbb{V}(\boldsymbol{\theta}^*_{lk})\end{aligned}$$

Since $\boldsymbol{\theta}^*_{lk}$ are iid for all l and k, it follows that $\mathbb{C}\text{or}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j}) = \frac{1}{1+\beta}$. On the other hand, if $j \neq j'$

$$\begin{aligned}\mathbb{C}\text{ov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) &= \mathbb{C}\text{ov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j} | G_j = G_{j'} = G^*_k, \boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j} = \boldsymbol{\theta}^*_{lk}) \mathbb{P}(G_j = G_{j'} = G^*_k, \boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{i'j}) + \\ &\quad \mathbb{C}\text{ov}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j} | G_j \neq G_{j'} \text{ or } \boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j}) \mathbb{P}(G_j \neq G_{j'} \text{ or } \boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{i'j}) \\ &= \frac{1}{(1+\alpha)(1+\beta)} \mathbb{V}(\boldsymbol{\theta}^*_{lk})\end{aligned}$$

Appendix C

Proof of theorem 2

Let $P^{\infty\infty}(\boldsymbol{\theta})$ be the joint probability measure induced by the nDP for the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$, and let $P^{LK}(\boldsymbol{\theta})$ be the corresponding joint measure under the LK truncation. Then

$$\begin{aligned}
\int \left| P^{LK}(\mathbf{y}) - P^{\infty\infty}(\mathbf{y}) \right| d\mathbf{y} &\leq \int\int p(\mathbf{y}|\boldsymbol{\theta}) \left| P^{LK}(d\boldsymbol{\theta}) - P^{\infty\infty}(d\boldsymbol{\theta}) \right| d\mathbf{y} \\
&= \int \left| P^{LK}(d\boldsymbol{\theta}) - P^{\infty\infty}(d\boldsymbol{\theta}) \right| \\
&= 2 \sup_{A \in \Theta} \left| P^{LK}(A) - P^{\infty\infty}(A) \right|
\end{aligned}$$

where the last equality is due to Scheffe's lemma. This means that the total variation distance between the true and approximated marginal densities can be bounded by the total variation distance between the priors. Now,

$$\begin{aligned}
\sup_{A \in \Theta} \left| P^{LK}(A) - P^{\infty\infty}(A) \right| &\leq 2 \left(1 - \mathbb{P}\left[\zeta_j \leq K - 1 \; \forall \, j, \xi_{ij} \leq L - 1 \; \forall \, i, j\right]\right) \\
&= 2 \left(1 - \mathbb{P}\left[\zeta_j \leq K - 1 \; \forall \, j\right] \times \right. \\
&\qquad \left. \mathbb{P}\left[\xi_{ij} \leq L - 1 \; \forall \, i, j | \zeta_j \leq K - 1 \; \forall \, j\right]\right)
\end{aligned}$$

Consider first the case $L = \infty$ and $K < \infty$. Then

$$\mathbb{P}\left[\xi_{ij} \leq L - 1 \; \forall \, i, j | \zeta_j \leq K - 1 \; \forall \, j\right] = 1$$

and

$$\mathbb{P}\left[\zeta_j \leq K-1 \; \forall \; j\right] = \mathbb{E}\left\{\left[\sum_{s=1}^{K-1} \pi_s^*\right]^J\right\}$$

$$\geq \left[\sum_{s=1}^{K-1} \mathbb{E}(\pi_s^*)\right]^J$$

by Jensen's inequality. Now,

$$\mathbb{E}(\pi_s^*) = \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{s-1} \;\Rightarrow\; \sum_{s=1}^{K-1} \mathbb{E}(\pi_s^*) = 1 - \left(\frac{\alpha}{1+\alpha}\right)^{K-1}$$

And therefore

$$\mathbb{P}\left[\zeta_j \leq K-1 \; \forall \; j\right] \geq \left[1 - \left(\frac{\alpha}{1+\alpha}\right)^{K-1}\right]^J$$

If $L < \infty$ and $K = \infty$. Then

$$\mathbb{P}\left[\zeta_j \leq K-1 \; \forall \; j\right] = 1$$

and

$$\begin{aligned}
\mathbb{P}\left[\xi_{ij} \leq L-1 \; \forall \; i,j | \zeta_j \leq K-1 \; \forall \; j\right] &= \mathbb{P}\left[\xi_{ij} \leq L-1 \; \forall \; i,j\right] \\
&= \sum_{(m_1,\ldots,m_J)\in C_J} \mathbb{P}\left[\xi_{ij} \leq L-1 \; \forall \; i,j | (m_1,\ldots,m_J)\right] \times \\
&\qquad \mathbb{P}\left[(m_1,\ldots,m_J)\right]
\end{aligned}$$

where $(m_1,\ldots,m_J) \in C_J$ is an assignment of J distributions to atoms $\{G_j^*\}_{k=1}^{\infty}$ such that there are m_1 distinct distributions appearing only once, m_2 that occur exactly twice and so on, and C_J is the set of all such possible assignments. From Antoniak (1974)

$$\mathbb{P}\left[(m_1,\ldots,m_J)\right] = \frac{J!\Gamma(J+\alpha)}{\prod_{j=1}^J m_j! j^{m_j}} \frac{\alpha^{\sum_{j=1}^J m_j}}{\Gamma(\alpha)}$$

and since $\{G_k^*\}_{k=1}^K$ are in turn independent samples from a DP,

$$\begin{aligned}
\mathbb{P}\left[\xi_{ij} \leq L-1 \; \forall \; i,j | (m_1,\ldots,m_J)\right] &= \prod_{j=1}^{J} \left\{ \mathbb{E}\left[\left(\sum_{l=1}^{L-1} w_{l1}^* \right)^{jn} \right] \right\}^{m_j} \\
&\geq \prod_{j=1}^{J} \left[\left(\sum_{l=1}^{L-1} \mathbb{E}(w_{l1}^*) \right)^{jn} \right]^{m_j} \\
&= \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{n \sum_{j=1}^{J} j m_j} \\
&= \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{nJ}
\end{aligned}$$

since $\sum_{j=1}^{J} j m_j = J$ for any configuration $(m_1,\ldots,m_J)$. Therefore,

$$\begin{aligned}
\mathbb{P}\left[\xi_{ij} \leq L-1 \; \forall \; i,j | \zeta_j \leq K-1 \; \forall \; j\right] &\geq \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{nJ} \sum_{(m_1,\ldots,m_J) \in C_J} \mathbb{P}\left[(m_1,\ldots,m_J)\right] \\
&= \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{nJ}
\end{aligned}$$

Finally, the case $K < \infty$ and $L < \infty$ combines both results. As before

$$\mathbb{P}\left[\zeta_j \leq K-1 \; \forall \; j\right] \geq \left[1 - \left(\frac{\alpha}{1+\alpha} \right)^{K-1} \right]^{J}$$

Since K is finite, the expressions of Antoniak (1974) cannot be used in this case. However, we do not need an explicit expression for $\mathbb{P}\left[(m_1,\ldots,m_J)\right]$ since we only need its sum, which is 1. Therefore

$$\mathbb{P}\left[\xi_{ij} \leq L-1 \; \forall \; i,j | \zeta_j \leq K-1 \; \forall \; j\right] \geq \left[1 - \left(\frac{\beta}{\beta+1} \right)^{L-1} \right]^{nJ}$$

as before.

Appendix D

Pólya urn schemes for the nDP

For the nDP, it is possible to explicitly integrate out the random distributions $\{G_k^*\}_{k=1}^{\infty}$ to obtain Pólya urn schemes on $(\boldsymbol{\Theta}, \mathcal{B})$. Conditional on the top level indicators $\boldsymbol{\zeta}$, all observation within the same cluster are iid samples from a random distribution following a $\mathsf{DP}(\beta H)$. Therefore, marginalizing the unknown distribution, the full conditional distribution for one single draw is given by

$$(\boldsymbol{\theta}_{ij}|\zeta_j = k, \boldsymbol{\zeta}_j^-, \boldsymbol{\theta}_{ij}^-) \sim \sum_{(i',j')\in\mathcal{N}_j^k} \frac{1}{\beta + |\mathcal{N}_j^k| + n_j - 1}\delta_{\boldsymbol{\theta}_{i'j'}} + \sum_{i'\neq i} \frac{1}{\beta + |\mathcal{N}_j^k| + n_j - 1}\delta_{\boldsymbol{\theta}_{i'j}} + \frac{1}{\beta + |\mathcal{N}_j^k| + n_j - 1}H \tag{D.1}$$

where $\mathcal{N}_j^k = \{(i', j') : \zeta_{j'} = k\ ,\ j' \neq j\}$, $|\mathcal{N}|$ denotes the number of elements in $\mathcal{N}$ and the $-$ superscript denotes the corresponding vector with the subscripted component removed. The first term in the expression represents the probability of $\boldsymbol{\theta}_{ij}$ being equal to some observation belonging to the same cluster but not the same center j, the second is the probability of it being equal to some other observation in center j and the third is the probability of it being different from all other observations. Therefore, this Pòlya urn allows us to sample $\boldsymbol{\xi}|\boldsymbol{\zeta}$.

In principle, we can use a similar argument to construct a sampler for $\boldsymbol{\zeta}$. It is easy to see that the joint distribution for $(\boldsymbol{\theta}_{1j}, \dots m\boldsymbol{\theta}_{n_j,j})$ is given by

$$(\boldsymbol{\theta}_{1j}, \dots, \boldsymbol{\theta}_{n_j,j} | \zeta_j = k, \boldsymbol{\zeta}_j^-, \boldsymbol{\theta}_j^-) \sim \prod_i^{n_j} \left[\sum_{(i',j') \in \mathcal{N}_j^k} \frac{1}{\beta + |\mathcal{N}_j^k| + i - 1} \delta_{\boldsymbol{\theta}_{i'j'}} + \sum_{i'=1}^{i} \frac{1}{\beta + |\mathcal{N}_j^k| + i - 1} \delta_{\boldsymbol{\theta}_{i'j}} + \frac{\beta}{\beta + |\mathcal{N}_j^k| + i - 1} H \right] \tag{D.2}$$

Therefore, integrating over the unknown indicators $\boldsymbol{\zeta}_j^-$ we get

$$\begin{aligned}(\boldsymbol{\theta}_{1j}, \dots, \boldsymbol{\theta}_{n_j,j} | \boldsymbol{\theta}_j^-) \sim \sum_{k \neq j} \frac{1}{\alpha + J - 1} \prod_i^{n_j} & \left[\sum_{(i',j') \in \mathcal{N}_j^k} \frac{1}{\beta + |\mathcal{N}_j^k| + i - 1} \delta_{\boldsymbol{\theta}_{i'j'}} \right. \\ & \left. + \sum_{i'=1}^{i} \frac{1}{\beta + |\mathcal{N}_j^k| + i - 1} \delta_{\boldsymbol{\theta}_{i'j}} + \frac{\beta}{\beta + |\mathcal{N}_j^k| + i - 1} H \right] \\ & + \frac{\alpha}{\alpha + J - 1} \prod_{i=1}^{n_j} \left[\sum_{i'=1}^{i-1} \frac{1}{\beta + i - 1} \delta_{\boldsymbol{\theta}_{i'j}} + \frac{\beta}{\beta + i - 1} H \right]\end{aligned} \tag{D.3}$$

In theory, this new Pólya urn allows us to sample $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ simultaneously. However, as the sample sizes $n_1, \dots, n_J$ grow, the number of terms involved in each of the sums increases exponentially, making the use of D.3 impractical for real applications.

A slightly different approach is to generate a Pólya urn in $(\mathbf{P}, \mathcal{C})$ (remember the nomenclature from Sections 1.1 and 2.1.2), which is simply

$$P(G_j | \zeta_j^-) = \sum_{j' \neq j} \frac{1}{\alpha + 1} \delta_{G_{j'}} + \frac{\alpha}{\alpha + 1} \mathsf{DP}(\beta H) \tag{D.4}$$

Combining (D.1) and (D.4) we can generate an alternative sampler that truncates the distributional atoms G_k^* but not the upper level stick-breaking construction in

(2.2). Unfortunately, the trick behind the retrospective sampler in Roberts and Papaspiliopoulos (2007) cannot be used in this case.

Appendix E

Proof of corollary 1

By Bayes theorem,

$$\begin{aligned}\lim_{K,L\to\infty} p^{LK}(\boldsymbol{\theta}|\mathbf{y}) &= \lim_{K,L\to\infty} \frac{p(\mathbf{y}|\boldsymbol{\theta})p^{LK}(\boldsymbol{\theta})}{p^{LK}(\mathbf{y})} \\ &= \frac{p(\mathbf{y}|\boldsymbol{\theta})\lim_{K,L\to\infty} p^{LK}(\boldsymbol{\theta})}{\lim_{K,L\to\infty} p^{LK}(\mathbf{y})} \\ &= p^{\infty\infty}(\boldsymbol{\theta}|\mathbf{y})\end{aligned}$$

Appendix F

Reordering pairing probabilities

This algorithm is inspired by the ideas underlying hierarchical clustering and microarray analysis, and relies on the interpretation of pairing probabilities as inverse distances. In principle, it can help in interpreting the results from any soft-clustering algorithm. In what follows, we assume that there are J distributions labeled from 1 to J and that p_{ij} is a point estimate of the probability of distributions i and j being in the same cluster. The algorithm follows the steps:

- Initialize the sets $A_s = \{i^*, j^*\}$ (the set of included labels), $\Omega = \{1, 2, \ldots, J\}$ and $A_s^c = \Omega / A_s$ (the set of non-included labels) where $(i^*, j^*) = \arg\max_{i,j}\{p_{i,j}\}$. That is, the first two distributions are those that happen to be closer to each other in the set.

- For all $j \in A_s^c$, calculate p^*_{j,A_s}, the "inverse distance" of each element of A_s^c to the set A_s, which is given by $p^*_{j,A_s} = \max_{j \in A_s}\{p_{ij}\}$. That is, the distance between any non-included element and the set of included elements is the distance to the closest element.

- Let $A_{s+1} = A_s \cup \{j^*\}$ and $A^c_{s+1} = \Omega / A_{s+1}$ where $j^* = \arg\max_{j \in A_s^c}\{p^*_{j,A_s}\}$. Therefore, the next element to be added is the one closest to the current set of

included elements.

- Repeat (2) until A_s^c is empty.

Any possible tie is broken randomly, and the permutation of the labels is given by the order in which they have been chosen to build the sets $\{A_J\}$.

Appendix G

Covariance structure in the discrete-time DDP

Note that our model implies that $y_{it}|\boldsymbol{\theta}_{it}, \sigma_i^2 = \mathbf{F}_t\boldsymbol{\theta}_{it} + \epsilon_{it}$ with errors $\epsilon_{it} \sim \mathsf{N}(0, \sigma_i^2)$ independent for every pair (i, t). Then

$$\begin{aligned}\mathbb{C}\text{ov}(y_{i,t}, y_{i',t+k}) = \mathbb{C}\text{ov}(\mathbf{F}_t\boldsymbol{\theta}_{it}, \mathbf{F}_{t+k}\boldsymbol{\theta}_{i',t+k}) + \mathbb{C}\text{ov}(\epsilon_{it}, \mathbf{F}_{t+k}\boldsymbol{\theta}_{i',t+k}) \\ + \mathbb{C}\text{ov}(\mathbf{F}_t\boldsymbol{\theta}_{it}, \epsilon_{i',t+k}) + \mathbb{C}\text{ov}(\epsilon_{it}, \epsilon_{i',t+k})\end{aligned}$$

where the last three terms are zero as long as either $i \neq i'$ or $k \neq 0$. Therefore

$$\mathbb{C}\text{ov}(y_{i,t}, y_{i',t+k}) = \mathbf{F}_t\mathbb{C}\text{ov}(\boldsymbol{\theta}_{it}, \boldsymbol{\theta}_{i',t+k})\mathbf{F}'_{t+k}$$

Now,

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}_{it}\boldsymbol{\theta}_{i',t+k}) &= \mathbb{E}\left(\sum_{l=1}^{\infty} w_l^*\delta_{(\boldsymbol{\theta}_{lt}^*)}\sum_{l=1}^{\infty} w_l^*\delta_{(\boldsymbol{\theta}_{l,t+k}^*)}\right) \\ &= \mathbb{E}\left(\sum_{l=1}^{\infty}\sum_{r=1}^{\infty} w_l^* w_r^*\delta_{(\boldsymbol{\theta}_{lt}^*)}\delta_{(\boldsymbol{\theta}_{r,t+k}^*)}\right) \\ &= \mathbb{E}\left(\sum_{l=1}^{\infty} w_l^{*2}\delta_{(\boldsymbol{\theta}_{lt}^*)}\delta_{(\boldsymbol{\theta}_{l,t+k}^*)}\right) + \mathbb{E}\left(\sum_{l=1}^{\infty}\sum_{r=1,r\neq l}^{\infty} w_l^* w_r^*\delta_{(\boldsymbol{\theta}_{lt}^*)}\delta_{(\boldsymbol{\theta}_{r,t+k}^*)}\right) \\ &= \frac{1}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}_{1t}^*\boldsymbol{\theta}_{1,t+k}^*) + \frac{\alpha}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}_{1t}^*)\mathbb{E}(\boldsymbol{\theta}_{1,t+k}^*)\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{V}(\boldsymbol{\theta}_{it}, \boldsymbol{\theta}_{i',t+k}) &= \mathbb{E}(\boldsymbol{\theta}_{it}\boldsymbol{\theta}_{i',t+k}) - \mathbb{E}(\boldsymbol{\theta}_{it})\mathbb{E}(\boldsymbol{\theta}_{i',t+k}) \\
&= \frac{1}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}^*_{1t}\boldsymbol{\theta}^*_{1,t+k}) - \frac{1}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}^*_{1t})\mathbb{E}(\boldsymbol{\theta}^*_{1,t+k}) \\
&= \frac{1}{1+\alpha}\mathbb{C}\text{ov}(\boldsymbol{\theta}^*_{1t}, \boldsymbol{\theta}^*_{1,t+k}) \\
&= \left[\prod_{r=1}^{k} \mathbf{G}_{t+k-r+1}\right] \mathbb{V}(\boldsymbol{\theta}^*_t)
\end{aligned}$$

Since the pair $(\boldsymbol{\theta}^*_{1t}, \boldsymbol{\theta}^*_{1,t+k})$ is sampled from the baseline measure K_0.

Bibliography

Aguilar, O., Huerta, G., Prado, R., and West, M. (1999). Bayesian inference on latent structure in time series. In A. D. J.O. Berger, J.M. Bernardo and A. Smith, eds., *Bayesian Statistics 6*, 3–26. Oxford University Press.

Ait-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica* **64**, 527–560.

Ait-Sahalia, Y. and Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics* **116**, 9–47.

Albert, J. and Chib, S. (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.

Assuncao, R. (2003). Space varying coefficient models for small area data. *Environmetrics* **14**, 453–473.

Baird, D., Wilcox, A., Weinberg, C., Kamel, F., McConnaughey, D., Musey, P., and Collins, D. (1997). Preimplantation hormonal differences betweeen the conception and non-conception menstrual cycles of 32 normal women. *Human Reproduction* **12**, 2607–2613.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC.

Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of distributions in nonparametric problems. *The Annals of Statistics* **27**, 536–561.

Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for dirichiet process mixture models. *Journal of the American Statistical Association* **98**, 224–235.

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika* **92**, 419–434.

Bigelow, J. L. and Dunson, D. B. (2007). Posterior simulation across nonparametric models for functional clustering. *Journal of the Royal Statistical Society, Series B.* **, under revision.**

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Binder, D. A. (1981). Approximations to Bayesian clustering rules. *Biometrika* **68**, 275–285.

Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* **81**, 637–654.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1**, 121–144.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **31**, 307–327.

Box, G. E. P. and Jenkins, G. M. (1974). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 2nd edn.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of American Statistical Association* **93**, 961–976.

Burgess, J. F., Lourdes, V., and West, M. (2000). Profiling substance abouse provider trends in health care delivery systems. *Health Services and Outcomes Research Methodology* **1**, 253–276.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

Chaveesuk, R., Srivaree-Ratana, C., and Smith, A. (1999). Alternative neural network approaches to corporate bond rating. *Journal of Engineering Valuation and Cost Analysis* **2**, 117–131.

Chib, S. and Hamilton, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator. *Statistical Science* **6**, 404–419.

Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance* **16**, 519–542.

Cont, R. and Tankov, P. (2003). *Financial modelling with jump processes*. Chapman & Hall.

Dahl, D. (2003). An improved merge-split sampler for conjugate Dirichlet process mixture models. Tech. rep., Department of Statistics, University of Winsconsin.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer, Berlin.

DeIorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An anova model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

Delbaen, F. and Schachermayer, W. (2006). *The Mathematics of Arbitrage*. Springer-Verlag.

Diaconis, P. and Freedman, D. (1986a). On inconsistent Bayes estimates of location. *The Annals of Statistics* **14**, 68–87.

Diaconis, P. and Freedman, D. (1986b). On the consistency of Bayes estimates. *The Annals of Statistics* **14**, 1–26.

Doksum, K. A. (1974). Tail free and neutral random probabilities and their posterior distributions. *Annals of Probability* **2**, 183–201.

Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Coloques Internationaux du Centre National de la Recherche Scientifique, no. 13, 23–37. Centre National de la Recherche Scientifique.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **, *in press***.

Duffie, D., Pan, J., and Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68**, 1343–1376.

Dunson, D. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* **100**, 618–627.

Dunson, D. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7**, 551–568.

Dunson, D. B., Herring, A. H., and Mulheri-Engel, S. A. (2007a). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the Royal Statistical Society* ***, In press***.

Dunson, D. B. and Park, J.-H. (2007). Kernel stick-breaking processes. Tech. rep., Institute of Statistics and Decision Sciences - Duke University.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007b). Bayesian density regression. *Journal of the Royal Statistical Society, Series B.* **69**, 163–183.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science* **11**, 89–121.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation. *Econometrica* **50**, 987–1008.

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 West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* **90**, 577–588.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In D. Dey, P. Müller, and D. Sinha, eds., *Practical nonparametric and semiparametric Bayesian statistics*, 1–22. Springer - Verlag (Berlin, New York).

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica* **14**, 715–745.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* **50**, 201–220.

Fan, J. Q., Hickman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of American Statistical Association* **90**, 141–150.

Fan, J. Q., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

Fan, J. Q. and Yim, T. H. (2004). A cross validatation 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.

Fidler, A. T. and Bernstein, J. (1999). Infertility: From a personal to a public health problem. *Public Health Reports* **114**, 494–511.

Fisher, R. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society* **85**, 87–94.

Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.

Gelfand, A. E., Guindani, M., and Petrone, S. (2007). Bayesian nonparametric modelling for spatial data using Dirichlet processes. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 8*, 1–26. Oxford University Press.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test* **13**, 263–312.

Ghosal, S., Ghosh, J., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27**, 143–158.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. New York: Springer-Verlag.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62**, 10–18.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Griffin, J. E. and Steel, M. F. J. (2006a). Nonparametric inference in time series problems. In *Valencia Statistics 8*.

Griffin, J. E. and Steel, M. F. J. (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.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **101**, 18–29.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.

Hjort, N. (2000). Bayesian analysis for a generalized Dirichlet process prior. Tech. rep., University of Oslo.

Huang, Z., Chen, H., Hsu, C., and Chen, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems* **37**, 543–558.

Hull, J. (2005). *Options, Futures and Other Derivatives*. John Wiley & Sons.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

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

Jacod, J. and Shiryaev, A. (1987). *Limit Theorems for Stochastic Processes*, vol. 288 of *Grundlehren der mathematischen Wissenschaften*. Springer - Verlag, Berlin.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *Journal of business and Economic Statistics* **12**, 371–389.

Jain, S. and Neal, R. M. (2000). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Tech. rep., Department of Statistics, University of Toronto.

Kacperczyk, M., Damien, P., and Walker, S. G. (2003). A new class of Bayesian semiparametric models with applications to option pricing. Tech. rep., University of Michigan Bussiness School.

Kim, C. J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics* **60**, 1–22.

Kleinman, K. and Ibrahim, J. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine* **17**, 2579–2596.

Kottas, A., Branco, M. D., and Gelfand, A. E. (2002). A nonparametric Bayesian modeling approach for cytogenetic dosimetry. *Biometrics* **58**, 593–600.

Kottas, A. and Gelfand, A. E. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* .

Lamon, E. I., Carpenter, S., and Stow, C. (1998). Forecasting pcb concentrations in lake michigan salmonids: A dynamic linear model approach. *Ecological Applications* **8**, 659–668.

Lau, J. W. and Green, P. (2006). Bayesian model based clustering procedures. Tech. rep., Department of Mathematics, University of Bristol.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics* **20**, 1222-1235.

Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics* **22**, 1161–1176.

Laws, D. J. and O'Hagan, A. (2002). A hierarchical Bayes model for multilocation auditing. *Journal of the Royal Statistical Society, Series D* **51**, 431–450.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2005). Mixtures of g-priors for Bayesian variable selection. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics* **12**, 351–357.

Luan, Y. and Li, H. (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. *Commnunications 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*, 50–55.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Tech. rep., Ohio State University, Department of Statistics.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *The Canadian Journal of Statistics* **27**, 251–267.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

Mauldin, R., Sudderth, W. D., and Williams, S. (1992). Pólya trees and random distributions. *The Annals of Statistics* **20**, 1203–1221.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

Melick, W. and Thomas, C. (1997). Recovering an asset's implied pdf from option prices: an application to crude oil during the Gulf War. *Journal of Financial and Quantitative Analysis* **32-1**, 91–115.

Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* **4**, 141–183.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68**, 179–199.

Mukhopadhyay, S. and Gelfand, A. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Müller, P., Quintana, F., and Rosner, G. (2004). Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–.

Ongaro, A. and Cattaneo, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics and Probability Letters* **67**, 33–45.

Paciorek, C. J. (2006). Misinformation in the conjugate prior for the linear model with implications for free-knot spline modelling. *Bayesian Analysis* **1**, 375–383.

Paddock, S. M. (1999). *Randomized Pólya Trees: Bayesian Nonparametrics for Multivariate Data Analysis*. Ph.D. thesis, Duke University, Durham, N. C.

Paddock, S. M., Ruggeri, F., Lavine, M., and M., W. (2003). Randomized Pólya tree models for nonparametric Bayesian inferene. *Statistia Sinica* **13**, 443–460.

Panigirtzoglou, N. and Skiadopoulos, G. (2004). A new approach to modeling the dynamics of implied distributions: Theory and evidence from the S&P500 options. *Journal of Banking and Finance* **28**, 1499–1520.

Pennell, M. L. and Dunson, D. B. (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* **62**, 1044–1052.

Pesaran, H., Smith, R., and Im, K. (1995). Dynamic linear models for heterogeneous panels. Cambridge Working Papers in Economics 9503, Faculty of Economics (formerly DAE), University of Cambridge. available at http://ideas.repec.org/p/cam/camdae/9503.html.

Petrone, S., Guindani, M., and Gelfand, A. E. (2006). Finite mixtures of spatial processes. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In T. S. Ferguson, L. S. Shapeley, and J. B. MacQueen, eds., *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, 245–268. Hayward, CA:IMS.

Protter, P. (1990). *Stochastic Integration and Differential Equations*. Springer-Verlag.

Quintana, F. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B.* **65**, 557–574.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory.* MIT Press, Cambridge.

Ramoni, M., Sebastiani, P., and Kohane, P. (2002). Cluster analysis of gene expression dynamics. *Procedings of the National Academi of Sciences* **99**, 9121–9126.

Rasiel, E. (2003). *Hedging with a Smile: A Behavioral Explanation for the Volatility Put Skew.* Ph.D. thesis, Fuqua, Duke University.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning.* The MIT Press.

Ray, S. and Mallick, B. K. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society, Series B.* **68**, 305–332.

Rebonato, R. (2004). *Volatility and Correlation: The Perfect Hedger and the Fox.* John Wiley & Sons.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53**, 233–243.

Roberts, G. and Papaspiliopoulos, O. (2007). Retrospective Markov chain Monte Carlo. *Biometrika* ***, In press.***

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.

Silverman, B. (1986). *Density Estimation.* Chapman and Hall, London.

Tang, Y. and Ghosal, S. (2006). A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities. Tech. rep., Department of Statistics - North Carolina State University.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Doklady* **4**, 1035–1038.

Truong, Y., Kooperberg, C., and Stone, C. (2005). *Statistical modeling with spline functions: Methodology and theory*. Springer.

Tu, C. (2006). *Bayesian nonparametric modeling using Lévy process priors with applications for function estimation, time series modeling and spatio-temporal modeling*. Ph.D. thesis, Institute of Statistics and Decision Sciences, Duke University.

Uhlig, H. (1997). Bayesian vector-autoregressions with stochastic volatility. *Econometrica* **65**, 59–73.

Venners, S. A., Wang, X., Chen, C., Wang, L., Dafang, C., Guang, W., Huang, A., Ryan, L., O'Connor, J. F., Lasley, B., Overstreet, J., Wilcox, A., and Xiping, X. (2004). Paternal smoking and pregnancy loss: A prospective study using a biomarker of pregnancy. *American Journal of Epidemiology* **159**, 993–1001.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Vidakovic, B. (1999). *Statistical modeling by wavelets*. New York: Wiley.

Wakefield, J., Zhou, C., and Self, S. (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*, 721–732. Oxford University Press.

Walker, S. G. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society, Series B* **63**, 811–821.

Walker, S. G. and Mulliere, P. (1997). Beta-Stacey processes and a generalization of the Pólya-urn scheme. *Annals of Statistics* **25**, 1762–1780.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B* **93**, 159–174.

West, M. (1995). Bayesian time series: Models and computations for the analysis of time series in the physical sciences. In *XV Workshop on Maximum Entropy and Bayesian Methods*.

West, M. (1997). Time series decomposition. *Biometrika* **84**, 489–494.

West, M., Aguilar, O., and Lourdes, V. (1998). Va hospital quality monitors: 1988-1997. Tech. rep., Duke University - Institute of Statistics and Decision Sciences.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer - Verlag, New York, second edition edn.

West, M., Prado, R., and Krystal, A. (1999). Evaluation and comparison of EEG traces: Latent structure in non-stationary time series. *Journal of the American Statistical Association* **94**, 1083–1095.

Wilcox, A. J., Weinberg, C., O'Connor, J. F., Baid, D. D., Schlatterer, J., E., C. R., Armstrong, E. G., and Nisula, B. C. (1998). Incidence of early loss of pregnancy. *New England Journal of Medicine* **319**, 189–194.

Wolpert, R. and Ickstadt, K. (2004). Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes. *Inverse Problems* **20**, 6, 1759–1771.

Wolpert, R. L. and Ickstadt, K. (1998a). Poisson/Gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.

Wolpert, R. L. and Ickstadt, K. (1998b). Simulation of Lévy random fields. In D. Dey and P. Müller, eds., *Practical nonparametric and semiparametric Bayesian statistics*, 227–242. Springer.

Wolpert, R. L., Ickstadt, K., and Hansen, M. B. (2003). A nonparametric Bayesian approach to inverse problems. In J. Bernardo, M. Bayarri, A. Dawid, J. Berger, D. Heckerman, A. Smith, and M. West, eds., *Bayesian Statistics 7*. Oxford University Press.

Wu, H. and Zhang, J. T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of American Statistical Association* **97**, 883–897.

Biography

Abel Rodríguez was born in Caracas, Venezuela, on July 16, 1975. After graduating from college with degrees in Industrial Engineering from Universidad Simón Bolivar and Law from Universidad Católica Andrés Bello in Venezuela, he pursued his passion for applied mathematics and completed a Master of Sciences in Statistics at Universidad Simón Bolívar. There, he worked with Luis Raúl Pericchi on Bayesian objective model selection and dynamic linear models. Following the completion of his master, Abel took charge of his families business for four years while also serving as visiting assistant professor at Universidad Simón Bolívar. Having been bitten by the academic bug, he decided in 2003 to return to school to complete his Ph.D. in statistics at Duke University, where he also obtained a Master of Arts in Economics and a Certificate in Bioinformatics and Computational Biology.