Bayesian Inference for Models with Infinite-Dimensionally Generated Intractable Components

Isadora Antoniano Villalobos

School of Mathematics, Statistics and Actuarial Science

University of Kent

A thesis submitted for the degree of

Doctor of Philosophy

November 20, 2012

To my family...

Acknowledgements

I would like to acknowledge all the people that made this possible.

My friends and family here and there, thank you for being with me through good times and bad; for always reminding me of the things that are really important in life. Guada, Gio, Vero, Moni, Claudia, my friends before and my friends still. Giota, Vasilis, Guru and Jose, Eleanna, Gelly, always around, making sure I remember there is more to life than just the work. Vasso, my dear friend, without you I may not have survived this; you and Antonio were my family in exile, keeping me sane. Juan Carlos, you left but you never left me, and your friendship always brought a shine to my days. Cristiano, coffee just doesn't taste the same without you. Anton, Vero, Gaby, Riccardo, you walked me out of the tunnel and into the light, it was around you that I started writing and it is back to you that I go now. Ivonne, Enrique, Ruben, you are you are the foundation, the beginning, the love.

To all my professors, who along the way have become colleagues and friends, I thank you for all your teachings, trust and support. Raul, Eduardo, Manuel, always willing to give a good word for me, and more importantly, to me. Ramses, you know your special roll in this play, you started it!

To my supervisor, Prof. Stephen Walker, thank you for sharing your knowledge, your experience, your ideas; without you, this would not be.

Four years of work and research were funded by the Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexican government, through the scholarship 196371/304461.

Abstract

In recent years, great effort has been placed on the development of flexible statistical models, which can capture the rich and diverse structures found in real data. Complex models are often intractable, and they require non trivial techniques for inference. In the Bayesian setting, the most common intractability problem is related with normalizing constants which cannot be calculated directly. In this case, MCMC methods are a useful tool for posterior simulation of the model parameters, and many ideas have been developed to enable the construction of the chains with the desired stationary densities. Frequently, ideas applied for posterior simulation from doubly-intractable distributions involve an approximation error; general exact methods are only available for models in which both the data and the parameters take values in fnite-dimensional spaces.

In the present work we propose a novel idea, based on a series expansion representation of the intractable functions, to enable MCMC simulation for models in which either the data or the parameters are infinite-dimensional. We achieve this by introducing a suitable set of latent variables with unknown and possibly infinite dimension. The MCMC construction is then made for a tractable latent model, from which the density of interest can be recovered through marginalization.

We illustrate the applicability of the method in various situations. We show that the latent variable construction of the retrospective rejection sampler commonly known as exact simulation algorithm for diffusions, is a particular case of the latent variable construction we propose. We provide an idea for an alternative exact simulation and

inference scheme, through a Markov chain construction. We also present two related nonparametric mixture models, for time series and regression analysis. Their novelty is in the construction of the mixture weights, which gives them great flexibility but introduces an intractable component generated by the infinite-dimensional parameters; we show how our methodology can be applied to enable MCMC inference for these models. We also show how our ideas can be used for inference when the power likelihood for nonparametric mixture models is used; a problem which is of interest in many settings and, to our knowledge, has not been solved without the introduction of some approximation error.

Finally, we discuss the matter of Bayesian consistency for Markov models. Unrelated to the driving theme of the thesis, the problem naturally arises from some of the models studied. We make a first step towards a general result for strong consistency which can be used both for discretely observed diffusions and for the time series model we propose.

Contents

List of Figures

Introduction

The first word in the title of this work is Bayesian, we therefore begin by considering a basic Bayesian model: a likelihood function $f(y_{1:n}|\theta)$ for a sample of observations, $y_{1:n} := (y_i)_{i=1}^n$, each modelled as a realization of a random variable Y_i, taking values on a state space $\mathbb{Y}$; and some prior distribution Π for the parameter $\theta \in \Theta$. Bayesian inference is then carried out based on the posterior distribution

$$\Pi_n(\theta|y_{1:n}) \propto f(y_{1:n}|\theta)\Pi(\theta). \tag{1}$$

For simplicity of notation, we assume throughout that all densities exist, with respect to some reference measure on $\mathbb{Y}^n \times \Theta$. Furthermore, we freely denote by Π both prior distribution and density, allowing the interpretation to be inferred from the context; the same liberty is taken with the use of ν, which denotes the generic reference measure with respect to which densities are defined, on the adequate spaces.

The title of the thesis also mentions intractable components, because we focus on models for which the likelihood function has the representation

$$f(y_{1:n}|\theta) = g(y_{1:n}, \theta)h(y_{1:n}, \theta), \tag{2}$$

where the function g is tractable but h is not, either because there is no analytic expression for it or because its evaluation is too computationally expensive for any practical application, hence making it an intractable component. As will become clear in later chapters, we are using the term tractable in a wide sense, referring not necessarily to functions which can be evaluated directly, but to functions which can be addressed using methods previously established in the literature.

Finally, we are interested in models which are either nonparametric, or defined on infinite-dimensional state spaces, or both. In other words, either the parameter θ or each observation y_i, is an infinite-dimensional object. Thus, we say that the intractable component is infinite-dimensionally generated.

There is extensive literature regarding the problem of intractable components in statistical models (see e.g. DiCiccio *et al.*, 1997; Evans & Swartz, 1995; Smith, 1991). However, most of it is concerned with the approximation of intractable normalizing constants when both the observations and the parameters are finite-dimensional. Noteworthy exceptions are found in the context of discretely observed diffusions (see Sorensen, 2004), and models involving some nonparametric priors, as we mention in the next chapter; but the results are specific to the models studies and not applicable in other situations. Both for general and particular models, three mainstream approaches can be identified: analytic approximation, usually based on Laplace transforms and other mathematical representations; numerical integration or some form of adaptive quadrature method based on classical analysis techniques; and Monte Carlo simulation methods, which use samples drawn from the distribution of interest to estimate features of it. In all of these methods, choices must be made which determine the quality of approximation that can be achieved. In high-dimensional situations application of such methods may be very computationally demanding; in the infinite dimensional set-up, it is not clear how they could be employed.

The idea we propose is simple. Rather than trying to approximate the intractable component, h, we replace it by a latent structure based on a power series expansion.

We start by factorizing the likelihood function in a standard way

$$f(y_{1:n}|\theta) = \prod_{i=1}^{n} f(y_i|y_{1:i-1}), \tag{3}$$

where y_0 is considered as a fixed known point, an artifice to simplify notation. In this case, the tractable and intractable components of equation (2) can also be

factorized, and we have

$$f(y_i|y_{1:i-1},\theta) = g_i(y_{1:i},\theta)h_i(y_{1:i},\theta). \tag{4}$$

Assuming that each h_i can be represented by some adequate series expansion

$$h_i(y_{1:i},\theta) = \sum_{k_i=0}^{\infty} c_{i,k_i}(\theta)h_{i,k_i}(y_{1:i},\theta), \tag{5}$$

in term of a sequence $(h_{i,k_i})_{k_i\geq 0}$ of fully specified functions, we propose using the indices $k_{1:n} = (k_1,\ldots,k_n)$ as latent variables. We incorporate them into the likelihood expression (3), thus obtaining an extended model

$$f(y_{1:n},k_{1:n}|\theta) = \prod_{i=1}^{n} g_i(y_{1:i},\theta)c_{i,k_i}(\theta)h_{i,k_i}(y_{1:i},\theta). \tag{6}$$

The dependence of each of the functions g_i and h_{i,k_1} on the complete set of variables $y_{1:i}$ represents only the most general case. More commonly, the dependence structure assumed by the model, simplifies these expressions, so that only a fixed number of variables $y_{i-m:i} = (y_{i-m},\ldots,y_i)$ is required for their evaluation. For example, if independence between observations is assumed, then $f(y_i|y_{1:i-1}) = f(y_i)$, therefore $g_i(y_{1:i},\theta) = g_i(y_i,\theta)$ and $h_{i,k_i}(y_{1:i},\theta) = h_{i,k_i}(y_i,\theta)$. Similar simplifications apply when Markov dependence of some order is assumed.

While the representation of equation (5) may not always be available for an arbitrary function h_i, it covers a wide spectrum of the intractable component problems that can be found in the literature. In the following chapters, we present a variety of models for which the method works. We focus on two cases:

i) By defining $c_{i,k_i}(\theta) = [r(\theta)]^{k_i}/k_i!$ for all $i = 1,\ldots,n$, $k_i \in \mathbb{N}$, some fixed, known function $r : \Theta \to [0,\infty)$; and adequate functions $(h_{i,k_i})_{k_i\geq 0}$, we deal with an exponential intractable component,

$$h_i(y_{1:i},\theta) = \int \exp\{r(\theta)b_i(y_{1:i},\theta,\lambda)\}\mathrm{d}\nu(\lambda). \tag{7}$$

This is relevant in the context of inference for discretely observed diffusions (see e.g. Beskos *et al.*, 2006b), where λ is a continuous function and the reference measure ν is a Weiner measure.

Notice that, in this case,

$$f(k_i|y_{1:n},\theta) \propto \frac{[r(\theta)]^{k_i}}{k_i!}\int \left[b_i(y_{1:i},\theta,\lambda)\right]^{k_i} \mathrm{d}\nu(\lambda) \tag{8}$$

may be intractable. However, conditional on the observations $y_{1:i}$, the model parameter θ and the auxiliary variable λ, each latent variable k_i has a Poisson distribution with mean parameter $r(\theta)b_i(y_{1:i},\theta,\lambda)$.

ii) By making, for every $\theta \in \Theta$ and every $i = 1,\ldots,n$, $c_{i,0}(\theta) = 1$, $c_{i,1}(\theta) = \alpha$ and $c_{i,k_i}(\theta) = \alpha^{(k_i)}/k_i! := \alpha(\alpha+1)\ldots(\alpha+k_i-1)/k_i!$, for $k_i > 1$ and some known, fixed $0 < \alpha < 1$; and assuming

$$h_{i,k_i}(y_{1:i},\theta) = \left[1 - b_i(y_{1:i},\theta)\right]^{k_i}, \tag{9}$$

for some bounded function $b_i : \mathbb{Y}^i \times \Theta \to [0,1]$, we obtain an adequate representation for functions of the type

$$h_i(y_{1:n},\theta) = \frac{1}{\left[b_i(y_{1:n},\theta)\right]^{\alpha}}. \tag{10}$$

Intractable components of this form are common in the literature, and in the following chapters we illustrate this with three nonparametric models in the contexts of time series analysis, regression analysis and power-likelihood estimation for i.i.d. observations.

In this case, the conditional distribution of k_i given $y_{1:i}$ and θ is the negative binomial with parameters α and $1 - b_i(y_{1:i},\theta)$, i.e.

$$f(k_i|y_{1:i},\theta) = \frac{\Gamma(k_i+\alpha)}{k_i!\Gamma(\alpha)}\left[1 - b(y_{1:i},\theta)\right]. \tag{11}$$

In particular, when $\alpha = 1$, then $k_i|y_{1:i},\theta$ is a geometric random variable, and we can write

$$h_i(y_{1:i},\theta) = \mathbb{E}[k_i|y_{1:i},\theta] + 1. \tag{12}$$

Before we can apply standard techniques designed for inference involving functions of infinite dimensional objects, such as $h_{i,k_i}(y_{1:i},\theta)$ for the latent model (6), we introduce additional auxiliary variables.

Let $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$, be some measurable space and assume we can define, for each $i = 1, \ldots, n$ and $l = 1, \ldots, k_i$ a function $b_{i,l} : \mathbb{Y}^i \times \Theta \times \mathbb{S}^{k_i} \to [0, \infty)$ such that

$$h_{i,k_i}(y_{1:i}, \theta) = \int_{\mathbb{S}^{k_i}} \prod_{l=1}^{k_i} b_{i,l}(y_{1:i}, \theta, s_{i,1:k_i}) \, \mathrm{d}\nu(s_{i,1:k_i}). \tag{13}$$

Notice that we are again favouring simplicity in the notation, trusting that any ambiguity in the use of b to denote different functions is resolved by the variables involved. We use the same principle when using f to denote densities and Π to denote priors, regardless of the random variables and spaces on which they are defined.

Under assumption (13) we introduce a set of auxiliary variables $s_{1:n,1:k_i} = \{s_{i,l} : i = 1, \ldots, n; l = 1, \ldots, k_i\}$, and arrive at the extended latent model

$$f(y_{1:n}, k_{1:n}, s_{1:n,1:k_i}|\theta) = \prod_{i=1}^{n} g_i(y_{1:i}, \theta) c_{i,k_i}(\theta) \prod_{l=1}^{k_i} b_{i,l}(y_{1:i}, \theta, s_{i,1:k_i}), \tag{14}$$

from which the original likelihood (3) can be recovered by integrating over the $k_{1:n}$ and $s_{1:n,1:k_i}$. However, in this last expression, the dimension of the state space $\mathbb{S}^{k_i}$ of latent variables is itself random, as it depends on k_i. This poses an issue for inference which we resolve following the ideas of Godsill (2001). That is, we consider infinite-dimensional latent variables $s_{1:n,1:\infty}$, each $s_{i,1:\infty}$ defined on $(\mathbb{S}^\infty, \mathcal{B}(\mathbb{S}^\infty))$, and a full latent model

$$f(y_{1:n}, k_{1:n}, s_{1:n,1:\infty}|\theta) = g(y_{1:n}, \theta) \prod_{i=1}^{n} c_{i,k_i}(\theta) \left(\prod_{l=1}^{k_i} b_{i,l}(y_{1:i}, \theta, s_{i,1:k_i}) \right) \left(\prod_{l>k_i} \Pi(s_{i,l}) \right) \tag{15}$$

where $\Pi(s)$ denotes a completely known density on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$. At this point, inference can be carried out using basic MCMC methods, such as the Gibbs Sampler and the Metropolis-Hastings algorithm, by observing that, at any iteration of a Markov chain update scheme, the $k_{1:n}$ are given, so only a finite number $s_{1:n,1:k_i}$ of latent variables is needed.

Summarizing, we are concerned with inference for doubly intractable distributions with infinite-dimensionally generated intractable components; we achieve

it through the introduction of adequate infinite-dimensional latent variables. We do not present output diagnostics or study theoretical rates of convergence of the proposed Markov chain schemes. Our aim is simply to propose a means of making MCMC simulation possible for some doubly intractable models for which, to our knowledge, there is no available methodology. We acknowledge this is simply a starting point and much work may yet be done. However, as it stands, our ideas are widely applicable, and we illustrate this through a series of examples, ranging from univariate continuous processes to multivariate regression models and from independent and identically distributed observations to time series.

Only general ideas have been presented in this Introduction. Many details and considerations must be made concerning each specific example. We address them in the following Chapters, hoping that the illustrations will serve to clarify.

Outline of the thesis

In Chapter 1 we present some background material, relevant for the present work. It is divided into three sections. The first one provides a review of some Bayesian models currently used in the context of independent and identically distributed observations, discretely observed continuous-time Markov processes, time series analysis and regression analysis. The second gives a review of auxiliary variable constructions and MCMC methods, for simulation and posterior inference for complex and intractable models. The third is a brief exposition of current results on Bayesian consistency for i.i.d. observations.

The next four Chapters provide examples of intractable models for which our auxiliary variable approach is applicable. Chapter 2 focuses on discretely observed diffusions. In Chapter 3 we propose a nonparametric model for stationary time series, for which both the transition and the invariant densities have an infinite mixture representation. A similar model is developed in Chapter 4 in the context of nonparametric regression, for which the covariate space may include combinations of both continuous and discrete variables. Chapter 5 is concerned with inference for infinite mixture models, when a smoothed version of the likelihood, a power likelihood, is used. In each Chapter, we provide a latent model for which inference is feasible via MCMC posterior simulation for the model parameters.

Chapter 6 is somewhat different. We develop a new result for posterior consistency, in the context of Bayesian estimation of the transition density of a time homogeneous Markov process. This is not directly related with the latent model approach studied in this thesis, but it is relevant to some of the results and models presented in previous Chapters.

Each of the Chapters 2 to 6 ends with a discussion of the results and methods presented within, as well as some ideas for extending them. Chapter 7 provides a more general discussion of the overall results and methods developed in the thesis, some relations between the models studied in the previous Chapters, as well as some ideas for future work which involve the combination of some of the models and ideas found in different chapters.

Chapter 1

Background

In this Chapter, we present some background material on models and methods currently found in the literature. We do not intend to cover the subjects extensively, but rather to provide a context for the present work, as well as the basis over which we build our own models and results.

We begin with an overview of some statistical models. It is followed by a section on auxiliary variable schemes, resulting in latent models which make inference possible, simpler or more efficient. We then discuss some of the algorithms currently available for simulation and posterior inference via Markov Chain Monte Carlo methods. Finally, we define Bayesian posterior consistency and review some of the current results regarding asymptotic properties of Bayesian models.

1.1 Statistical Models

Our starting point is a sample, $y_{1:n} = (y_1, \ldots, y_n)$. For $i = 1, \ldots, n$, each observation, y_i, is considered as a realization of a random variable Y_i. Two elements define a Bayesian statistical model. First, the joint density f of $(Y_1, \ldots, Y_n)$, which characterizes the random mechanism generating the observations. Since this density is assumed to be unknown, a family $\mathcal{F}$ of density functions is defined, containing all "candidate" densities. The second element of the model is a probability measure Π over $\mathcal{F}$, describing the uncertainty about such mechanism and incorporating any prior belief about it. Bayes theorem can then be used to up-

date the prior into a posterior distribution, thus learning about the phenomenon of interest.

We distinguish here between two types of models, depending on the size of the $\mathcal{F}$ space and, therefore, the nature of the prior Π imposed on it. When each density $f_\theta \in \mathcal{F}$ can be indexed by some finite-dimensional parameter $\theta \in \Theta$, the prior Π is a probability measure on the parameter space Θ which, in turn, induces the prior on $\mathcal{F}$; this is known as a parametric model. A nonparametric model is defined when the prior Π is a probability measure defined on the space $\mathcal{P}$ of probability measures over $\mathbb{Y}$. If we consider the support of Π as the subset of $\mathcal{P}$ for which probability measures have a well defined density, this induces a prior on a functional space $\mathcal{F}$ of densities, too large to be indexed by what is commonly considered a parameter. In practice, it is common to use a representation for each density $f \in \mathcal{F}$ in terms of an infinite-dimensional parameter, over which the prior Π is defined. This induces a prior on $\mathcal{F}$ and on $\mathcal{P}$, hence Bayesian models with infinite dimensional parameters are also known as nonparametric.

The capacity of the model to explain a complex phenomenon about which little is known a priori, in other words, its flexibility, depends on the size of the space of densities $\mathcal{F}$ under consideration. For parametric models, the larger the dimension of θ, the larger the family of densities indexed by it. In the limit, an infinite dimensional parameter is sufficient for representing entire functional spaces. Therefore, nonparametric models are considered more flexible than parametric ones.

The problem of defining a prior Π on $\mathcal{F}$ is closely related to the problem of parametrizing the space, either by a finite or infinite-dimensional parameter θ. If the parameter space Θ is finite, the flexibility of the model, that is, the size of $\mathcal{F}$, depends on the parametrization itself. Diffusion models are an example of complex parametric models. For nonparametric models, a simpler representation may be found for each element of $\mathcal{F}$; the complexity in this case falls on the definition of the prior on the infinite-dimensional Θ.

There are many ways to specify the family $\mathcal{F}$. Here, we consider three of them, where the distinction is made with respect to the type of dependence structure assumed for the data:

i) Independent and identically distributed observations.

Each function $f \in \mathcal{F}$ is a density on the state space $\mathbb{Y}$, and for any sample size n, the likelihood function for $y_{1:n} = (y_i)_{i=1}^n$ can be represented as the n-fold product

$$f(y_{1:n}) = \prod_{i=1}^{n} f(y_i). \tag{1.1}$$

The main assumption in this case is that the random variables $(Y_i)_{i=1}^n$ are independent and identically distributed (i.i.d.) according to f. In fact, from the Bayesian point of view, the observations are only conditionally independent given their common density f, something related to the concept of exchangeability. However, we use the term i.i.d. observations to refer to this type of model, as is commonly done in the Bayesian literature.

We discuss some existing nonparametric models for i.i.d. observations in Section 1.1.1.

ii) Observations with a Markov type dependence.

Each function $f \in \mathcal{F}$ is a conditional density and for any sample size n, the likelihood function for $y_{1:n} = (y_i)_{i=1}^n$ is again a product,

$$f(y_{1:n}) = \prod_{i=1}^{n} f(y_i | y_{i-1}, y_{i-2}, \ldots, y_{i-m}). \tag{1.2}$$

The main assumption in this case is that the random variables $(Y_i)_{i=1}^n$ are dependent and each Y_i is conditionally independent on the rest, given $(Y_j)_{j=i-m}^{i-1}$. This is known as an order m Markov dependence structure, or simply Markov when $m = 1$. Some considerations regarding the initial m data points are needed in this case, as the expression (1.2) depends on $(y_{1-m}, \ldots, y_0)$. Unless otherwise stated, the initial points are assumed to be fixed and known. In other words, the first observations enter the likelihood expression as fixed, known quantities and not as realizations of random variables.

We present a large family of parametric models for this type of data and discuss some nonparametric models in Section 1.1.2.

iii) Observations dependent on covariates.

The sample is defined by pairs of data, $(y, x)_{1:n}$, where each y_i represents a realization of the variable of interest, while x_i, the covariate, provides additional information about the behaviour of Y_i. Each function $f \in \mathcal{F}$ is again a conditional density, and the likelihood function for a sample of size n takes the form

$$f(y_{1:n}|x_{1:n}) = \prod_{i=1}^{n} f(y_i|x_i). \tag{1.3}$$

The variables $(Y_i)_{i=1}^n$ are assumed to be independent, but the mechanism generating each Y_i is allowed to vary, depending on the value of the corresponding covariate x_i. Formally, the random variables $(Y_i)_{i=1}^n$ are conditionally independent given the $(x_i)_{i=1}^n$, and variables with common covariate values are i.i.d. The covariates may be modelled either as fixed or random values, however we consider here the case of non random covariates only.

Models of this type, used to capture the way in which each random variable Y_i depends on the covariate value x_i, are known as Regression models. In Section 1.1.3 we discuss some of the nonparametric regression models present in the literature.

1.1.1 Nonparametric Models for i.i.d. Observations

We begin by considering a parametric model for independent and identically distributed random variables. This provides what is perhaps the most basic and frequently used construction for the likelihood function of a sample of size n, as the n-fold product of a single function, evaluated at each data point y_i,

$$f(y_{1:n}|\theta) = \prod_{i=1}^{n} f(y_i|\theta). \tag{1.4}$$

Calling this a model for i.i.d. observations is arguably an abuse of terminology, since the likelihood expression (1.4) implies only that the observations arise from random variables which are conditionally independent, given the parameter θ. In fact, the assumption behind this model is somewhat milder than that of independence. It is enough to assume the observations possess a form of symmetry known as exchangeability.

To formalize and set the notation, assume each observation y_i is a realization of a measurable random variable Y_i defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, called the sample space, and taking values on a complete and separable metric space $(\mathbb{Y}, d)$, known as the state space, with Borel σ-algebra $\mathcal{B}(\mathbb{Y})$. The space $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ of probability measures over $\mathbb{Y}$, is again complete and separable under the metric of weak convergence, and we can therefore define a probability measure Π over it.

Let ν be a σ-finite measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ with respect to which densities are defined. In in the present work, most random variables take values on the p-dimensional Euclidian space $\mathbb{Y} \subseteq \mathbb{R}^p$, on a discrete space $\mathbb{Y} \subseteq \mathbb{Z}^p$, or a product of them. Therefore ν is the Lebesgue measure, the counting measure or a product of them.

Denote by $\mathcal{F}$ the set of density functions over $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$, and by $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\} \subset \mathcal{F}$ a set of densities parametrized by $\theta \in \Theta$. A random variable Y is distributed according to $P \in \mathcal{P}$ or has density $f \in \mathcal{F}$ if, for every $B \in \mathcal{B}(\mathbb{Y})$

$$\mathbb{P}[Y \in B] = P(B) = \int_B f(y)\mathrm{d}\nu(y). \tag{1.5}$$

Let $\mathbb{Y}^\infty$ denote the infinite product space of $\mathbb{Y}$, with corresponding Borel $\sigma-$algebra $\mathcal{B}(\mathbb{Y}^\infty) = (\mathcal{B}(\mathbb{Y}))^\infty$. For each probability measure $P \in \mathcal{P}$, we denote by P^∞ the corresponding product measure over $\mathbb{Y}^\infty$, with density f^∞.

Finally, let $\delta_y \in \mathcal{P}$ denote Dirac's delta measure on $y \in \mathbb{Y}$, that is, a probability measure with all the mass accumulated on y. The corresponding density is the indicator function $\mathbf{1}_{\{y\}} \in \mathcal{F}$, given by

$$\mathbf{1}_{\{y\}}(\tilde{y}) = \begin{cases} 1 & \text{if } \tilde{y} = y; \\ 0 & \text{otherwise.} \end{cases} \tag{1.6}$$

We are now ready to define the concept of exchangeability and present some models commonly used when this property is assumed.

Definition 1 (Exchangeability) *A finite set $(Y_i)_{i=1}^n$ of random variables is called exchangeable if and only if every permutation of them has the same joint distribution. A sequence $(Y_i)_{i\geq 1}$ is exchangeable if every finite subset is exchangeable.*

Exchangeability is a sensible assumption in many situations, when the order in which the observations are received and incorporated into the model does not affect the information they contain regarding the mechanism that generates them. The use of the term i.i.d. in this case is justified by the following theorem (see e.g. Schervish, 1995, Chapter 1), stating that a sequence of exchangeable random variables is conditionally i.i.d. given a probability measure known as de Finetti's measure, and vice versa.

Theorem 1 (de Finetti's Representation Theorem) *A sequence* $Y = (Y_i)_{i \geq 1}$ *of random variables taking values on* $\mathbb{Y}$ *is exchangeable if and only if there exists a probability measure* Π *over* $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ *such that, for any* $B = \bigotimes_{i=1}^{\infty} B_i \in \mathcal{B}(\mathbb{Y}^{\infty})$,

$$\mathbb{P}[Y \in B] = \mathbb{P}[Y_i \in B_i;\, i \geq 1] = \int_{\mathcal{P}(\mathbb{Y})} P^{\infty}(B)\Pi(\mathrm{d}P).$$

Furthermore, the de Finetti measure Π *for the sequence is unique and equal to the limit of the empirical distributions,*

$$\Pi = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}.$$

Therefore, if exchangeability is assumed, the the choice of a likelihood function given by expression (1.4) is justified, and a Bayesian model for exchangeable sequences can be represented in a hierarchical form as

$$\begin{aligned} Y_i &\overset{iid}{\sim} P; \\ P &\sim \Pi. \end{aligned} \tag{1.7}$$

In this case, we use $f_P \in \mathcal{F}$ to denote the density corresponding to a probability measure $P \in \mathcal{P}$.

When the model is parametric, we may write

$$\begin{aligned} Y_i &\overset{iid}{\sim} P_\theta; \\ \theta &\sim \Pi, \end{aligned} \tag{1.8}$$

since the prior on $(\Theta, \mathcal{B}(\Theta))$ induces a prior on the parametric family $\{P_\theta : \theta \in \Theta\} \subset \mathcal{P}$. The densities in this case are denoted by f_θ.

A variety of well known models are available in the parametric case. Nonparametric models are more complicated and here we present some of the ideas commonly used to define them.

Let us first consider a finite state space, $\mathbb{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_J\}$. Every probability $P \in \mathcal{P}$ can be expressed as

$$P = \sum_{j=1}^{J} w_j \delta_{\tilde{Y}_j}, \tag{1.9}$$

for some weights $0 \leq w_j \leq 1$ such that $\sum_j w_j = 1$ and points $\tilde{Y}_j = \tilde{y}_j \in \mathbb{Y}$. If the fixed points and weights are replaced by random variables, the distribution over them defines a prior probability measure Π over $\mathcal{P}$. A simple way to do so is to assume the points $(\tilde{Y}_j)_{j=1}^J$ are i.i.d. random variables taking values in $\mathbb{Y}$, and distributed according to some probability P_0. An independent probability measure may be defined on the simplex $\{w_1, \ldots, w_j \in (0,1) : \sum_j w_j = 1\}$. P_0 can be chosen as a simple parametric measure; the distribution for the weights requires a more careful selection, to guarantee that they add up to one. If both distributions have full support, the prior Π they induce on $\mathcal{P}$ will also have full support. A possible choice for the distribution of the weights is the Dirichlet distribution defined below.

Definition 2 (Dirichlet Distribution) *Let $\tilde{w}_1, \ldots, \tilde{w}_J$ be a set of random variables, such that $\tilde{w}_j \overset{ind}{\sim} \mathrm{Ga}(\gamma_j, 1)$, where $\gamma_j \geq 0$ for every $0 \leq j \leq J$ and $\sum_j \gamma_j > 0$. The Dirichlet distribution with parameter $\gamma_{1:J} = (\gamma_1, \ldots, \gamma_J)$ is the joint distribution of the random variables $(w_1, \ldots, w_J)$ defined by*

$$w_j = \frac{\tilde{w}_j}{\sum_{j'=1}^{J} \tilde{w}_{j'}},$$

and it is denoted by $(w_{1:J}) \sim \mathrm{Dir}(\cdot|\gamma_{1:J})$.

A generalization of this idea for an infinite sample space $\mathbb{Y}$ gives place to the Dirichlet Process, possibly the most widely known and used model in Bayesian nonparametrics.

1.1.1.1 Dirichlet Process and Stick-Breaking Priors

When the state space $\mathbb{Y}$ is not finite but countable, it is still possible to represent every probability measure over it as a weighted sum of Dirac's delta measures over the elements of the space,

$$P = \sum_{j=1}^{\infty} w_j \delta_{\tilde{Y}_j}, \tag{1.10}$$

However, a more careful prior specification on the weights $w = (w_j)_{j \geq 1}$ is required to guarantee that $\sum_{j=1}^{\infty} w_j = 1$ and $w_j > 0$ for every j. One idea is to use the representation commonly known as Stick-Breaking, in which the weights are defined through a sequence of independent Beta distributed variables, $v_j \sim \text{Be}(\alpha_j, \zeta_j)$, by making $w_1 = v_1$ and for $j > 1$,

$$w_j = v_j \prod_{j'<j} (1 - v_{j'}). \tag{1.11}$$

This multiplicative structure for the definition of the weights ensures that they add up to 1 (see Ishwaran & James, 2001), whenever the parameters for the Beta-distributed variables satisfy the condition

$$\sum_{j=1}^{\infty} \log(1 + \frac{\alpha_j}{\zeta_j}) = \infty. \tag{1.12}$$

The prior on the weights is complemented by an independent base measure P_0 from which the atoms $(\tilde{Y}_j)_{j \geq 1}$ are assumed to be independently distributed. The prior Π is the joint distribution for the weights and the atoms, or more formally, the probability measure on $\mathcal{P}$ induced by it. As with the finite state space case, Π has full support whenever P_0 does.

When $\alpha_j = 1$ and $\zeta_j = \zeta > 0$ for every j, this corresponds to a representation of the Dirichlet process given by Sethuraman (1994). This is, however, not the only characterization of the process, which owes its name to the first definition given by Ferguson (1973). He specified a set of finite dimensional distributions for a stochastic process, based on the Dirichlet distribution, and proved the Kolmogorov consistency conditions to guarantee the existence of the process.

Definition 3 (Dirichlet Process) *Let P_0 be a probability measure over $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$, and let $\zeta \in (0, \infty)$. A random measure $P \in \mathcal{P}$ has a Dirichlet Process distribution with parameter ζP_0 if for every $j = 1, 2, \ldots$ and every measurable partition $(\mathbb{Y}_1, \ldots, \mathbb{Y}_j)$ of $\mathbb{Y}$, the vector $(P(\mathbb{Y}_1), \ldots, P(\mathbb{Y}_j))$ has a Dirichlet distribution, $\text{Dir}(\cdot|\zeta P_0(\mathbb{Y}_1), \ldots, \zeta P_0(\mathbb{Y}_j))$. We denote this by $P \sim \text{DP}(\cdot|\zeta P_0)$.*

Intuitively, each path $P = (p_j)_{j \geq 1}$ of a Dirichlet process constitutes a probability measure on $\mathbb{Y}$, where each p_j represents the probability assigned to a point $\tilde{y}_j \in \mathbb{Y}$. Therefore, the law of the process constitutes a probability measure Π over $\mathcal{P}$.

Since its original introduction, the Dirichlet process has been widely studied and used. Its popularity is somewhat related to its many characterizations which, through adequate generalizations, allow the definition of new processes which can, in turn, be used as distributions for random probability measures. In the same paper where he introduced the process, Ferguson (1973) presented an equivalent definition, related to the Poisson-Dirichlet distribution (see Pitman, 1996, for more details) which is worth mentioning. However, we do not provide it, as it is not relevant to the present work.

Blackwell & MacQueen (1973) introduced yet another characterization of the Dirichlet process in terms of a generalized Polya urn. We present it here, because it allows the simulation of a sample from a Dirichlet process, through a closed expression for the predictive distribution.

Theorem 2 *Let $(Y_n)_{n \geq 1}$ be a sequence of random variables with distribution defined by a generalized Polya urn with parameter ζP_0, i.e.*

$$\mathbb{P}[Y_1 \in \cdot] = P_0(\cdot)$$

$$\mathbb{P}[Y_{n+1} \in \cdot \mid Y_1, \ldots, Y_n] = \frac{\zeta P_0(\cdot) + \sum_{i=1}^{n} \delta_{Y_i}(\cdot)}{\zeta + n}.$$

Then

(a) The sequence $\{P_n\}_{n \geq 1}$ of probability measures with support in $\mathbb{Y}$, defined by

$$P_n = \frac{\zeta P_0 + \sum_{i=1}^{n} \delta_{Y_i}}{\zeta + n}$$

converges a.s. to a discrete probability measure P supported on $\mathbb{Y}$.

(b) $P \sim \mathrm{DP}(\cdot \mid \zeta P_0)$.

(c) Given P, the random variables $Y_1, Y_2, \ldots$ are conditionally independent with distribution P.

We present below some well known results and properties of the Dirichlet process.

Lemma 1 *If $P \sim \mathrm{DP}(\cdot \mid \zeta P_0)$, then P is almost surely a discrete measure.*

This is clear from the stick-breaking representation of the process, but not from Definition 3. This gives evidence of the relevance of the various characterizations of the process, since each one may be more convenient for proving different properties.

Lemma 2 *Let $Y_{1:n}$ be a sample of size n from a Dirichlet process P with parameter ζP_0, i.e.*

$$Y_i | P \overset{iid}{\sim} P,$$
$$P \sim \mathrm{DP}(\cdot | \zeta P_0).$$

Then, the conditional distribution of P given the sample is again a Dirichlet process, i.e.

$$P \mid Y_{1:n} \sim \mathrm{DP}\left(\cdot \middle| \zeta P_0 + \sum_{i=1}^{n} \delta_{Y_i}\right).$$

In other words, if a Dirichlet process is used as a prior for a nonparametric model, then the posterior is also a Dirichlet process.

Different representations of the Dirichlet process have been generalized to define new nonparametric priors. We focus on the stick-breaking representation (1.10) and the construction of the weights given by (1.11). In the following chapters, we consider the Stick-breaking priors obtained when $\alpha_j = \alpha$ and $\zeta_j = \zeta$ for all j.

An alternative nonparametric model, is obtained when the weights are constructed based one single beta distributed random variable, $v \sim \text{Be}(\alpha, \zeta)$, through a geometric structure

$$w_j = v(1-v)^{j-1}. \tag{1.13}$$

This process, known as the geometric stick-breaking (GSB) prior, was defined by Fuentes-García *et al.* (2010), who prove it has the same support as the MDP model. Its proposed advantage would be a reduction in the variability of the weights, due to the simplification of their construction, thus improving estimation. In fact, the use of the GSB prior may be interpreted as the removal of a hierarchical level of the nonparametric model structure, achieved by substituting the weights of the Dirichlet process, by their expected values. More clearly, the expected values of the weights in the Dirichlet process are given by

$$\mathbb{E}[w_j] = \frac{1}{\zeta+1}\left(\frac{\zeta}{\zeta+1}\right)^{j-1}, \tag{1.14}$$

which is a reparametrization of expression (1.13), for the Geometric stick-breaking weights.

All these stick-breaking constructions, as well as other nonparametric priors based on normalized stochastic processes (see e.g. Lijoi *et al.*, 2005) share the limitation of assigning probability 1 to the space of discrete probability measures. This is enough when countable state spaces are considered, but when $\mathbb{Y}$ is uncountable, more flexible models are desirable. Many efforts have been made to define nonparametric priors supported on sets of continuous probability measures. The first and probably most used solution is once more a generalization of the Dirichlet process.

1.1.1.2 Nonparametric Mixture Models

Consider now an uncountable state space $\mathbb{Y}$ and a parametric family $\mathcal{K}_\Theta = (K_\theta : \theta \in \Theta) \subset \mathcal{F}$ of density functions over it. Notice that we have changed the notation, using K_θ to denote parametric densities, also called kernels, in order to clearly distinguish them from general, possibly nonparametric densities, denoted by f.

Basic results of linear algebra and functional analysis may be used to define the subspace $\mathcal{F}(\mathcal{K}_\Theta) \subset \mathcal{F}$ generated by $\mathcal{K}_\Theta$ as the set of all densities over $\mathbb{Y}$ which can be represented as a convex combination of elements of $\mathcal{K}_\Theta$. In other words,

$$\mathcal{F}(\mathcal{K}_\Theta) = \left\{ f = \sum_{j=1}^{\infty} w_j K_{\theta_j} : \forall\, j,\ K_{\theta_j} \in \mathcal{K}_\Theta, w_j > 0 \text{ and } \sum_{j=1}^{\infty} w_j = 1 \right\}. \quad (1.15)$$

In particular, when $\Theta = \mathbb{Y}$ and $K_\theta = \mathbf{1}_{\{\theta\}}$, then $\mathcal{F}(\mathcal{K}_\Theta) = \mathcal{F}_\mathcal{D}$ is the space of all discrete densities over $\mathbb{Y}$; but if every K_θ is a continuous density, then so is every $f \in \mathcal{F}(\mathcal{K}_\Theta)$.

As done before, for the finite and countable state space cases, a prior Π on $\mathcal{P}$ is induced by defining independent priors over the parameter space Θ and the simplex

$$\left\{ (w_j)_{j=1}^{\infty} \subset [0,1]^\infty : \sum_{j=1}^{\infty} w_j = 1 \right\}, \quad (1.16)$$

together with the choice of the parametric family of kernels $\mathcal{K}_\Theta$. Furthermore, Π assigns probability 1 to the subset of probability measures with densities in $\mathcal{F}(\mathcal{K}_\Theta)$. Therefore, in order to define a prior on the set $\mathcal{F}_\mathcal{C}$ of continuous density functions it is enough to choose a family of continuous kernels. A common choice is $K_\theta(\cdot) = \mathrm{N}(\cdot|\mu, \sigma^2)$, the normal density function with mean μ and variance σ^2, where $\theta = (\mu, \sigma)$. For $\Theta = \mathbb{R} \times \{\sigma\}$, where $\sigma > 0$ is any positive number, the family $\mathcal{K}_\Theta$ of normal densities is a basis for the subspace $\mathcal{F}_c \subset \mathcal{F}$ of continuous densities. Therefore, when the Gaussian kernel is used, $\mathcal{F}(\mathcal{K}_\Theta) = \mathcal{F}_\mathcal{C}$, since any continuous density over $\mathbb{Y}$ can be expressed as a convex combination of normal density functions.

From the previous section, we know that the Dirichlet process can be used to define the desired prior, by first defining an a.s. discrete random probability measure

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}, \quad (1.17)$$

on the parameter space Θ, which, in turn, induces a prior on $\mathcal{F}(\mathcal{K}_\Theta)$. We use the notation

$$f_P(y) = f(y|w_{1:\infty}, \theta_{1:\infty}) = \sum_{j=1}^{\infty} w_j K(y|\theta_j), \quad (1.18)$$

to make the dependence between each $f \in \mathcal{F}(\mathcal{K}_\Theta)$ and the choice of the weights and particles in P explicit.

This model is known as the Mixture of Dirichlet Process (MDP) prior and it can also be represented in a hierarchical way as

$$\begin{aligned} Y_i|\theta_j &\overset{iid}{\sim} K(\cdot|\theta_j), \\ \theta_j|P &\overset{iid}{\sim} P, \\ P &\sim \mathrm{DP}(\cdot|\zeta P_0). \end{aligned} \tag{1.19}$$

This MDP model was first introduced by Lo (1984), who noticed that a continuous density over the sample space $\mathbb{Y}$ can be defined as a convolution of measures, since

$$f_P(\cdot) = \int_\Theta K(\cdot|\theta)\,\mathrm{d}P(\theta), \tag{1.20}$$

defines a continuous density, whenever K is continuous, regardless of the choice of the mixing probability measure, P. If a prior is assigned to P which gives probability one to the set of discrete measures, the above equation becomes (1.18).

Different Stick-breaking priors on P result in diferent nonparametric mixture models, all of which can be represented by equations (1.18) and (1.17). Therefore, throughout this thesis, we use the term nonparametric mixture model with parametric kernel $K(\cdot|\theta)$ and a stick-breaking prior with parameters (α_j, ζ_j) and base measure P_0, to refer to the complete model

$$\begin{aligned} y_i|w_{1:\infty}, \theta_{1:\infty} &\overset{iid}{\sim} f_P \\ f_P(\cdot) &= \sum_{j=1}^{\infty} w_j K(\cdot|\theta_j); \\ w_1 = v_1 \quad \text{and} \quad w_j &= v_j \prod_{j'<j}(1 - v_{j'}), \forall\, j > 1; \\ v_j &\overset{iid}{\sim} \mathrm{Be}(\cdot|\alpha_j, \zeta_j); \\ \theta_j &\overset{iid}{\sim} P_0, \end{aligned} \tag{1.21}$$

The assumption $\alpha_j = \alpha$ and $\zeta_j = \zeta$ for all j is used to simplify notation, and all results can be extended for the more general choice of stick-breaking parameters.

Nonparametric mixture models with almost surely discrete random mixing probability measures are not the only way to define a prior over the set $\mathcal{F}$ of

continuous densities. An interesting alternative involves the use of normalized continuous processes (Nieto-Barajas *et al.*, 2004; Regazzini *et al.*, 2003). However, for the purpose of this thesis, we focus on nonparametric mixtures with the stick-breaking representation.

1.1.2 Markov Models

We consider two large families of Markov processes commonly used as statistical models: real valued diffusions, in continuous time; and nonparametric time series, in discrete time.

1.1.2.1 Real Valued Diffusion Processes

Diffusion processes have been widely studied in the context of probability theory and in many other areas, ranging from the natural sciences like biology or genetics, to the realms of economics and finance. One of the main features of this family of stochastic processes, the continuity of their paths, makes them attractive models for several phenomena.

There are different ways to define diffusions, and they are all closely related with Brownian motion, the predecessor and simplest of all real valued diffusions.

Definition 4 *A continuous time, real-valued stochastic process* $\{W_t : t \geq 0\}$ *is called a Brownian motion with drift parameter* μ*, diffusion parameter* σ^2*, and started at* $y \in \mathbb{R}$ *if the following conditions hold*

i) $W_0 = y$.

ii) The process has independent increments, i.e. for every $n \in \mathbb{N}$ *and* $0 \leq t_0 \leq \ldots \leq t_n < \infty$*, the increments* $W_{t_n} - W_{t_{n-1}}, W_{t_{n-1}} - W_{t_{n-2}}, \ldots, W_{t_1} - W_{t_0}$ *are independent random variables.*

iii) For every $s \geq 0$ *and* $t > 0$*, the increment* $W_{t+s} - W_s$ *is a normally distributed random variable with mean* μt *and variance* $\sigma^2 t$.

iv) The mapping $t \mapsto W_t$ *is almost surely continuous.*

If $\mu = 0$, $\sigma^2 = 1$ *and* $y = 0$ *the process is called standard Brownian motion.*

Even the fact that such a process exists is not trivial, due to the non countable nature of the product spaces involved and the continuity condition on the paths. Norbert Wiener provided the first proof of the existence of a process satisfying all conditions in Definition 4; for this reason, Brownian motion is also known as Wiener process. The proof found most commonly in the literature constructs a standard Brownian motion as the limit of a sequence of adequately scaled random walks. Such construction induces a probability measure $\mathbb{W}$ on the space $\mathcal{C}_{[0,\infty)}$ of continuous real valued functions on $[0,\infty)$, with Borel σ-algebra $\mathcal{B}(\mathcal{C}_{[0,\infty)})$, under the topology of weak convergence. Therefore $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}), \mathbb{W})$ is known as the canonical probability space for Brownian motion, and the probability $\mathbb{W}$ is called the Wiener measure.

The definition of standard Brownian motion, together with some properties of the model allow the construction of other Brownian motion processes. Specifically, if $W = \{W_t : t \geq 0\}$ is a standard Brownian motion on any probability space, then the process $Y = \{Y_t : t \geq 0\}$ defined by $Y_t = y + \sigma W_t + \mu t$ is a Brownian motion started at y, with drift and diffusion parameters μ and σ^2, respectively. In particular, we denote by $\mathbb{W}^y$ the measure induced by Y on $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}))$, when $\mu = 1$ and $\sigma^2 = 1$, giving place to the following definition.

Definition 5 *A Brownian family is a stochastic process $W = \{W_t : t \geq 0\}$ adapted to a filtration $\{\mathcal{A}_t : t \geq 0\}$ on a measurable space $(\Omega, \mathcal{A})$, and a family of probability measures $\{\mathbb{P}^y : y \in \mathbb{R}\}$ such that W is a Brownian motion started at y under the probability measure $\mathbb{P}^y$.*

In fact, it is enough to have $\mathbb{P}^y[W_0 = y] = 1$. This idea can be generalized, resulting in a wider definition of Brownian motion.

Definition 6 *An adapted process $W = \{W_t, \mathcal{A}_t : t \geq 0\}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, which is a Brownian motion under $\mathbb{P}$ and such that $\mathbb{P}[W_0 \in B] = P_0(B)$ for every $B \in \mathcal{B}(\mathbb{R})$, is called Brownian motion with initial distribution P_0.*

Clearly, when $P_0 = \delta_y$ for some $y \in \mathbb{R}$, we simply have $\mathbb{P} = \mathbb{P}^y$.

From the definition, we know that Brownian motion is a process with continuous paths and stationary independent increments. The converse is also true, any process with such properties is a Brownian motion.

It can be shown that a Weiner process is a time homogeneous (strong) Markov process. A Brownian motion on the canonical space is therefore completely defined by the distribution P_0 of the initial point W_0, and a family $\{f_t : t > 0\}$ of transition densities. In fact, an adapted process $Y = \{Y_t : t \geq 0\}$ on $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}), \{\mathcal{A}_t\}_{t\geq 0}, \mathbb{P})$ such that, for every $B \in \mathcal{B}(\mathbb{R})$

i) $\mathbb{P}[W_0 \in B] = P_0(B)$ and

ii) for every $t, s > 0$

$$\mathbb{P}[W_{t+s} \in B | W_s = y_0] = \mathbb{P}[W_t \in B | W_0 = y_0] = \int_B \mathrm{N}(y|\mu t + y_0, \sigma^2 t)\mathrm{d}y,$$

is a Brownian motion with drift coefficient μ, diffusion coefficient σ^2 and initial distribution P_0, where $f_t(\cdot) = \mathrm{N}(\cdot|\mu, \sigma^2)$ denotes the normal density function with mean μ and variance σ^2.

This means that the finite dimensional distributions of Brownian motion are all multivariate Gaussian distributions. A Brownian motion started at y, with drift coefficient μ and diffusion coefficient σ^2 is, therefore, a Gaussian process with mean function $\mu(t) = \mu t$ and covariance function $\sigma(s,t) = \sigma^2 \min\{s,t\}$.

Brownian motion has many important and interesting properties. We consider here only two of them, which are essential to the definition of diffusion processes in particular, and to the development of stochastic calculus in general. A Weiner process $W = \{W_t : t \geq 0\}$ has unbounded variation and finite quadratic variation. Formally, for any $T \in [0, \infty)$, consider a sequence $(\mathcal{T}_n)_{n\geq 1}$ of partitions of $[0, T]$, i.e.

$$\mathcal{T}_n = \{0 = t_0^n \leq \ldots \leq t_n^n = T\}.$$

Assume that, for every $n \geq 1$, $\mathcal{T}_n \subset \mathcal{T}_{n+1}$, and

$$\max_i \{t_i^n - t_{i-1}^n\} \to 0 \quad \text{as } n \to \infty.$$

Then

$$\lim_n \sum_{i=1}^n |B_{t_i} - B_{t_{i-1}}| = \infty \quad \text{a.s.}, \tag{1.22}$$

$$\lim_n \sum_{i=1}^n (B_{t_i} - B_{t_{i-1}})^2 = T \quad \text{in } L^2. \tag{1.23}$$

The unbounded variation (1.22) means that the Riemann-Lebesgue-Stieltjes theory of integration does not allow us to define an integral of the form

$$\int_0^t g(s)\mathrm{d}W_s. \tag{1.24}$$

However, the definition of this integral and more general ones, where the integrator is any stochastic process with finite quadratic variation, give rise to the field of stochastic calculus. They are known as Itô integrals and they are defined for integrands in a family of stochastic processes known as supermartingales, which includes, in particular, continuous functionals of Brownian motion.

The theory of Itô calculus, or stochastic calculus, is extensive and mathematically complex, so we do not discus it here. We merely present some of the results that are most important for the definition of diffusions and our use of them as statistical models. In particular, we focus on real-valued diffusions, so we only require the stochastic calculus version of some of the main results of univariate standard calculus. We begin by presenting Itô's formula, which substitutes the fundamental theorem of calculus.

Lemma 3 (Itô's formula) *Let $h : \mathbb{R} \to \mathbb{R}$ be a twice continuous differentiable function and let $W = \{W_t : t \geq 0\}$ be a Weiner process. Then, for all $t \geq 0$*

$$h(W_t) - h(W_0) = \int_0^t h'(W_s)\mathrm{d}W_s + \frac{1}{2}\int_0^t h''(W_s)\mathrm{d}s. \tag{1.25}$$

Rearranging the terms in the above expression, we may write the stochastic integral of $h' = \mathrm{d}h(y)/\mathrm{d}y$ with respect to Brownian motion, as

$$\int_0^t h'(W_s)\mathrm{d}W_s = h(W_t) - h(W_0) - \frac{1}{2}\int_0^t h''(W_s)\mathrm{d}s, \tag{1.26}$$

the familiar expression $h(W_t) - h(W_0)$, from the fundamental theorem of calculus, plus a compensation term, which is a Lebesgue-Stiltjes integral. This result is essential to the construction of the latent likelihood expression we present in the next section, to allow simulation and inference for diffusion models.

Another fundamental result is Girsanov's formula, which extends the principle of change of measure to Itô integration.

Theorem 3 (Girsanov-Cameron-Martin) *Let $W = \{W_t : t \geq 0\}$ be a standard Brownian motion defined on the canonical space $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}), \mathbb{W})$, with the natural filtration $\{\mathcal{A}_t\}_{t\geq 0}$. Let $Y = \{Y_t : t \geq 0\}$ be an adapted process on the same space, such that*

$$\mathbb{W}\left[\int_0^t Y_s^2 \mathrm{d}s < 0\right] = 1, \quad 0 \leq t < \infty. \tag{1.27}$$

Define, for each $t \geq 0$,

$$\tilde{W}_t = W_t - \int_0^t Y_s \,\mathrm{d}s, \tag{1.28}$$

and assume the stochastic process $Z = \{Z_t : t \geq 0\}$ given by

$$Z_t = \exp\left\{\int_0^t Y_s \,\mathrm{d}W_s - \frac{1}{2}\int_0^t Y_s^2 \,\mathrm{d}s\right\} \tag{1.29}$$

is a martingale.

Then, for every $t \in [0, \infty)$, the process $\{\tilde{W}_s : 0 \leq s \leq t\}$ is a Brownian motion on $(\mathcal{C}_{[0,\infty)}, \mathcal{A}_t, \mathbb{P}_t)$, adapted to $\{\mathcal{A}_t\}_{0\leq s\leq t}$, where $\mathbb{P}_t$ is defined as

$$\mathbb{P}_t(A) = \int_A Z_t(\omega)\mathrm{d}\mathbb{W}(\omega). \quad A \in \mathcal{A}_t. \tag{1.30}$$

The theorem can be expressed more generally, not only for the canonical space. However, this choice guarantees the existence of a probability measure $\mathbb{P}$ on $\mathcal{A}_\infty = \mathcal{B}(\mathcal{C}_{[0,\infty)})$ such that, restricted to $\mathcal{A}_t$ it coincides with $\mathbb{P}_t$. Therefore, the complete process $\tilde{W} = \{\tilde{W}_t : t \geq 0\}$ is a Brownian motion with respect to $\mathbb{P}$. Furthermore, the measures $\mathbb{P}$ and $\mathbb{W}$ are mutually absolutely continuous when restricted to $\mathcal{A}_t$, with Radon-Nykodim derivative given by $Z_t = \mathrm{d}\mathbb{P}_t/\mathrm{d}\mathbb{W}$. However, $\mathbb{P}$ and $\mathbb{W}$ are not, in general, absolutely continuous on the complete $\mathcal{B}(\mathcal{C}_{[0,\infty)})$.

Before the development of Itô calculus, diffusions were defined as continuous time Markov processes characterized by their infinitesimal generators. Consider a real valued time homogeneous Markov family $Y = \{Y_t : t \geq 0\}$, $\{\mathbb{P}^y : y \in \mathbb{R}\}$ on a filtered space $(\Omega, \mathcal{A}, \{\mathcal{A}_t\}_{t\geq 0})$. Denote by $\mathcal{C}_2$ the set of real valued, twice continuously differentiable functions on $\mathbb{R}$ and, for $h \in \mathcal{C}_2$, let

$$\mathbb{E}_y\big[h(Y_t)\big] = \int_\Omega h\big(Y_t(\omega)\big)\mathrm{d}\mathbb{P}^y(\omega). \tag{1.31}$$

The infinitesimal generator of Y is a linear operator $\mathcal{G}$ defined by

$$\mathcal{G}h(y) = \lim_{t \downarrow 0} \frac{\mathbb{E}_y\big[h(Y_t)\big] - h(y)}{t}, \quad \forall y \in \mathbb{R}, \tag{1.32}$$

for every integrable function h (see e.g. Ethier & Kurtz, 1986; Lamperti, 1977, for more on generators of Markov processes).

Let $\mathcal{D}$ be the second order differential operator associated with the drift coefficient $\alpha : \mathbb{R} \to \mathbb{R}$. and the diffusion coefficient $\sigma : \mathbb{R} \to (0, \infty]$, i.e.

$$\mathcal{D}h(y) := \frac{1}{2}\sigma^2(y)\frac{\mathrm{d}^2 h(y)}{\mathrm{d}y^2} + \alpha(y)\frac{\mathrm{d}h(y)}{\mathrm{d}y}. \tag{1.33}$$

Definition 7 *Let $Y = \{Y_t : t \geq 0\}$, $\{\mathbb{P}^y : y \in \mathbb{R}\}$, $(\Omega, \mathcal{A}, \{\mathcal{A}_t\}_{t \geq 0})$ be a real valued time homogeneous Markov family. Then Y is called a diffusion process if the following conditions hold*

i) Y has (a.s.) continuous sample paths.

ii) For every bounded $h \in \mathcal{C}_2$, with bounded continuous first and second order derivatives,

$$\mathcal{G}h = \mathcal{D}h. \tag{1.34}$$

iii) For every $y \in \mathbb{R}$

$$\mathbb{E}_y\big[Y_t - y\big] = t\alpha(y) + o(t); \tag{1.35}$$

$$\mathbb{E}_y\big[(Y_t - y)^2\big] = t\sigma^2(y) + o(t). \tag{1.36}$$

The drift coefficient α can be interpreted as the instantaneous expected rate of change of the process, while σ^2 represents the instantaneous rate of change of the process variance. Diffusion processes are therefore defined in terms of infinitesimal characteristics, which cannot be captured by discretization. This constitutes one of the problems for inference on discretely observed diffusion models, which reflects in the intractability of the resulting likelihood functions, as we explain at the end of this section.

The construction of diffusion processes, from Definition 7 follows an analytical approach. Assuming that the transition densities

$$f_t(y|y_0)\mathrm{d}y = \mathbb{P}^{y_0}[Y_t \in \mathrm{d}y] \quad \text{for each } y, y_0 \in \mathbb{R} \text{ and } t > 0, \tag{1.37}$$

for the Markov family exist, they must satisfy the forward and backward Kolmogorov equations. It follows from condition (1.34) that, fixing $y_0 \in \mathbb{R}$, the forward equation is given by

$$\frac{\partial f_t(y|y_0)}{\partial t} = \mathcal{G}^* f_t(y|y_0) = \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y) f_t(y|y_0)\right] - \frac{\partial}{\partial y}\left[\alpha(y) f_t(y|y_0)\right]; \tag{1.38}$$

while the backward equation is, for a fixed $y \in \mathbb{R}$

$$\frac{\partial f_t(y|y_0)}{\partial t} = \mathcal{G} f_t(y|y_0) = \frac{1}{2}\sigma^2(y)\frac{\partial^2}{\partial y_0^2} f_t(y|y_0) - \alpha(y)\frac{\partial}{\partial y_0} f_t(y|y_0). \tag{1.39}$$

A diffusion process can then be defined by finding a solution to the above, known as Fokker-Planck-Kolmogorov equations, since the family of transition densities characterizes the finite-dimensional distributions of a Markov process with fixed initial point.

When diffusion processes are used as statistical models, it is assumed that a sample $y_{1:n} = (y_1, \ldots y_n)$ is a partially observed realization of a diffusion path. Formally, we consider a diffusion process $Y = \{Y_t : t \geq 0\}$, started at a known, fixed point $Y_0 = y_0 \in \mathbb{R}$, and assume each y_i is a realization of Y_{t_i}, for times $t_1 < t_2 < \ldots < t_n$. This is commonly referred to as a discretely observed diffusion.

The likelihood for the sample is the density associated to the finite dimensional distribution of $(Y_{t_1}, \ldots, Y_{t_n})$ under the probability measure $\mathbb{P}^{y_0}$, evaluated at $(y_1, \ldots y_n)$. We assume the drift and diffusion coefficients characterizing the diffusion process satisfy all necessary conditions for the existence of the transition densities $\{f_t : t > 0\}$. The likelihood function, then, takes the form

$$f(y_{1:n}) = \prod_{i=1}^{n} f_{\Delta_i}(y_i|y_{i-1}), \tag{1.40}$$

where $\Delta_i = t_i - t_{i-1}$.

Provided sufficient smoothness conditions on α and σ, the existence of a solution to the Fokker-Planck-Kolmogorov can be guaranteed. However, in all but a few cases, such solutions do not have an analytic form, and therefore, the discretely observed diffusion model has an intractable likelihood function.

The advent of Itô calculus provided a new way to define diffusion processes, as solutions to stochastic differential equations (SDEs), and the constraints on the drift and diffusion coefficient for such solutions to be well defined are milder than those required for the existence of the transition densities. Thus, the use of stochastic calculus transformed diffusion processes into an extensive and rich family of processes. Although we are only interested in a subset of it, for which densities exist, we rely on the construction of diffusions as weak solutions to stochastic differential equations (SDEs), through the use of Girsanov's formula, in order to deal with the intractability of the model. In this section, we discus the construction of diffusion processes only. In the next section, we present a latent variable extension, based in this construction, which Beskos *et al.* (2006a) developed to enable the simulation of diffusion paths, as well as inference for discretely observed diffusion models.

Consider a Brownian family $Y = \{Y_t : t \geq \infty\}$, $\{\mathbb{W}^y\}_{y\in\mathbb{R}}$, defined on the canonical space $\left(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)})\right)$, with natural filtration, and a continuous, thus Borel-measurable function, $\alpha : \mathbb{R} \to \mathbb{R}$, such that

$$|\alpha(y)| \leq C(1 + |y|), \quad \forall\, y \in \mathbb{R}, \tag{1.41}$$

for some $C > 0$. In this case, it can be shown that

$$Z_t = \exp\left\{\int_0^t \alpha(Y_s)\,\mathrm{d}Y_s - \frac{1}{2}\int_0^t \alpha^2(Y_s)\,\mathrm{d}s\right\} \tag{1.42}$$

is a martingale under each $\mathbb{W}^y$. Applying Girsanov's theorem, the process

$$W_t = Y_t - Y_0 - \int_0^t \alpha(Y_s)\mathrm{d}s \tag{1.43}$$

is a Brownian motion started at $W_0 = 0$, under the measure $\mathbb{P}^y$ defined by $\mathrm{d}\mathbb{P}^y/\mathrm{d}\mathbb{W}^y = Z_t$ on $\mathcal{A}_t$. Rearranging terms gives

$$Y_t = Y_0 + \int_0^t \alpha(Y_s)\mathrm{d}s - W_t, \tag{1.44}$$

which in differential notation, corresponds to the SDE

$$\mathrm{d}Y_t = \alpha(Y_t)\mathrm{d}t - \mathrm{d}W_t; \quad Y_0 = y. \tag{1.45}$$

The definition of the Brownian motion on the canonical space, guarantees the existence of the unique measure $\mathbb{P}^y$, for each y, so that the complete process $Y = \{Y_t : t \geq 0\}$ is defined on $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}), \mathbb{P}^y)$. It can be shown that Y satisfies the conditions of Definition 7, making it a diffusion process with drift coefficient α and diffusion coefficient 1.

This idea may be generalized to define a diffusion process with general diffusion coefficient σ, as a weak solution to an SDE

$$\mathrm{d}Y_t = \alpha(Y_t)\mathrm{d}t - \sigma(Y_t)\mathrm{d}W_t. \tag{1.46}$$

The construction of diffusions as solutions to stochastic differential equations was suggested by P. Lévy and developed by K. Itô. The SDE (1.46) is said to have a *strong solution* when a process Y can be constructed, which satisfies the equation with respect to a given filtration and a given Brownian Motion. A *weak solution* exists when the probability space, the filtration and the driving Brownian motion are part of the solution and not part of the statement of the problem. For more detail on this type of constructive definition of diffusion processes and the difference between strong and weak solutions see e.g. Karatzas & Shreve (1991); Revuz & Yor (1999). Throughout the present work, we are only interested in weak solutions, since they are sufficient for a diffusion process to be used as a statistical model.

For simplicity, we limit our analysis to discretely observed diffusion models for which the diffusion coefficient is constant $\sigma = 1$ and the drift coefficient belongs to a parametric family $\{\alpha_\theta : \theta \in \Theta\}$. Thus, the likelihood of the model takes the shape

$$f(y_{1:n}|\theta) = \prod_{i=1}^{n} f_{\Delta_i}(y_i|y_{i-1}, \theta). \tag{1.47}$$

The Bayesian model is completed by a prior Π on the parameter space Θ, which induces a prior on the space $\mathcal{F}$ of, possibly intractable, transition densities implicitly defined by the choice of diffusion and drift coefficients.

1.1.2.2 Nonparametric Time Series Models

In the most general sense, time series are countable sequences of random variables recorded over time. This defines an order which is assumed to be relevant, in other words, even though i.i.d. or exchangeable sequences of data, formally constitute time series, the term is used only for stochastic processes in which the order matters. The law of the process is thus, generally described in terms of conditional distributions of the ordered sequence. Therefore, given a sample $y_{0:n} = (y_0, \ldots, y_n)$,the likelihood for a time series model has the form

$$f(y_{0:n}) = f_0(y_0, \ldots y_{m-1}) \prod_{i=m}^{n} f_i(y_i | y_{i-1}, \ldots, y_{i-m}). \tag{1.48}$$

In other words, in a time series model, the observations are assumed to be realizations from a discrete time stochastic process with order $m \leq n$ Markov dependence. If the process is time homogeneous, the densities are time invariant, so we may write

$$f(y_{0:n}) = f(y_0, \ldots y_{m-1}) \prod_{i=m}^{n} f(y_i | y_{i-1}, \ldots, y_{i-m}). \tag{1.49}$$

A general practice is to assume the first observations, $y_{0:m-1}$ are fixed, so that $f_0(y_{0:m-1}) = 1_{\{y_{0:m-1}\}}$; or that such density is fully known. In either case, the $f_0(y_{0:m-1})$ may be removed from the likelihood expression. An alternative, is to assume the process is stationary, and therefore fully specified by the conditional density $f(y_i | y_{i-1}, \ldots, y_{i-m})$. In this case, $f_0(y_{0:m-1})$ may be included in the likelihood expression, but model specification is only required for the conditional densities, since the invariant density is uniquely defined by them.

A case which deserves special attention occurs when $m = 1$, and the time series model is simply a discrete time Markov model. The simplest and most commonly used is the normal linear autoregressive model, AR(1), where each observation y_i is assumed to be a realization of a random variable Y_i with a dependence structure given by

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathrm{N}(\cdot | 0, \omega^2), \tag{1.50}$$

for some $\beta_0, \beta_1 \in \mathbb{R}$ and $\omega > 0$. When $|\beta_1| < 1$, the process is stationary. Clearly, both the transition structure and the invariant distribution for this type of model are very limited, but many more flexible models have arisen as generalizations of this simple idea. In particular, the nonparametric model we propose in Chapter 3 originates on the simple AR(1) model.

In the context of statistical inference, time homogeneous order m Markov processes are commonly known as autoregressive models, since they can be constructed as regression models, in which the covariates for each observation y_i are the previous m observations $y_{i-m}, \ldots y_{i-1}$. From this perspective, autoregressive models are defined through the conditional densities

$$f(y_i|y_{i-1}, \ldots, y_{i-m}) \tag{1.51}$$

and they are as flexible as the regression form chosen for such conditional densities. In the AR(1) case

$$f(y_i|y_{i-1}, \beta, \omega) = \mathrm{N}(y_i|\beta_0 + \beta_1 y_{i-1}, \omega^2). \tag{1.52}$$

More general forms of regression lead to more general transition densities and therefore to more flexible dependence structures of the stochastic process they define. Therefore, one approach for the definition of nonparametric process is to construct nonparametric conditional densities to define the transition mechanism. In this sense, any one of the flexible regression models presented in the next section can be used to define a time series model. However, when flexible transition densities are defined, desirable properties of the resulting process, such as stationarity are difficult to verify and so this type of constructions focus on transition density estimation and prediction, with no regard for the stationarity of the process.

Examples of this type of construction are given by Müller *et al.* (1997), who define the transition density of an autoregressive model as a semiparametric finite mixture of AR(1) models, assigning a Drichlet Process prior on the model parameters. Tang & Ghosal (2007a) define a mixture of Dirichlet process model for the transition density, with a Gaussian kernel, and the correlation structure introduced through a hyperbolic tangent link function as the mean for the parametric kernels.

The construction of flexible transition structures is only as complicated as the definition of flexible conditional densities in regression models. It is additional requirements, such as stationarity that make the problem more challenging. Until recently the construction of parametric stationary time series with non-normal invariant distributions was a difficult enough task. Pitt *et al.* (2002) and Pitt & Walker (2005) provide a latent variable construction for stationary processes with general parametric invariant densities. Their idea is to define a joint density for an observation y and a latent variable s, as $f_\theta(y,s) = f_\theta(y|s)f_\theta(s)$ from which the conditional $f_\theta(s|y)$ can be obtained via Bayes theorem. A stochastic process is then defined via a Gibbs sampling scheme in which the observation y_i is updated from the conditional $f_\theta(y_i|s_i)$, while the latent variable is updated from the conditional $f_\theta(s_i|y_{i-1})$. The process has a stationary density given by the marginal of $f_\theta(y,s)$, namely

$$f_\theta(y) = \int f_\theta(y|s)f_\theta(s)\mathrm{d}\nu(s), \tag{1.53}$$

and the transition density is simply the conditional distribution

$$f_\theta(y_i|y_{i-1}) = \frac{f_\theta(y_i,y_{i-1})}{f_\theta(y_i)} = \frac{\int f_\theta(y_i|s)f_\theta(s|y_{i-1})\,\mathrm{d}\nu(s)}{\int f_\theta(y|s)f_\theta(s)\,\mathrm{d}\nu(s)}. \tag{1.54}$$

For a large family of parametric densities, an adequate latent variable may be chosen to ensure the desired stationary density. The transition may not have a closed form, but the Gibbs sampling construction ensures the models are naturally suited for MCMC inference.

More flexible models are needed to accommodate the complex dynamics observed in real life data. Mena & Walker (2005) generalize the above construction by considering the latent structure to be a random probability measure. In this case, the joint density is expressed as $f(y,\mathrm{d}P) = f_P(y)\Pi(P)$. Once again, a stationary Markov process is defined through this structure, with stationary and transition densities given by

$$f(y) = \int f_P(y)\mathrm{d}\Pi(P); \tag{1.55}$$

$$f(y_i|y_{i-1}) = \frac{f(y_i,y_{i-1})}{f(y_{i-1})} = \frac{1}{f(y_{i-1})}\int f_P(y)\mathrm{d}\Pi(P|y_{i-1}). \tag{1.56}$$

The integrals, in this case, constitute nonparametric mixtures, defined in terms of a nonparametric prior Π and the corresponding posterior $\Pi(\cdot|y)$ given an observation. The expressions are intractable and inference methods are only available for particular cases of the general construction. In order to enable inference through MCMC methods, Mena & Walker (2005) consider a joint density $f(y_i, y_{i-1})$ directly defined as

$$f(y_i, y_{i-1}) = \int f_P(y_i) f_P(y_{i-1}) \mathrm{d}\Pi(P). \tag{1.57}$$

In other words, conditional on the random probability P, the observations are independent and identically distributed, and their dependence is induced only through the correlation structure of the Gibbs construction for the distribution P. over time. Π is a Bayesian nonparametric prior and, specifically, Mena & Walker (2005) base it on the Gaussian process prior of Leonard (1978) and Lenk (1991). This model results in a transition density which is the predictive density function given a single observation from the Bayesian model, i.e.

$$f(y_i|y_{i-1}) = \int f_P(y_i) \mathrm{d}\Pi(P|y_{i-1}). \tag{1.58}$$

This can be nonparametric since the $\Pi(\cdot|y)$ is a probability measure that can accommodate two functions; one being the mean density $f(y)$ and another to do with the variance process $V(y)$, based on $\int P^2(y)\mathrm{d}\Pi(P)$. Then $f(y_i|y_{i-1})$ is a function of $(f(y_{i-1}), V(y_{i-1}))$. On the other hand, the stationary density is given by the parametric mean density of the process,

$$f(y) = \int f_P(y) \mathrm{d}\Pi(P). \tag{1.59}$$

While stationarity is a desirable property which facilitates estimation of relevant quantities, it is difficult to construct stationary models for which both the transition mechanism and the invariant density are sufficiently flexible. Therefore, attempts at defining flexible models often result in a compromise between flexibility and statistical properties.

In order to overcome the issue of the lack of flexibility of the stationary density, Martínez-Ovando & Walker (2011) propose a transition density defined as a nonparametric mixture, extending the Gibbs sampler model construction

from Pitt *et al.* (2002) and Mena & Walker (2005) described above. By adding a hierarchical level to the latent structure, they construct a process for which both the transition and the stationary densities are defined as nonparametric mixtures. The price to pay for the added complexity is a lack of interpretability which obscures the effects of the prior choices; and a model complexity which requires careful choices of the mixing components and probabilities to ensure feasibility of the MCMC inference procedures.

Other examples of nonparametric time series models define the transition density as a mixture of parametric conditional densities, i.e.

$$f(y_i|y_{i-1}) = \int_{\Theta} K_\theta(y_i|y_{i-1}) \mathrm{d}P(\theta|y_{i-1}), \tag{1.60}$$

for some conditional density function or kernel K_θ. In general, this type of models need not be stationary. Furthermore, constraints must be imposed on the structure of the dependent mixing measures $P_y(\cdot) = P(\cdot|y)$ and the corresponding priors, in order to ensure inference is feasible. We discuss this further in the following section and in Chapter 4, in the context of regression models. Suffice to say that the constructions proposed by Müller *et al.* (1996) and Martínez-Ovando & Walker (2011) provide evidence of the difficulty in defining nonparametric transitions for which nonparametric stationary distributions exist.

The idea of using latent structures to induce time dependence has been widely explored, even to an extreme in which the latent structure is itself the object of interest in the estimation procedure. The area of hidden Markov models has a place of its own in Bayesian literature and many models have been proposed (see e.g. Cappé *et al.*, 2005). Nonparametric extensions (Van Gael *et al.*, 2008) allow for great flexibility in the transitions, but emphasis is placed on inference for the latent structure, so inference for the transition density for the observations may not be possible.

Finally, in many models, flexibility is achieved by forcing non stationarity and non homogeneity of the transition mechanism over time. It is common in this case to define transition densities through dependent mixture models in the manner of nonparametric regression models, incorporating time as a covariate. Some models of this type can be found in Griffin & Steel (2006, 2011); Zhu *et al.* (2005) and Williamson *et al.* (2010); more can be obtained from the nonparametric regression

models described in the following section, by incorporating time and/or previous observations as covariates for the transition density of a stochastic process.

There is an extensive literature regarding the definition of time series models and the methods used for statistical inference, both in the classical and the Bayesian settings. We do not cover all of it here, as we are only interested in the more flexible nonparametric ideas. In Chapter 3 we propose a time homogeneous autoregressive stationary model with fully nonparametric transition and invariant densities, which can be generalized to obtain higher order Markov dependence as well as a dependence structure changing over time.

1.1.3 Nonparametric Regression Models

The contents of this section constitute the introduction of Antoniano-Villalobos *et al.* (2012).

The standard linear regression model assumes a response variable $y \in \mathbb{Y}$ is related to some covariate $x \in \mathbb{X}$ through a linear function with additive normal errors, that is

$$y = \beta X + c; \quad c \sim \mathrm{N}(c|0, \sigma^2),$$

where, for a p-dimensional covariate x, β is a $(p+1)$-dimensional vector of constant coefficients, and we define $X = (1, x)$.

This is, however, a limited and unrealistic model in most applications. Real life data exhibit a more complicated relation between covariates and response variables, so there is a need to construct models that allow for a more flexible dependence structure. One of the most popular approaches consist in representing the regression function as a linear combination of basis functions, such as splines or wavelets (Denison *et al.*, 2002; Dimatteo *et al.*, 2001). Another common practice, when more flexibility is desired, is to place a Gaussian Process prior on the unknown regression function (Rasmussen & Williams, 2006), thus defining a nonparametric model.

These models achieve flexibility for the mean function, however, they are still limited, in the sense that they only allow for a basic structure of the errors. Many data sets present departures from classical distributional assumptions, such

as normality or the uni-modality of error distributions. It is common to observe non standard variances, skewness and unconventional tail behaviour in different regions of the covariate space, $\mathbb{X}$. To capture such behaviour, nonparametric approaches for modelling the conditional density $f(y|x)$ in its entirety, are becoming increasingly popular.

As stated in Section 1.1.1.2, a flexible model for independent and identically distributed observations can be defined as an infinite mixture of parametric models, given by

$$f_P(y) = \sum_{j=1}^{\infty} w_j K(y|\theta_j), \tag{1.61}$$

where $K(\cdot|\theta)$ is a parametric family of density functions defined on $\mathbb{Y}$ and P is an almost surely discrete random probability measure on the parameter space Θ, characterized by some atoms $\theta_j \in \Theta$, and weights $w_j \geq 0$, such that $\sum_j w_j = 1$ (a.s.).

For covariate dependent density estimation, the mixture model can be adapted by allowing the mixing distribution P to depend on the covariate value x, and replacing the parametric kernel $K(y|\theta)$ with some parametric regression model $K(y|x,\theta)$, such as a linear regression model. Hence, for every $x \in \mathbb{X}$,

$$f_{P_x}(y|x) = \int K\big(y|x,\theta(x)\big)\mathrm{d}P_x\big(\theta(x)\big). \tag{1.62}$$

As in the i.i.d. case, the Bayesian model is completed by assigning a prior distribution on the family $\{P_x\}_{x\in\mathbb{X}}$ of covariate dependent mixing probability measures. If, for every x, the prior gives probability one to the set of discrete probability measures, then each mixing distribution admits a representation defined by a weighted sum of atom masses,

$$P_x = \sum_{j=1}^{\infty} w_j(x)\delta_{\theta_j(x)},$$

and

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x)K\big(y|x,\theta_j(x)\big), \tag{1.63}$$

where $\theta_j(x) \in \Theta$, and the weights $w_j(x) \geq 0$ are such that $\sum_j w_j(x) = 1$ (a.s.) for all $x \in \mathbb{X}$.

A first proposal along the lines of model (1.62) was given by Cifarelli & Regazzini (1978), with a focus on discrete covariates. They induce dependence between a finite number of random probability measures, through the base measure of a Dirichlet Process. Their proposal extends Antoniak's (1974) mixture of Dirichlet Processes, by defining, for some finite N and $\mathbb{X} = \{1, \ldots, N\}$,

$$P_1, \ldots, P_N | u(1), \ldots, u(N) \sim \prod_{x=1}^{N} \mathrm{DP}\Big(\zeta(x) P_0(\cdot | u(x))\Big),$$

where ζ is a function on $\mathbb{X}$ taking values in $(0, \infty)$, and for some distribution Π, $u(1), \ldots, u(N) \overset{iid}{\sim} \Pi$.

In terms of equation (1.63), this implies that the weights are allowed to vary with x, but are constructed independently across x. Thus, dependence of the conditional densities for different covariate values is induced through the covariate dependent atoms, given by

$$\theta_j(x) | u(x) \overset{ind}{\sim} P_0\big(\cdot | u(x)\big). \tag{1.64}$$

To allow continuous covariates, Muliere & Petrone (1993) extend this idea by assuming $u(x) = (\beta, \sigma^2)\ \forall\ x \in \mathbb{X}$ and

$$\theta_j(x) | u(x) \overset{ind}{\sim} \mathrm{N}(\cdot | X\beta, \sigma^2), \tag{1.65}$$

where $X = (1, x)$. The limitation of this construction is the restrictive nature of the induced dependence.

The general model (1.63) is introduced by MacEachern (1999; 2000), assuming a Dirichlet Process prior as the marginal distribution of P_x. This choice is justified by the connection of the DP with finite mixture models, its simple prior elicitation and large support, as well as the availability of computational procedures for inference.

MacEachern's general class of models is now known as Dependent Dirichlet Processes (DDP). The basic assumption underlying their construction is that

each $\{w_j(x)\}_{x\in\mathbb{X}}$ is a stochastic processes, with a correlation across j given by the stick breaking construction,

$$w_1(x) = v_1(x). \quad \text{and} \quad w_j(x) = v_j(x)\prod_{j'<j}(1 - v_{j'}(x)),\ j > 1. \tag{1.66}$$

where the $\{v_j(x)\}_{x\in\mathbb{X}}$ are independent processes such that, marginally

$$v_j(x) \overset{ind}{\sim} \text{Be}(1, \zeta(x)) \text{ for } j = 1, 2, \ldots,$$

for some function $\zeta : \mathbb{X} \to (0, \infty)$. Moreover, the $\{\theta_j(x)\}_{x\in\mathbb{X}}$ are independent stochastic processes with marginal distribution P_{0x}, and independent of the $v_j(x)$.

A popular version of the general model, the single-weight DDP is obtained when $w_j(x) = w_j$ for all $x \in \mathbb{X}$. Its attractiveness results from the fact that inference can be carried out using any of the well established algorithms for DPM models mentioned in Section 1.2.2.2. Single-weight DDP mixtures have been successfully applied to address a wide range of problems, from classical regression (MacEachern, 2000, 2001) to ANOVA (De Iorio *et al.*, 2004), spatial modelling (Gelfand *et al.*, 2005), time series analysis (Rodriguez & ter Horst, 2008), discriminant analysis (Cruz-Mesía *et al.*, 2007), longitudinal analysis (Müller *et al.*, 2005), and survival analysis (De Iorio *et al.*, 2009; Jara *et al.*, 2010).

Recent developments explore the use of covariate dependent weights. To simplify computations and ease interpretation, atoms are usually assumed not to depend on the covariates, and are therefore referred to as single-particle DDPs. It can be argued that both the single-weight and the single-particle versions of the model have a large enough support to describe the variability found in real data (see Barrientos *et al.*, 2012). The general model with covariate dependent weights and atoms, on the other hand, is usually considered too flexible for effective estimation.

The main constraint for the construction of DDP models with covariate dependent weights, is the need to specify a prior such that $\sum_j w_j(x) = 1$ a.s. for all $x \in \mathbb{X}$, which is non trivial for an infinite number of positive weights. The stick-breaking representation (1.66) propsed by MacEachern is justified by the need to satisfy this constraint. A wide variety of models present in the literature follow this structure and differ only in the construction of the $v_j(x)$.

One of the first approaches to covariate dependent weight mixture models, developed by Griffin & Steel (2006), incorporates dependency in the weights by reordering i.i.d. beta random variables, $\{v_j\}$, according to some concept of distance in the covariate space. They successfully apply this idea to stochastic volatility and spatial modelling; but do not discuss how to handle discrete covariates.

Dunson & Park (2008) introduce the kernel stick-breaking approach for the construction of covariate dependent weights, where $v_j(x) = v_j K(x|\psi_j)$ for some kernel function on $\mathbb{X}$, with parameter ψ_j. They use this idea in and epidemiological study; Reich & Fuentes (2007) apply it to a spatial data-set concerning hurricane wind fields. Both examples involve continuous covariates only; to incorporate discrete covariates, adequate kernels must be specified.

Another common model defines $v_j(x) = \ell(g(x;\psi_j))$, where $\ell : \mathbb{R} \to [0,1]$ is a monotone, differentiable link function and g is a real-valued function on $\mathbb{X}$. Common choices for g, are simple linear functions, linear combinations of basis functions, and Gaussian Processes (see e.g. Chung & Dunson, 2009; Dunson & Rodríguez, 2011; Ren *et al.*, 2011). Applications of this approach include stochastic volatility models and image segmentation. Alternative options for g must be explored if discrete covariates are present.

Other proposals focus exclusively on discrete covariates (see for example, Müller *et al.*, 2004; Rodriguez *et al.*, 2008; Teh *et al.*, 2006).

An interesting idea that has received recent attention in the literature is to model the joint distribution of y and x through a nonparametric mixture of density functions on $\mathbb{Y} \times \mathbb{X}$. Inference is carried out for the joint density, via the usual methods for nonparametric mixture models. Conditional density estimates are then obtained from the posterior inference based on the joint model. However, as stated by Müller & Quintana (2004), this approach "wrongly introduces an additional factor for the marginal of x in the likelihood and thus provides only approximate inference". In fact, including this additional factor, forces a fit of the marginal distribution of x, thus degrading the performance of the conditional density estimate. This approach was first introduced by Müller *et al.* (1996), and subsequently studied and employed by Hannah *et al.* (2011); Kang & Ghosal (2009); Park & Dunson (2010); Shahbaba & Neal (2009) and Müller & Quintana (2010).

1.2 MCMC Methods and Latent Variables

Most of the existing literature regarding intractable components is concerned with the approximation of intractable normalizing constants when both the observations and the parameters are finite-dimensional.

Consider a Bayesian parametric model

$$Y_i \overset{iid}{\sim} f(\cdot|\theta)$$
$$\theta \sim \Pi.$$

The posterior density, given a sample $y_{1:n} = (y_1, \ldots, y_n)$ is

$$\Pi^n(\theta) = \frac{\Pi(\theta)\prod_{i=1}^n f(y_i|\theta)}{\int_\Theta \Pi(\theta)\prod_{i=1}^n f(y_i|\theta)\mathrm{d}\nu(\theta)}. \tag{1.67}$$

The prior Π is sometimes chosen to be conjugate with the likelihood, to guarantee the expression above has a closed form. However, in general, the posterior density can be known only up to proportionality and the integral in the denominator constitutes an intractable normalizing constant.

Three mainstream approaches can be identified to deal with this type of problem: analytic approximation, usually based on Laplace transforms and other mathematical representations (see e.g. DiCiccio *et al.*, 1997); numerical integration or some form of adaptive quadrature method based on classical analysis techniques (see e.g. Evans & Swartz, 1995); and Monte Carlo simulation methods, which use samples drawn from Π^n to estimate relevant features of the distribution. While analytic approximation and numerical integration may be convenient for some distributions, they are not always available and their statistical properties are difficult to establish. For these and other reasons, Monte Carlo simulation is considered a more adequate approach for statistical analysis, specially in the Bayesian context (for more on the advantages of the Monte Carlo approach see e.g. Smith & Roberts, 1993).

Since drawing samples from complex and often high dimensional distributions directly may not be possible, estimation is usually achieved through a family of methods commonly known by the acronym MCMC, which stands for Markov Chain Monte Carlo.

The idea behind MCMC methods is the following. Suppose we wish to estimate some quantity associated to some distribution, in this case, Π^n. Then a stationary Markov chain is constructed with equilibrium density equal to Π^n. Independently of the initial point of the chain, if enough time is allowed, convergence should occur, so that the simulated values can eventually be regarded as a sample from the desired marginal distribution. Such sample can then be used to generate Monte Carlo estimates of the quantities of interest, relying on asymptotic properties of the Markov process.

One such asymptotic result tells us that the sequence of random variables $\{\theta_i\}_{i\geq 0}$ which constitute the Markov chain, converges in distribution to a random variable θ distributed according to Π^n. Formally,

$$\theta_i \xrightarrow[i\to\infty]{\text{D}} \theta \sim \Pi^n.$$

This allows, with some caution, the use of the Markov chain realizations as a sample from the desired distribution. Of course, successive realizations are correlated, therefore an adequate spacing may be allowed between consecutive sample elements to generate an approximately i.i.d. sample from Π^n. Alternatively, independent runs of the Markov chain may be used to generate a sample of the desired size.

The second commonly applied result in the context of MCMC states that, for any integrable function h,

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta_i) \xrightarrow[N\to\infty]{\text{a.s.}} \mathbb{E}[h(\theta)],$$

where the integrability, the almost sure convergence and the expectation are all with respect to the invariant measure of the Markov chain, in this case Π^n. This allows the use of consecutive realizations from a single run of the Markov chain to calculate an ergodic average of the function of interest, obtaining an unbiased estimate for the expectation.

There is by now an extensive literature on theoretical results concerning the convergence of MCMC methods, as well as output analysis and convergence diagnostics to ensure that, in practice, the Monte Carlo error of the estimates can be

considered small or negligible. Some useful references on this subject are Besag & Green (1993); Robert & Casella (2004); Smith & Roberts (1993); Tierney (1994).

However, before studying the asymptotic properties of an MCMC scheme, a Markov Chain with the desired stationary distribution must be constructed. There are two basic standard methods commonly used for this, known as the Gibbs sampler (GS) and Metropolis-Hastings (MH) algorithms.

For a p-dimensional parameter $\theta = (\theta_1, \ldots, \theta_p)$, the Gibbs sampler involves successive sampling from the full conditional distributions $\Pi^n(\theta_j|\theta_{-j})$, where θ_{-j} stands for the vector θ from which the component j has been removed. A more elaborate version of the algorithm allows for blocks of variables to be sampled simultaneously, from the conditional distributions given the variables not included in the block. The choice of blocks should minimize the correlation structure of successive states of the Markov chain. Thus, bigger blocks improve the speed of convergence of the chain to the equilibrium distribution, at the price of sampling from multivariate distributions. In practice, a compromise solution must be found between efficiency and the possibility to sample directly from high-dimensional multivariate conditional distributions. Even in the simplest form of the algorithm, for many models, sampling directly from the full conditional distributions is not possible or the correlation structure of the resulting Markov chain is so strong as to make the convergence of the MCMC scheme unfeasible. Furthermore, a sampling scheme of this type can only be applied for finite-dimensional parameters.

The Metropolis-Hastings algorithm avoids the problem of simulating exactly from a full conditional distribution. Each realization of the Markov chain is updated by generating an observation θ' from a proposal distribution q, which may depend on the current state θ. The proposal is then accepted as the new state for the chain with a probability which, in order to ensure the resulting Markov chain has the desired limiting distribution, is calculated as

$$\min\left\{1, \frac{\Pi^n(\theta')q(\theta|\theta')}{\Pi^n(\theta)q(\theta'|\theta)}\right\}. \tag{1.68}$$

Calculation of the intractable normalizing constant for the posterior distribution, that is, the marginal distribution for the data, is therefore not required. Since the acceptance probability depends on the choice of q, the performance of the algorithm can be improved by a suitable choice of the proposal distribution.

However, in many more complex statistical models, the intractability is originated by the normalizing constant of the likelihood function. We discuss this problem in the following Section.

1.2.1 MCMC for Doubly-Intractable Distributions

Assume the sample consists of i.i.d. observations, each of them distributed according to a parametric density which is known up to proportionality. In other words,

$$f(y|\theta) = \frac{g(y,\theta)}{Z(\theta)}, \tag{1.69}$$

where

$$Z(\theta) = \int_{\mathbb{Y}} g(y,\theta)\mathrm{d}\nu(y) \tag{1.70}$$

is intractable.

For simplicity, we consider posterior inference for a single observation, i.e. $n = 1$. Given a prior Π on the parameter space Θ, the posterior density for the parameter is given by

$$\Pi^n(\theta) = \frac{\Pi(\theta)f(y|\theta)}{f(y)} = \frac{\Pi(\theta)g(y,\theta)/Z(\theta)}{\int_\Theta \left[\Pi(\theta)g(y,\theta)/Z(\theta)\right]\mathrm{d}\nu(\theta)}. \tag{1.71}$$

In this expression, the integral in the denominator is an intractable normalizing constant. Furthermore, the $g(\theta)$ appearing in the numerator is the intractable normalizing constant for the conditional distribution of the data given the parameter. Therefore, in the literature, expressions of this type are sometimes called doubly-intractable distributions.

Traditional methods used for inference in the presence of single intractability are not applicable here. A Gibbs sampler can only be used when the target distribution is known at least up to proportionality. On the other hand, implementation of a Metropolis-Hastings update scheme would require the evaluation of the ratio

$$\frac{\Pi^n(\theta')q(\theta|\theta')}{\Pi^n(\theta)q(\theta'|\theta)} = \frac{\Pi(\theta')g(y,\theta')q(\theta|\theta')}{\Pi(\theta)g(y,\theta)q(\theta'|\theta)}\frac{Z(\theta)}{Z(\theta')}, \tag{1.72}$$

in order to calculate the acceptance probability. The expression, therefore depends on the ratio of intractable normalizing constants, which is not available.

Once again, a variety of methods have been proposed to deal with this problem and they can be divided into two main groups. The traditional approach consists of approximating the unknown ratio through the use of pseudo-likelihoods, importance sampling techniques or more elaborate approximate sampling schemes, such as bridge sampling or path sampling. See e.g. Gelman & Meng (1998) for a discussion of this type of methods and Andrieu & Roberts (2009) for more modern versions of the proposed algorithms.

Unfortunately, the use of approximate ratios leads to a Markov chain with a stationary distribution which only approximates the desired posterior Π^n. This may lead to problems in the Monte Carlo estimation, as shown by Murray & Ghahramani (2004) in the context of undirected graphical models. Approximation issues are more evident for high-dimensional parameters, for which accurate estimation of the intractable ratio is more challenging. In a nonparametric setting, where parameters are infinite-dimensional, it is unlikely that effective general methods of this type can be designed.

An alternative idea, proposed by Møller *et al.* (2006), enables exact MCMC simulation from Π^n through the introduction of an auxiliary variable s, with state space $\mathbb{Y}$, through the latent likelihood

$$f(s, y|\theta) = f(s|y, \theta) f(y|\theta) = f(s|y, \theta) \frac{g(y, \theta)}{Z(\theta)}. \tag{1.73}$$

MCMC simulation is implemented for the joint posterior distribution

$$\Pi^n(s, \theta) = \frac{\Pi(\theta) f(s|y, \theta) g(y, \theta)/Z(\theta)}{\int_{\mathbb{Y}} \int_{\Theta} \left[\Pi(\theta) f(s|y, \theta) g(y, \theta)/Z(\theta)\right] \mathrm{d}\nu(\theta, s)}. \tag{1.74}$$

The posterior density $\Pi^n(\theta)$ can be recovered from the above expression by marginalization over s, so a Markov chain with stable distribution defined by (1.74) would produce a sample with the desired marginal distribution.

The intractable constant $Z(\theta)$ still appears in this latent distribution. In order to deal with it, Møller *et al.* (2006) propose a MH scheme with proposal distribution given by

$$q(s', \theta'|s, \theta, y) = q(\theta'|\theta, y) q(s'|\theta'), \tag{1.75}$$

where $q(\theta'|\theta, y)$ represents the usual choice for the parameter update step. The additional term

$$q(s'|\theta') = \frac{g(s', \theta')}{Z(\theta')} \tag{1.76}$$

is designed to ensure that evaluation of the intractable ratio is not required for the calculation of the acceptance probability. The MH ratio in this case is given by

$$\frac{\Pi^n(s', \theta')q(s, \theta|s', \theta')}{\Pi^n(s, \theta)q(s', \theta'|s, \theta)} = \frac{\Pi(\theta')g(y, \theta')q(\theta|\theta', y)}{\Pi(\theta)g(y, \theta)q(\theta'|\theta, y)} \frac{g(s, \theta)f(s'|\theta', y)}{g(s', \theta')f(s|\theta, y)}. \tag{1.77}$$

The main assumption here is that an exact sample can be generated from the proposal distribution. In other words, it must be possible simulate observations from the model density $f(\cdot|\theta)$ for any possible parameter value $\theta \in \Theta$.

The proposal distribution q for the MH scheme is fixed, and so the performance of the algorithm in terms of the overall acceptance rate can only be affected by the choice of the target distribution, i.e. the conditional $f(s|\theta, y)$. A common choice is given by

$$f(s|\theta, y) = f(s|\hat{\theta}) = \frac{g(s, \hat{\theta})}{Z(\hat{\theta})}, \tag{1.78}$$

for some fixed value $\hat{\theta}$, which may be an estimate of the parameter based on some pseudo-likelihood approximation.

Murray *et al.* (2006) interpret this as a one-sample importance sampler where an estimate of the ratio of intractable normalizing constants is calculated as the ratio of two estimates given by

$$\frac{Z(\hat{\theta})}{Z(\theta')} \approx \frac{g(s', \hat{\theta})}{g(s', \theta')}, \quad s' \sim \frac{g(s', \theta')}{Z(\theta')}; \tag{1.79}$$

$$\frac{Z(\hat{\theta})}{Z(\theta)} \approx \frac{g(s, \hat{\theta})}{g(s, \theta)}, \quad s \sim \frac{g(s, \theta)}{Z(\theta)}. \tag{1.80}$$

They propose to improve the performance of the algorithm by substituting the single auxiliary variable s by a vector $s_{1:k} = \{s_1, \ldots, s_k\}$ of auxiliary variables, for some $k \geq 1$ chosen a priori and fixed throughout. They explore the performance of the algorithm as a function of k.

Alternative Markov chain constructions are also provided by Murray *et al.* (2006). which result in simpler and more efficient updating schemes. The idea

is to choose different proposal distributions for the MH scheme, which produce direct estimates of the ratio of normalizing constants $Z(\theta)/Z(\theta')$ instead of the ratio of estimates of the original algorithm. The choice of q ensures that the intractable component cancels out from the acceptance ratio, making the MCMC feasible. However, this methods still rely on the possibility of producing exact samples from $f(\cdot|\theta)$ in order to update the auxiliary variables.

In the following Chapters, we present some large families of models for which this assumption fails, due to the infinite-dimensional nature of the state space $\mathbb{Y}$ (Chapter 2) or the parameter space Θ (Chapters 3, 4 and 5). It is this type of models that provide a motivation for the present work, extending beyond the scope of normalizing constants to deal with other forms of intractability.

1.2.2 Latent Variables for MCMC Methods

The idea of extending a model by introducing latent variables, to enable or simplify MCMC simulation, is not limited to the Metropolis-Hastings updating schemes presented above. Auxiliary variable extensions have been used for many years within the conditional sampling step of Gibbs samplers. An overview of the early developments on the use of auxiliary variables for MCMC simulation can be found in Besag & Green (1993).

Assume one seeks to generate realizations from some density $f(y)$, which may be a full conditional within a Gibbs sampling scheme. In the most general setting, the variable of interest is augmented by an auxiliary variable s which may or may not have a physical interpretation. The conditional distribution $f(y|s)$ is specified, to produce a joint $f(y,s) = f(y)f(s|y)$. The desired sample can then be produced by constructing a Markov chain which converges to $f(y,s)$ and therefore, marginally, to $f(y)$. If a Gibbs sampling scheme is already in place, this is achieved simply by adding one step to the updating loop. Clearly, the usefulness of the method depends on the possibility of choosing a density which enables a simple simulation from both conditionals, $f(y|s)$ and $f(s|y)$.

Swendsen & Wang (1987) introduced a latent variable approach to improve the performance of the Gibbs sampler for the Potts model, a generalization of the popular Markov random field model known as the Ising model. A more general

methodology for enabling and simplifying MCMC simulation through the use of auxiliary variables was introduced by Damien *et al.* (1999). The method, commonly known as slice sampling, provides an alternative to the traditional Metropolis-Hastings and rejection-based methods. Their main contribution is showing that, for a general family of complex models, it is possible to introduce the latent variable s in such a way that direct simulation from the conditionals, specifically from $f(y|s)$ is possible.

The general idea behind the slice sampler is the following. Assume the target density $f(y)$ can be factorized as $f(y) \propto \pi(y)g(y)$, where π is a density and g is a non negative invertible function, in the sense that the sets $A_s = \{y : g(y) > s\}$ can be found. Damien *et al.* (1999) propose the introduction of a latent variable s with positive support, by defining the joint density

$$f(y, s) \propto \pi(y)\mathbf{1}\{s < g(y)\}. \tag{1.81}$$

In this case, the conditional density $f(s|y)$ is simply a uniform on $\big(0, g(y)\big)$ and the conditional $f(y|s)$ is $\pi(y)$ restricted to the set A_s. For a p-dimensional state space $\mathbb{Y}$, when sampling from the corresponding truncated density may be difficult, the problem is solved by updating each component consecutively, using the corresponding full conditionals. That is, it is necessary to sample from $\pi(y_j|y_{-j})$ restricted to the set $A_{j,s} = \{y_j : g(y) > s\}$, for which it is only required that $g(y)$ is invertible for given values of y_{-j}, for each $j = 1, \ldots, p$.

More flexibility may be achieved by substituting the uniform random auxiliary variable approach implicit in the use of $\mathbf{1}\{s < g(y)\}$, with a more general $\tilde{g}_j(s)\mathbf{1}\{s < \tilde{g}_2(y)\}$, which results in the need to sample from two truncated densities, corresponding to the full conditionals. There is no general way to choose the most convenient latent variable; suitable options depend on the context. However, Damien *et al.* (1999) provide many examples in which at least one choice may be found which allows direct sampling from the full conditionals.

Another way to define useful auxiliary variables for sampling from complex models, known as partial decoupling, is found in Higdon (1998). It is particularly useful in the context of Markov random fields.

Many other latent structures have been proposed, originating in specific models, with the purpose of improving the mixing or convergence properties of Markov

chain simulators, reduce the correlation of the samples or improve the quality of the estimates. We do not give here a full review on the subject, but present some particular simulation algorithms, based on latent variables, which are relevant for the present work.

1.2.2.1 Latent Variables for Truncated Density Simulation

In this Section, we illustrate the use of latent variables to simulate from a truncated density as part of an MCMC scheme. This is relevant for the applications we present in the following Chapters, which require the simulation from truncated normal and gamma densities. The results and methods in this section are taken from Damien & Walker (2001).

Let us first consider a truncated univariate normal density

$$f(y) \propto \exp(-y^2/2)\mathbf{1}\{a < y < b\}. \tag{1.82}$$

A latent variable s is introduced, via the joint density

$$f(y, s) \propto \mathbf{1}\{0 < s < \exp(-y^2/2)\}\mathbf{1}\{a < y < b\}, \tag{1.83}$$

which clearly has the desired marginal $f(y)$. Since $0 < s < \exp(-y^2/2)$ if and only if $|y| < \sqrt{-2\log s}$, the corresponding conditional distributions are given by

$$f(s|y) = \mathrm{U}\left(s|0, \exp(y^2/2)\right) \tag{1.84}$$

$$f(y|s) = \mathrm{U}\left(y|\max\left\{a, -\sqrt{-2\log s}\right\}, \min\left\{b, \sqrt{-2\log s}\right\}\right), \tag{1.85}$$

where $\mathrm{U}(\cdot|a, b)$ denotes the density function for the uniform distribution on (a, b).

When a sample from the truncated density (1.82) is required within a Gibbs sampling scheme, it may be obtained simply by adding an extra full conditional at each iteration. Thus, a rejection or MH step is substituted by the straightforward sampling of two uniformly distributed variables, which in many cases may be more efficient.

This idea is extended by Damien & Walker (2001) to the problem of sampling from a truncated multivariate normal density

$$f(y_1, \ldots, y_p) \propto \exp\left\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\} 1_A(y), \tag{1.86}$$

assuming the truncation region for y_i given all the rest can be written as (a_i, b_i). Once again, they introduce a latent variable s, through the joint density

$$f(y_1, \ldots, y_p, s) \propto \exp(-s/2)1\left\{s > \exp\left[-1/2(y-\mu)'\Sigma^{-1}(y-\mu)\right]\right\}1_A(y),$$

so the full conditional distributions are given by

$$\begin{aligned} f(s|y_1, \ldots, y_p) &\propto \exp(-s/2)1\left\{s > \exp\left[-1/2(x-\mu)'\Sigma^{-1}(x-\mu)\right]\right\}, \\ f(y_i|y_{-i}, s) &\propto 1\left\{y_i \in (a_i, b_i) \cap B_i\right\} \end{aligned}$$

where $B_i = \left\{y_i : (y-\mu)'\Sigma^{-1}(y-\mu) < s\right\}$. So once the bounds are found by solving the quadratic equation, the Gibbs sampler can be implemented by updating p uniform random variables and one truncated exponential. For the last one, the cdf inversion technique is suggested.

Another truncated density considered by Damien & Walker (2001) and used within the present work is the Gamma,

$$f(y) \propto y^{\alpha-1}\exp(-y)1\{a < y < b\}, \tag{1.87}$$

for some $0 \leq a < b \leq \infty$.

The latent variable extension proposed corresponds to the joint density

$$f(y, s) \propto y^{\alpha-1}1\{0 < s < \exp(-y)\}1\{a < y < b\}. \tag{1.88}$$

leading to full conditional distributions given by

$$\begin{aligned} f(s|y) &= \mathrm{U}\big(s|0, \exp(-y)\big), \\ f(y|s) &\propto y^{\alpha-1}1\left(a < y < \min\{b, -\log y\}\right), \end{aligned}$$

the second of which can once more be sampled using the cdf inversion technique.

In the following Chapters we use this idea to sample from full conditional densities within a more elaborate MCMC scheme, in cases where the truncated regions are more complex or where more general multivariate truncated densities are involved.

1.2.2.2 Slice Sampler for MDP Models

Sampling from the MDP model described in Section 1.1.1.2 is a complex problem, mainly because of the possibility of choosing form an infinite number of discrete mass points, arising from the Dirichlet process prior. The first algorithm to allow sampling from the posterior distribution defined by this model is owing to Escobar (1988). It is usually referred to as a marginal method, as it relies on integrating out the random distribution function, thus removing the infinite dimensionality problem. Many variations of this method have been defined over the years, for example by MacEachern & Müller (1998) and Neal (2000).

Even though marginal sampling methods may be sufficient for certain applications, it is sometime convenient to avoid the integration of the random measure. For example, when the random measure itself is an object of interest in the inference process. In such cases, it is preferable to sample using an MCMC scheme with includes the random measure in the updating process. The first algorithm of this type, known as conditional methods, was introduced by Ishwaran & Zarepour (2000), who proposed an approximation to the MDP model based on the hierarchical representation (1.19). Papaspiliopoulos & Roberts (2008) proposed an algorithm to produce an exact sample by using retrospective sampling techniques, while Walker (2007) and Kalli *et al.* (2011) achieve the same with a slice sampler.

In the following Chapters simulation for mixture models will be done using the slice sampling methodology. Therefore, we introduce the method, as presented by Kalli *et al.* (2011) for the MDP model with a Normal parametric kernel. This idea, as well as that of the retrospective sampler method, relies on the stick breaking representation (1.18). That is,

$$f(y|w_{1:\infty}, \theta_{1:\infty}) = \sum_{j=1}^{\infty} w_j K(y|\theta_j). \tag{1.89}$$

The prior Π is defined by $\theta_j \overset{iid}{\sim} P_0$ and, for a collection of random variables, $v_j \overset{ind}{\sim} \text{Be}(\alpha_j, \zeta_j)$ the weights are given by $w_j = v_j \prod_{j'<j}(1 - v_{j'})$, with $w_1 = v_1$.

The slice sampling method introduces suitable auxiliary variables, conditional on which, only a finite number of weights and particles needs to be sampled at

each step of an MCMC scheme for posterior simulation. The first latent variable to be introduced is a uniform random variable, u, taking values on $(0, 1)$. This results in the latent expression

$$f(y, u|w_{1:\infty}, \theta_{1:\infty}) = \sum_{j=1}^{\infty} K(y|\theta_j)\mathbf{1}\{u < w_j\}. \tag{1.90}$$

Let $A_u = \{j : w_j > u\}$. Since $\sum_j w_j = 1$, the weights must define a sequence decreasing to 0. Therefore, the cardinality $J_u = \sum_{j=1}^{\infty} \mathbf{1}\{w_j > u\}$ of A_u, is finite. Furthermore, conditional on u, the density for y is a finite mixture with J_y components:

$$f(y|u, w_{1:\infty}, \theta_{1:\infty}) = W_u^{-1} \sum_{j \in A_u} K(y|\theta_j), \tag{1.91}$$

where $W_u = \sum_{j \in A_u} w_j$. An additional latent variable, d may be introduced, to index the specific component from which the observation is generated. This results in the joint density

$$f(y, u, d|w_{1:\infty}, \theta_{1:\infty}) = K(y|\theta_j)\mathbf{1}\{u < w_d\}. \tag{1.92}$$

The important features of this expression are that given u, the index d can only take a finite number of values, and there are no sums involved, so the likelihood for n observations can be expressed as a simple product of terms. If we introduce a pair of latent variables (u, d) for each observation, the full likelihood for the model is given by the product

$$f(y_{1:n}, u_{1:n}, d_{1:n}|w_{1:\infty}, \theta_{1:\infty}) = \prod_{i=1}^{n} K(y_i|\theta_{d_i})\mathbf{1}\{u_i < w_{d_i}\}, \tag{1.93}$$

and the posterior distribution for the model can be identified as

$$\Pi^n(w_{1:\infty}, \theta_{1:\infty}|y_{1:n}, u_{1:n}, d_{1:n}) \propto \Pi(w_{1:\infty}, \theta_{1:\infty}) \prod_{i=1}^{n} K(y_i|\theta_{d_i})\mathbf{1}\{u_i < w_{d_i}\}. \tag{1.94}$$

Posterior simulation can now be carried out using a Gibbs sampler. However, a more efficient sampler (see Kalli *et al.*, 2011) can be defined if the uniform auxiliary variable of the latent model (1.92) is substituted by a non uniform term, resulting in the latent expression

$$f(y, u, d|w_{1:\infty}, \theta_{1:\infty}) = e^{\xi d} w_d K(y|\theta_d)\mathbf{1}\{u < e^{-\xi d}\}, \tag{1.95}$$

for some known $0 < \xi \leq 1$. A Gibbs sampler implemented for this method should control the size of the A_u and improve the mixing. In fact the terms $\left\{e^{\xi j}\right\}_{j \geq 1}$ may be substituted by any positive sequence, each of which will result in a different balance between algorithmic efficiency and computational time. For more information on suggested options, see Kalli *et al.* (2011).

The posterior distribution for the complete latent model is given by

$$\Pi^n(w_{1:\infty}\theta_{1:\infty}|y_{1:n}, u_{1:n}, d_{1:n}) \propto \Pi(w_{1:\infty}, \theta_{1:\infty}) \prod_{i=1}^{n} w_{d_i} e^{\xi d_i} K(y_i|\theta_{d_i}) \mathbf{1}\left\{u_i < e^{-\xi d_i}\right\}.$$

Posterior simulation can be carried out using a Gibbs sampler, at each step of which we need to update the latent variables $d_{1:n}$, $u_{1:n}$, and the variables weights and particles characterizing the density of interest, $w_{1:\infty}$, $\theta_{1:\infty}$. Clearly, it is not possible to update an infinite number of variables, but given the latent variables, it is enough to sample a sufficiently large, but finite, number of them. Exactly how many, can be inferred from the full conditionals distributions,

$$\Pi(\theta_j|\ldots) \propto P_0(\theta_j) \prod_{d_i=j} K(y_i|\theta_j); \tag{1.96}$$

$$\Pi(v_j \mid \ldots) = \mathrm{Be}(v_j \mid \hat{\alpha}_j, \hat{\zeta}_j); \tag{1.97}$$

$$\Pi(u_i \mid \ldots) = \mathrm{U}\left(u_i|0, e^{-\xi d_i}\right); \tag{1.98}$$

$$\Pi(d_i \mid \ldots) \propto w_{d_i} e^{\xi d_i} K(y_i|\theta_j) \mathbf{1}\left\{e^{-\xi d_i} > u_i\right\}; \tag{1.99}$$

where

$$\hat{\alpha}_j = \alpha_j + \sum_{i=1}^{n} \mathbf{1}\left\{d_i = j\right\} \tag{1.100}$$

$$\hat{\zeta}_j = \zeta_j + \sum_{i=1}^{n} \mathbf{1}\left\{d_i > j\right\}. \tag{1.101}$$

The weights are defined through the usual stick breaking construction (1.11).

In order to sample from $\Pi(d_i \mid \ldots)$ exactly, only a number $J_i = \lfloor -\xi^{-1} \log y_i \rfloor$ of the weights and particles is needed. Therefore, at any given iteration of the Gibbs sampler, we only need to update $J = \max_i\{J_i\}$ of them.

The best strategy for sampling the kernel parameters $\theta_j = (\mu_j, \sigma_j^2)$ will be determined by the base measure P_0 specification and the choice of kernel function

$K(\cdot|\theta)$. We discuss some adequate methods in the next Chapters, as required by different examples.

The method can be adapted to obtain a similar representation for the Geometric stick-breaking prior. It is this flexibility, along with the simplicity of the resulting updating procedure, that makes the slice sampler so useful for the latent variable estimation procedures we present in this thesis.

1.2.2.3 MCMC for Parameters of Unknown Dimension

Assume we have a family of models such that, for every $k \in \mathbb{K}$, the dimension of the parameter space Θ_k for the model likelihood $f(y_{1:n}|k, \theta_k)$, depends on the index k. In a Bayesian setting, uncertainty about the models is expressed through a prior $\Pi(k, \theta_k) = \Pi(k)\,\Pi(\theta_k|k)$. The posterior probability for model k is given by

$$\Pi_n(k|y_{1:n}) = \frac{1}{f(y_{1:n})} \int_{\Theta_k} f(y_{1:n}|k, \theta_k)\,\Pi(k)\,\Pi(\theta_k|k)\,\mathrm{d}\nu(\theta_k). \tag{1.102}$$

This expression, commonly known as the marginal likelihood for the model, does not have, in general a closed from. Therefore, inference is carried out through MCMC methods.

Early approaches (Chib, 1995; Chib & Greenberg, 1998) rely on independent MCMC simulation for each model k in order to estimate the marginal likelihoods and calculate Bayes factors for model selection. This idea, however, is only feasible when $\mathbb{K}$ is finite and relatively small.

An alternative idea (Carlin & Chib, 1995) is to implement MCMC simulation simultaneously over the indexing variable k and all possible model parameters. The compound space $\mathbb{K} \times \prod_{k\in\mathbb{K}} \Theta_k$ has a fixed dimension but it may be too big for the MCMC methods to be of use. Godsill (2001) proposes a general methodology based on Metropolis-Hastings and Gibbs sampling schemes, using the relationship between the different models to make the sampling more efficient. The method is particularly useful in the case when the models have some sort of nesting structure. It has the advantage of relying on basic MCMC simulation ideas, and therefore any convergence properties of the Markov chains may be verified in the standard way.

In the present work, we make use of Godsill's version of the reversible jump sampler (Green, 1995), based on a Metropolis-Hastings construction with a specific choice of proposal distribution.

We consider a fully nested, possibly infinite, family of models. Let $\mathbb{K} = \mathbb{N}$ and, for every $k > 0$, assume model k has a parameter $\theta_{1:k} = (\theta_1, \ldots \theta_k)$, such that $\theta_j \in \Theta$ for all j and for $k' < k$ the first k' elements of $\theta_{1:k}$ coincide with $\theta'_{1:k'}$. In this case, Godsill's algorithm uses a Metropolis sampling scheme for (k, θ) in the infinite dimensional space $\mathbb{N} \times \Theta^{\mathbb{N}}$. The proposal distribution for updating from a state $(k, \theta_{1:k})$ to a state $(k', \theta'_{1:k'})$ takes the form

$$p(k', \theta'_{1:k'}|k, \theta_{1:k}) = p_1(k'|k)p_2(\theta'_{1:k'}|\theta_{1:k})\Pi(\theta'_{k'+1:\infty}|\theta'_{1:k'}), \tag{1.103}$$

where

$$p_2(\theta'_{1:k'}|\theta_{1:k}) = \begin{cases} q(\theta'_{k+1:k'}|\theta_{1:k})\mathbf{1}_{\theta_{1:k}}(\theta'_{1:k}) & \text{if } k' > k \\ \mathbf{1}_{\theta_{1:k'}}(\theta'_{1:k'}) & \text{if } k' \leq k \end{cases}, \tag{1.104}$$

and $\Pi(\theta_{k+1:\infty}|\theta_{1:k})$ is a pseudo-likelihood, in the sense that

$$f(k, \theta|y_{1:n}) = f(k, \theta_{1:k}|y_{1:n})\Pi(\theta_{k+1:\infty}|\theta_{1:k}). \tag{1.105}$$

It therefore takes advantage of the nesting structure to minimize the number of variables that need to be sampled.

The acceptance probability corresponding to this proposal is given by

$$\begin{aligned} \alpha &= \min\left\{1, \frac{p(k, \theta_{1:k}|k', \theta'_{1:k'})f(k', \theta'_{1:k'}|y_{1:n})}{p(k', \theta'_{1:k'}|k, \theta_{1:k})f(k, \theta_{1:k}|y_{1:n})}\right\} \\ &= \min\left\{1, \frac{p_1(k|k')p_2(\theta_{1:k}|\theta'_{1:k'})f(k', \theta'_{1:k'}|y_{1:n})}{p_1(k'|k)p_2(\theta'_{1:k'}|\theta_{1:k})f(k, \theta_{1:k}|y_{1:n})}\right\}. \end{aligned} \tag{1.106}$$

If, $k' > k$, this becomes

$$\alpha = \min\left\{1, \frac{p_1(k|k')f(k', \theta'_{1:k'}|y_{1:n})}{p_1(k'|k)q(\theta'_{k+1:k'}|\theta_{1:k})f(k, \theta_{1:k}|y_{1:n})}\right\}; \tag{1.107}$$

when $k > k'$, we have

$$\alpha = \min\left\{1, \frac{p_1(k|k')q(\theta_{k'+1:k}|\theta'_{1:k'})f(k', \theta'_{1:k'}|y_{1:n})}{p_1(k'|k)f(k, \theta_{1:k}|y_{1:n})}\right\}. \tag{1.108}$$

So the acceptance ratio does not depend on any parameter value θ_j or θ'_j for $j > \max\{k, k'\}$ and therefore only a finite number of variables needs to be updated at any step of the algorithm.

1.2.3 Exact Simulation and Inference for Diffusions

Auxiliary variable schemes can also be useful outside the context of MCMC methods. We illustrate this in the present Section.

Recall the discretely observed diffusion model described in Section 1.1.2.1. Throughout this section, we use $y_{1:n} = (y_{t_1}, \ldots, y_{t_2})$ to denote a sample of size n, for fixed, known times $0 < t_1 < \ldots < t_n < \infty$. In other words, we assume each observation y_{t_i} is the observed value, at time t_i, of a single realization, or path, of a diffusion process $Y = \{Y_t : t \geq 0\}$, defined by an SDE

$$\mathrm{d}Y_t = \alpha_\theta(Y_t)\mathrm{d}t + \mathrm{d}W_t, \tag{1.109}$$

and started at some fixed, known $Y_0 = y_0$. We assume that, for every $\theta \in \Theta$, condition (1.41) is satisfied, so that the process is well defined and Girsanov's change of measure formula (Theorem 3) applies. Under this assumption, the transition densities for the process exist; in most cases, however, they do not have an analytic form. Therefore, the likelihood of the discretely observed diffusion model, given by

$$f(y_{1:n}|\theta) = \prod_{i=1}^{n} f_{\Delta_i}(y_i|y_{i-1}, \theta), \quad \Delta_i = t_i - t_{i-1}, \tag{1.110}$$

is intractable.

Many methods have been proposed to face this issue. Until recently, they all relied on different forms of approximation or interpolation techniques, including approximate simulation; analytic approximations of the transition density or the complete likelihood functions; and direct approximation of the maximum likelihood estimator. Some of these methods are described in Barndorff-Nielsen & Sorensen (1994); Bibby & Sorensen (1995); Kelly *et al.* (2004); and Sorensen (2004) gives a review of them.

An important theoretical breakthrough, was brought about by the definition of a method, known as the exact simulation algorithm, which allows the simulation of diffusion paths at arbitrary time points within a closed time interval $[0, t]$, with no approximation error. The exact simulation algorithm, first presented by Beskos & Roberts (2005), is a retrospective rejection sampler, based on a

factorization of the diffusion paths in terms of a finite set of points, known as the skeleton, connected by independent Brownian bridges. Initially designed for a limited family of diffusion processes with rather restrictive conditions on the drift and diffusion coefficients, the result was later extended by Beskos *et al.* (2006a), Beskos *et al.* (2006b) and Beskos *et al.* (2009), to cover most of the diffusion process commonly used for statistical modelling. In the present work, however, we focus on the simplest version of the algorithm (EA1 in Beskos *et al.*, 2006b), as it suffices for illustrative purposes.

Since, the finite dimensional distributions of the process are, as the transition density, generally unavailable, the first key for the exact simulation of diffusion paths, is to express the law of the diffusion of interest, in terms of a Brownian Motion, for which the transition densities are known and easy to simulate from. Beskos *et al.* (2006b) achieve this through the application of Girsanov's formula (1.29).

For every fixed $t > 0$ the density of the law $\mathbb{P}^{y_0}$ of the diffusion Y started at $Y_0 = y_0$, restricted to $\mathcal{A}_t$, with respect to the Weiner measure $\mathbb{W}^{y_0}$ is given by

$$\tilde{f}_t(\cdot|y_0, \theta) = \frac{\mathrm{d}\mathbb{P}_t^{y_0}(\theta)}{\mathrm{d}\mathbb{W}^{y_0}} = \exp\left\{\int_0^t \alpha_\theta(Y_s)\mathrm{d}Y_s - \frac{1}{2}\int_0^t \alpha_\theta^2(Y_s)\mathrm{d}s\right\}. \tag{1.111}$$

We use the notation $\tilde{f}$ to indicate this is not a transition density with respect to Lebesgue measure, but a density on $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}))$ with respect to a Weiner measure.

In order to deal with the stochastic integral in the above expression, assume that the drift coefficient α_θ is continuously differentiable and integrable, for every $\theta \in \Theta$. Denote by

$$A_\theta(u) = \int \alpha_\theta(u)\mathrm{d}u \tag{1.112}$$

some antiderivative of α_θ, so that $A'_\theta = \alpha_\theta$ for every $\theta \in \Theta$. Then, an adequate version of Itô's formula (1.26) can be applied and equation (1.111) above becomes

$$\tilde{f}_t(\cdot|y_0, \theta) = \exp\left\{A_\theta(Y_t) - A_\theta(Y_0) - \frac{1}{2}\int_0^t \left[\alpha_\theta^2(Y_s) + \alpha'_\theta(Y_s)\right]\mathrm{d}s\right\}. \tag{1.113}$$

Since we are considering stochastic processes defined on the canonical space $(\mathcal{C}_{[0,\infty)}, \mathcal{B}(\mathcal{C}_{[0,\infty)}))$, all random variables involved are defined by the coordinate

mapping, i.e. for each $y \in \mathcal{C}_{[0,\infty)}$ and $t \geq 0$, $Y_t(y) = y_t$. Therefore, for any continuous function y, we can write

$$\tilde{f}_t(y|y_0, \theta) = \exp\left\{ A_\theta(y_t) - A_\theta(y_0) - \frac{1}{2}\int_0^t [\alpha_\theta^2(y_s) + \alpha_\theta'(y_s)]\mathrm{d}s \right\}.$$

This expression is the base for a simple, but impossible, rejection algorithm, in which a Brownian motion path $y = \{y_s : 0 < s < t\}$ started at y_0 is simulated, and accepted with probability proportional to $\tilde{f}_t(y|y_0, \theta)$. The impossibility comes from the fact that, even though the finite dimensional distributions of Brownian motion are known multivariate normal distributions, this is enough only to simulate a finite number of points in a Brownian motion path. The integral in the above expression, however, depends on the complete function $y = \{y_s : 0 < s < t\}$, making it intractable.

Beskos *et al.* (2006a) deal with this problem by introducing a set of latent variables. In the simplest form of the exact simulation such variables are defined through an auxiliary homogeneous Poisson process in the following manner.

Assume that the drift coefficient of SDE (1.109) is such that, for every $\theta \in \Theta$ we can write

$$l(\theta) \leq \inf_{u \in \mathbb{R}} \left\{ [\alpha_\theta^2(u) + \alpha_\theta'(u)]/2 \right\}; \tag{1.114}$$

$$r(\theta) \geq \sup_{u \in \mathbb{R}} \left\{ [\alpha_\theta^2(u) + \alpha_\theta'(u)]/2 - l(\theta) \right\}, \tag{1.115}$$

for some $l : \Theta \to \mathbb{R}$ and $r : \Theta \to (0, \infty)$. It is then possible to define a bounded function $\varphi_\theta : \mathbb{R} \to [0, 1]$ as

$$\varphi_\theta(u) = \frac{1}{r(\theta)} \left(\frac{\alpha_\theta^2(u) + \alpha_\theta'(u)}{2} - l(\theta) \right). \tag{1.116}$$

The expression for $\tilde{f}$ can then be rewritten in terms of φ_θ as

$$\tilde{f}_t(y|y_0, \theta) = \exp\{ A_\theta(y_t) - A_\theta(y_0) - l(\theta) \} \exp\left\{ -r(\theta) \int_0^t \varphi_\theta(y_s)\mathrm{d}s \right\}. \tag{1.117}$$

The second key for the exact simulation of diffusion paths is the realization that

$$\exp\left\{ -r(\theta) \int_0^t \varphi_\theta(y_s)\mathrm{d}s \right\} \tag{1.118}$$

is the probability that a realization of a homogeneous Poisson point process on $[0, t] \times [0, 1]$, with intensity $r(\theta)$ has 0 points under the graph $s \mapsto \varphi_\theta(y_s)$. This allows the evaluation of the acceptance probability for a Brownian path proposal, based only on a finite number of points, generated retrospectively, at times determined by the Poisson process.

The exact simulation algorithm is therefore defined as follows:

i) Generate a realization of the Poisson process, i.e. a Poisson random variable k with mean parameter $t\, r(\theta)$, and, conditional on k a set $(\tau_1, \ldots, \tau_k)$ of i.i.d. uniform random variables on $[0, t]$ and a set $(u_1, \ldots, u_k)$ of i.i.d. uniform random variables on $[0, 1]$ and independent of the $\tau_{1:k}$.

ii) Simulate $(y_{\tau_1}, \ldots, y_{\tau_k})$ from the k-dimensional distribution of a Brownian motion started at y_0.

iii) If there are no points of the Poisson process under the graph $s \mapsto \varphi_\theta(y_s)$, in other words, if

$$\prod_{j=1}^{k} \mathbf{1}\left\{\varphi_\theta(y_{\tau_j}) < u_j\right\} = 1, \tag{1.119}$$

then accept the Brownian path as a realization of the diffusion process.

In reality, the algorithm accepts simultaneously any complete Brownian path $\{y_s : 0 \leq s \leq t\}$ passing through $(y_{\tau_1}, \ldots, y_{\tau_k})$. Therefore, for arbitrary times $0 < t_1 < \ldots < t_n \leq t$, the corresponding points of the diffusion path can be simulated via Brownian bridge interpolation between $(y_{\tau_1}, \ldots, y_{\tau_k})$. Since the finite dimensional distributions of a Brownian bridge are simply multivariate normal distributions with known mean vectors and covariance matrices, this second stage can be carried out without problems.

The accepted pairs $(\tau_j, y_{\tau_j})_{j=1}^{k}$ are known as the skeleton of the path, and conditional on the skeleton and the skeleton size, k, the rest of the path is simply a set of independent Brownian bridges. More about the factorization of diffusions into Brownian bridges, and milder conditions on the drift and diffusion coefficients, can be found in Beskos *et al.* (2008).

For every time interval $[0, t]$, the exact simulation algorithm accepts or rejects a complete skeleton simultaneously. Since the expected number of points of a proposed skeleton is $t\, r(\theta)$, the acceptance rate decreases as t increases. An optimal acceptance rate is achieved when $t = 1/r(\theta)$ (see Beskos *et al.*, 2006a). Therefore, for larger values of t, a good performance of the algorithm requires that the time interval of interest is split into smaller intervals and then the Markov property used to produce the complete path. Clearly, the number of such smaller intervals will grow with t, affecting the performance of the algorithm. In the next Chapter, we propose an alternative MCMC scheme suitable both for simulation of diffusion paths and Bayesian inference, which is not based on rejection sampling and therefore does not need to be adapted depending on the size of the time interval under consideration.

Beskos *et al.* (2006a) propose the use of the exact simulation algorithm for estimation of the parameter and the transition density for the model, mainly focusing on maximum likelihood estimators and their properties. To do so, they observe that the transition density of the diffusion, with respect to $\mathbb{W}^{y_0}$, can be obtained from equation (1.117), by integrating out the rest of the path. In other words, we can write

$$\tilde{f}_t(y_t|y_0,\theta) = \exp\{A_\theta(y_t) - A_\theta(y_0) - l(\theta)\}\mathbb{E}_{\mathbb{W}^{y_0}}\left[\exp\left\{-r(\theta)\int_0^t \varphi_\theta(y_s)\mathrm{d}s\right\}\Big|y_t\right].$$

And, since the density with respect to Lebesgue measure of any point of a Brownian motion path is a known univariate normal density, a change of measure leads to an expression for the transition density of the diffusion (with respect to Lebesgue measure),

$$\begin{aligned} f_t(y_t|y_0,\theta) = \mathrm{N}(y_t|y_0,t)\exp\{A_\theta(y_t) - A_\theta(y_0) - l(\theta)\} \\ \mathbb{E}_{\mathbb{W}^{y_0}}\left[\exp\left\{-r(\theta)\int_0^t \varphi_\theta(y_s)\mathrm{d}s\right\}\Big|y_t\right]. \end{aligned} \tag{1.120}$$

The expectation term in the above expression is intractable, but it coincides with the acceptance probability for the exact simulation algorithm, when y_t is fixed. This is the base for the inference methods studied by Beskos *et al.* (2006a,b, 2009).

Even though it is not stated explicitly by the authors, the exact simulation algorithm, defines a latent expression for the transition function, given by

$$f_t(y_t, k, u_{1:k}, \tau_{1:k}, y_{\tau_1}, \dots y_{\tau_k}|y_0, \theta) = \mathrm{N}(y_t|y_0, t)\exp\{A_\theta(y_t) - A_\theta(y_0) - l(\theta)\} \frac{[r(\theta)]^k}{k!}\prod_{j=1}^{k} 1\{\varphi_\theta(y_{\tau_j}) < u_j\}.$$

Or, integrating out the uniform random variables $u_{1:k}$,

$$f_t(y_t, k, \tau_{1:k}, y_{\tau_1}, \dots y_{\tau_k}|y_0, \theta) = \mathrm{N}(y_t|y_0, t)\exp\{A_\theta(y_t) - A_\theta(y_0) - l(\theta)\} \frac{[r(\theta)]^k}{k!}\prod_{j=1}^{k}[1 - \varphi_\theta(y_{\tau_j})]. \quad (1.121)$$

Beskos *et al.* (2006b) consider Bayesian estimation for discretely observed diffusions, using this expression. If a prior Π is defined on the parameter space Θ, inference can be carried out through MCMC methods in the following manner.

i) Initialize the Markov chain by choosing some value θ for the parameter.

ii) Through the use of the exact simulation algorithm, generate independent Skeletons for a diffusion path, between consecutive observations, given the current parameter value.

iii) Update the value of the parameter by sampling from the full conditional distribution, given the skeletons and the observations. The full conditional density is proportional to the prior Π multiplied by the product of the latent transition densities for all the data points. Since this density depends only on a finite number of points and can be evaluated up to proportionality, any usual MCMC simulation method can be used to generate the new θ.

In Chapter 2, we show how the latent model used for this inference method can be seen as a particular case of the general latent model given by expression (15). We provide an expression for the full model, using a change of notation, and propose an alternative MCMC algorithm which can be used both for simulation of diffusion sample paths and for Bayesian posterior inference.

1.3 Bayesian consistency

In a Bayesian nonparametric setting, the simplest scenario assumes $\{Y_i\}_{i\geq 1}$ is a sequence of i.i.d. random variables from a distribution with probability density f_0 defined on the sample space $\mathbb{Y}$. Inference begins by defining a prior Π on the set $\mathcal{F}$ of density functions over $\mathbb{Y}$. Each observation y_i is assumed to be a realization of the variable Y_i. We say the model is consistent at f_0 if the posterior probability accumulates all of its mass around f_0.

There is some disagreement between Bayesian statisticians about the usefulness of the consistency property, mostly arising from the different views about the justification behind Bayesian procedures. In short, some Bayesians do not agree that a single f_0 exists for which the modelled random variables are i.i.d., since, under an exchangeability assumption, de Finetti's theorem guarantees only conditional independence. In the present work, we study consistency as a property that some people may want to verify and refer the reader to Diaconis & Freedman (1986) for arguments supporting the relevance of consistency in Bayesian procedures.

Given a sample of size n, the posterior mass assigned to a set $A \subset \mathcal{F}$ is given by

$$\Pi^n(A) = \frac{\int_A R_n(f)\Pi(\mathrm{d}f)}{\int R_n(f)\Pi(\mathrm{d}f)}, \tag{1.122}$$

where

$$R_n(f) = \prod_{i=1}^{n} \frac{f(y_i)}{f_0(y_i)} \tag{1.123}$$

is the likelihood ratio between f and f_0. As it stands, if the posterior Π^n accumulates its mass around f_0 as n grows, this would only describe a behaviour for the particular sample at hand. A more useful property would give us information on the behaviour of the posterior, regardless of the particular sequence observed.

Doob (1949) showed that, under weak conditions, consistency follows for Π almost every observed sequence. This, however, is not enough for a practical interpretation, since the true density can fall on a null set of the prior, in which case consistency fails for f_0. Therefore, a stricter and more formal definition of consistency is required.

Effectively, the idea of Bayesian consistency is that, as more data is gathered, it should be possible to identify the true density f_0 generating the data more accurately, for almost every sequence we may observe, i.e. almost surely with respect to the joint law of the complete sequence.

To formalize, if we denote by $\mathbb{P}_0$ the probability measure corresponding to f_0, we may think of the sequence $\{Y_i\}_{i\geq 1}$ as defined on the product space $\mathbb{Y}^\infty$, with joint probability measure $\mathbb{P}_0^\infty$. We say that the Bayesian model with prior Π is consistent if for every neighbourhood B of f_0, we have

$$\Pi^n(B^c|Y_0,\ldots,Y_n) \xrightarrow[n\to\infty]{} 0 \quad \text{a.s. } [\mathbb{P}_0^\infty]. \tag{1.124}$$

In this case, the posterior distribution is considered as a random object, due to its dependence on the random sample $Y_{1:n} = \{Y_1,\ldots,Y_n\}$. However, once this has been clarified, we use the notation $y_{1:n}$, even when probability statements refer, more formally to the random variables $Y_{1:n}i$. Throughout the rest of this chapter, all such probability statements are made with respect to $\mathbb{P}_0^\infty$.

Clearly, the concept of consistency depends on the definition of B, that is, on the topology imposed on the functional space $\mathcal{F}$. Different topologies lead to different types of consistency and we consider the two most relevant cases below.

1.3.1 Weak Consistency

A common topology to consider when dealing with functional spaces is the weak topology, associated to the concept of weak convergence. It is said that $B \subset \mathcal{F}$ is a weak neighbourhood of f_0 if it contains a set of the form

$$\left\{ f \in \mathcal{F} : \left| \int \phi_i f - \int \phi_i f_0 \right| < \varepsilon, i = 1,\ldots,\kappa \right\}, \tag{1.125}$$

where the $(\phi_i)_{i=1}^\kappa$ are bounded continuous functions, for some $\kappa \in \mathbb{N}$.

A density $f \in \mathcal{F}$ is in the weak support of the prior if every weak neighbourhood B of f has positive prior probability, i.e. $\Pi(B) > 0$.

Diaconis & Freedman (1986) proved that even when f_0 is in the weak support of the prior, weak consistency does not follow. A stronger condition is required, to guarantee the prior assigns enough mass on tighter neighbourhoods of f_0.

Such tighter neighbourhoods are defined in terms of the topology induced by the Kullback-Lebiler divergence on $\mathcal{F}$.

For every pair of functions $f_1, f_2 \in \mathcal{F}$, the Kullback-Leibler divergence from f_1 to f_2 is given by

$$K(f, \tilde{f}) = \int f_2 \log \frac{f_2}{f_1}. \tag{1.126}$$

This does not define a distance since, in particular, it is not symmetric. However, the Kullback-Leibler divergence can be used to define a system of neighbourhoods in a space of density functions, thus inducing a topology. For every $f \in \mathcal{F}$, a Kullback-Leibler neighbourhood of f is constructed as a countable union of balls of the form

$$B_K(f, \varepsilon) = \{\tilde{f} \in \mathcal{F} : K(\tilde{f}, f) < \varepsilon\}. \tag{1.127}$$

A density $f \in \mathcal{F}$ is in the Kullback-Leibler support of the prior if every weak neighbourhood B_K of f has positive prior probability, i.e. $\Pi(B_K) > 0$. If f_0 is in the Kullback-Leibler support of the prior, that is

$$\Pi[B_K(f_0, \varepsilon)] > 0 \quad \forall \varepsilon > 0, \tag{1.128}$$

it is said that the Bayesian model satisfies the Kullback-Leibler property.

Schwartz (1965) showed that the Kullback-Leibler property is a sufficient condition for weak consistency. This condition is stronger than the requirement of f_0 being in the weak support of the prior, since Kullback-Leibler neighbourhoods of a function are contained in weak neighbourhoods.

Weak consistency is a desirable property when inference is focused on estimating specific quantities related to the density, such as means or variances. However, as noted by Barron *et al.* (1999), if the interest of the inference is the density itself, it is convenient to seek convergence in some stronger sense, as weak neighbourhoods of f_0 may contain densities which do not truly resemble f_0.

1.3.2 Strong consistency

A Bayesian model is said to be strongly consistent when the posterior density accumulates all of its mass around strong neighbourhoods of the true density f_0

as n grows. The strong topology in a functional space is usually defined with respect to the L_1 distance, given by

$$L_1(f_1, f_2) = \int_{\mathbb{Y}} |f_1(y) - f_2(y)| \mathrm{d}\nu(y). \tag{1.129}$$

When limited to a space of density functions $\mathcal{F}$, the L_1 distance is equivalent to the total variation metric on the corresponding space $\mathcal{P}$ of probability measures, defined by

$$d_T(P_1, P_2) = \sup_{B \in \mathcal{B}(\mathbb{Y})} |P_1(B) - P_2(B)|. \tag{1.130}$$

The advantage of using the L_1 metric is that functional analysis results guarantee the separability of $\mathcal{F}$ with respect to L_1. Furthermore, it can be shown that the Hellinger H distance on $\mathcal{F}$, given by

$$H^2(f_1, f_2) = \frac{1}{2} \int_{\mathbb{Y}} \left[\sqrt{f_1(y)} - \sqrt{f_2(y)}\right]^2 \mathrm{d}\nu(y) = 1 - \int_{\mathbb{Y}} \sqrt{f_1(y) f_2(y)}\, \mathrm{d}\nu(y), \tag{1.131}$$

is topologically equivalent to the L_1 distance, with

$$H^2(f_1, f_2) \leq L_1(f_1, f_2) \leq \sqrt{2}\, H(f_1, f_2), \tag{1.132}$$

for every $f_1, f_2 \in \mathcal{F}$. This implies that convergence with respect to the L_1 distance and convergence with respect to the Hellinger distance are equivalent. This last one being more manageable in many calculations, strong consistency for density estimation is usually defined in terms of the Hellinger distance.

Following that convention, we say that the Bayesian model is strongly consistent when

$$\Pi_n(A_\varepsilon) \to 0 \quad \text{a.s. for all } \varepsilon > 0.$$

where

$$A_\varepsilon = \{f \in \mathcal{F} : H(f, f_0) > \varepsilon\}$$

is a set of densities ε-bounded away from f_0 with respect to the Hellinger distance.

The Kullback-Leibler property ensures the prior probability accumulated around f_0 is large enough so that, as the sample size n grows, the posterior probability assigned to weak neighbourhood does not vanish to zero. However, in order to achieve strong consistency, it is also necessary to ensure that the prior does not

concentrate too much mass on densities which can track the data. Therefore, more restrictive conditions are required for a model to be strongly consistent.

Recall the posterior mass assigned to a A_ε is given by

$$\Pi^n(A_\varepsilon) = \frac{\int_{A_\varepsilon} R_n(f)\Pi(\mathrm{d}f)}{\int R_n(f)\Pi(\mathrm{d}f)}. \tag{1.133}$$

Establishing convergence of this ratio proves a challenging task. However, The denominator does not depend on the set A_ϵ and therefore its treatment is independent of the topology defined on $\mathcal{F}$, so different approaches treat the numerator and the denominator separately.

The first result providing sufficient conditions for strong consistency is due to Barron *et al.* (1999). They show that, under the Kullback-Leibler property

$$\int R_n(f)\Pi(\mathrm{d}f) > \exp(-nc) \quad \text{a.s.}, \tag{1.134}$$

for any $c > 0$ and sufficiently large n. Therefore, the denominator cannot decrease to zero faster than at exponential rate. In order to guarantee strong consistency, a second condition is required, to guarantee the convergence to zero of the numerator at a faster rate. Barron *et al.* (1999) and Ghosal *et al.* (1999) provide such condition, namely, the existence of a sequence $(\mathcal{F}_n)_{n\geq 1} \subset \mathcal{F}$ such that for every large n

i) $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_2)$

ii) $J(\delta, \mathcal{F}_n) < n\beta$,

for some $c_1, c_2 > 0$ and $0 < \delta, \beta$ sufficiently small. The difference between the two results is in the definition of the $J(\delta, \mathcal{F}_n)$. For Ghosal *et al.* (1999), it denotes the L_1 entropy, i.e. the minimum k such that $\mathcal{F}_n$ can be expressed as a union of k balls of L_1-size δ around f_0; while for Barron *et al.* (1999) it denotes the bracketed entropy and is therefore slightly more restrictive.

The increasing sequence $(\mathcal{F}_n)_{n\geq 1} \subset \mathcal{F}$ is called a sieve and under the above condition and the Kullback-Leibler property, Ghosal *et al.* (1999) construct a sequence of uniformly consistent test to prove $f = f_0$ against $f \in \mathcal{F}_n \cap A_\varepsilon$. Therefore, we refere to this as the sieve and uniformly consistent test approach to consistency.

The construction of sieves and tests is often difficult, thus the relevance of an alternative approach proposed by Walker (2003, 2004). The introduction of this idea requires some additional notation.

Let

$$L_{nA} = \int_A R_n(f)\Pi(\mathrm{d}f)$$

denote the integrated likelihood ratio over a measurable subset $A \subset \mathcal{F}$. Then, the posterior mass assigned to A can be expressed as

$$\Pi^n(A) = \frac{L_{nA}}{I_n}, \tag{1.135}$$

where $I_n = L_{n\mathcal{F}} = \int R_n(f)\Pi(\mathrm{d}f)$.

The predictive density, with posterior restricted to the set A is given by

$$f_{nA}(y) = \int_A f(y)\mathrm{d}\Pi_A^n(f); \quad y \in \mathbb{Y},$$

where

$$\mathrm{d}\Pi_A^n(f) = \frac{\mathbf{1}(f \in A)\mathrm{d}\Pi^n(f)}{\int_A \mathrm{d}\Pi^n(f)}.$$

Assuming the existence of a sequence $(f_j)_{j\geq 1} \subset A_\varepsilon^c$ such that, for some $\delta < \varepsilon$,

$$A \subset \bigcup_{j=1}^{n} A_j; \quad A_j = \{f \in \mathcal{F} : H(f, f_j) < \delta\} \quad \forall\, j$$

and

$$\sum_{j=1}^{\infty} \sqrt{\Pi(A_j)} < \infty, \tag{1.136}$$

Walker (2004) proves that

$$L_{nA_\varepsilon} = \int R_{nA_\varepsilon}(f)\Pi(\mathrm{d}f) < \exp(-nd) \quad \text{a.s.}$$

for any $0 < d < -\log(\delta + 1 - \varepsilon)$ and sufficiently large n.

Combined with the exponential bound (1.134) provided by the Kullback-Leibler condition, this implies strong consistency. Since the space $\mathcal{F}$ of densities

is known to be separable with respect to the Hellinger distance, a countable cover for A_ε of Hellinger-size δ is always available. Therefore, this result adds to the Kullback-Leibler property a single condition on the prior, given by expression (1.136).

The key to Walker's result is the identity

$$\frac{L_{n+1A}}{L_{nA}} = \frac{f_{nA}(y_{n+1})}{f_0(y_{n+1})}. \tag{1.137}$$

In Chapter 6 we use an analogous expression to find sufficient conditions for strong consistency in the context of transition density estimation for Markov models.

It can be seen that $\{L_{nA}\}_{n\geq 0}$ defines a martingale and, even though this is not relevant to the consistency result, Walker's method has come to be known as the martingale approach.

For general models, conditions for strong consistency may hold, but be difficult to verify. Moreover, even if the conditions provided above fail, it does not follow that the model is not strongly consistent, since all of the results for strong consistency found in the literature provide conditions which are sufficient but not necessary. Walker & Hjort (2001) argue that, in this cases, it may be preferable to base Bayesian estimation on a consistent sequence $(Q_n)_{n\ ge1}$ of pseudoposterior distributions.

For each $n \in \mathbb{N}$ and some $\alpha \in (0,1)$, they define a probability measure Q_n by

$$Q_n(A) = \frac{\int_A R_n^{1-\alpha}(f)\,\Pi(\mathrm{d}f)}{\int R_n^{1-\alpha}(f)\,\Pi(\mathrm{d}f)}$$

and then use the sieve and uniformly consistent test approach of Barron *et al.* (1999) to prove that the Kulback-Leibler property alone guarantees that

$$Q_n(A_\varepsilon) \to 0 \quad \text{a.s. for all } \varepsilon > 0.$$

In other words, inference based on the $(1-\alpha)$-power likelihood results in strong consistency estimates for the true density f_0.

In Chapter 5 we deal with the problem of Bayesian inference for this type of power likelihood, for a large family of Bayesian nonparametric mixture models, and this last result provides one of the motivations.

1.3.3 An Important Counterexample

As we have mentioned above, existing results for strong consistency provide only sufficient conditions. This may raise the question of whether any condition, other than the Kullback-Leibler property is necessary. In this section we present an interesting example constructed by Barron *et al.* (1999) to show that the Kullback-Leibler property is not enough to guarantee posterior consistency when nonparametric densities are involved.

The idea is to define a prior which assigns equal probability to a set $\mathcal{F}_\Theta$ of continuous densities and a set $\mathcal{F}_*$ of piecewise constant densities. The roll of the first set is to ensure the Kullback-Leibler property is satisfied. while the second ensures posterior probability does not accumulate almost surely on arbitrarily small Hellinger neighbourhoods of the true density for the data.

The example starts by assuming we have a sequence $(Y_n)_{n\geq 1}$ of i.i.d. random variables uniformly distributed on $[0,1]$, so $f_0(x) = 1$. To construct the prior, first consider, for each positive integer N, the following partition of $[0,1]$

$$I_N = \left\{[0, 1/2N^2), [1/2N^2, 2/2N^2), \ldots, [(2N^2-1)/2N^2, 1]\right\}. \tag{1.138}$$

Let $\mathcal{F}_N$ be the set of all density functions which are constant on every interval of I_N and take only the values 0 and 2. Then, the cardinality of $\mathcal{F}_N$ is $q_N = \binom{2N^2}{N^2}$. The prior will assign equal mass

$$\Pi(f) = \frac{1}{Cq_N 2N^2} \tag{1.139}$$

to every function $f \in \mathcal{F}_N$, where C is a normalizing constant,

$$C = \sum_{N=1}^{\infty} \frac{1}{N^2}. \tag{1.140}$$

Making $\mathcal{F}_* = \bigcup_{N=1}^{\infty} \mathcal{F}_N$, this means exactly $1/2$ of the prior probability is accumulated on $\mathcal{F}_*$.

The rest of the prior mass is assigned to the parametric family

$$\mathcal{F}_\Theta = \left\{f_\theta = \exp(\theta + \sqrt{2\theta}\Phi^{-1}) : \theta \in (0,1)\right\}, \tag{1.141}$$

with probability induced by the density on the parameter space:

$$\Pi(\theta) \propto \exp\left(-1/\theta\right) \mathbf{1}\left\{0 < \theta < 1\right\}, \tag{1.142}$$

where Φ denotes the standard normal cumulative distribution function.

For every $f_\theta \in \mathcal{F}_\Theta$, the Kullback-Leibler divergence to the true density is $K\left(f_\theta, f_0\right) = 0$, so the Kullback-Leibler property is satisfied. At the same time, the squared Hellinger distance between f_0 and any density $f \in \mathcal{F}_*$ is $H^2\left(f, f_0\right) = 2 - \sqrt{2}$ and Barron *et al.* (1999) prove that

$$\limsup_{n\to\infty} \Pi^n(\mathcal{F}_*) = 1 \quad \text{a.s.} \tag{1.143}$$

Therefore, the model is not strongly consistent. In Chapter 5 we illustrate this lack of consistency via MCMC estimation of the Hellinger distance between the true density generating the data and the estimated predictive density.

Chapter 2

Discretely Observed Diffusions

Consider a discretely observed diffusion model defined by an SDE

$$\mathrm{d}Y_t = \alpha_\theta(Y_t)\mathrm{d}t + \mathrm{d}W_t. \tag{2.1}$$

In general, the transition function

$$f_t(y_t|y_0,\theta) \tag{2.2}$$

is intractable, as explained in Section 1.1.2.1. However, the exact simulation algorithm presented in Section 1.2.3, defines auxiliary variables which result in the latent expression (1.121). In the present Chapter, we show how this latent expression can be viewed as a particular case of the general auxiliary variable scheme described in the Introduction, and which constitutes the object of study of the present work.

We propose an alternative MCMC algorithm, based on the complete latent model, which can be used both for simulation and for Bayesian inference. Since the model is the same and no approximation is used, apart from the usual Monte Carlo error, the results obtained in this manner are equivalent to those obtained via the original exact simulation method. Our algorithm, however, is not based on the simultaneous acceptance or rejection of complete sample paths and is therefore equally applicable, regardless of the length of the time interval $[0,t]$ under consideration.

2.1 The Latent Model

We begin by considering the likelihood function for the diffusion model, given a sample $y_{0:n} = (y_{t_1}, \ldots, y_{t_n})$, with known observation times $0 = t_0 < t_1 < \ldots < t_n \leq T$, i.e.

$$f(y_{0:n}|\theta) = \prod_{i=1}^{n} f_{\Delta_i}(y_{t_i}|y_{t_{i-1}}, \theta), \quad \Delta_i = t_i - t_{i-1}, \tag{2.3}$$

where y_0 is considered to be fixed. The Bayesian model is completed by the definition of a prior Π on the parameter space Θ.

We assume throughout this chapter that the model satisfies all the conditions required for the application of the EA(1) algorithm presented in Section 1.2.3. In other words, for every $\theta \in \Theta$, a weak solution to the SDE (2.1) can be constructed through the application of the Girsanov-Carmeron-Martin change of measure formula (Theorem 3). Furthermore, the drift coefficient α_θ is continuously differentiable and integrable, with a tractable expression for the antiderivative

$$A_\theta(u) = \int \alpha_\theta(u)\mathrm{d}u; \tag{2.4}$$

tractable functions $l : \Theta \to \mathbb{R}$ and $r : \Theta \to (0, \infty)$ can be found such that

$$l(\theta) \leq \inf_{u \in \mathbb{R}} \left\{ \left[\alpha_\theta^2(u) + \alpha_\theta'(u)\right]/2 \right\}; \tag{2.5}$$

$$r(\theta) \geq \sup_{u \in \mathbb{R}} \left\{ \left[\alpha_\theta^2(u) + \alpha_\theta'(u)\right]/2 - l(\theta) \right\}. \tag{2.6}$$

As before, we define a bounded function $\varphi_\theta : \mathbb{R} \to [0, 1]$ given by

$$\varphi_\theta(u) = \frac{1}{r(\theta)} \left(\frac{\alpha_\theta^2(u) + \alpha_\theta'(u)}{2} - l(\theta) \right). \tag{2.7}$$

In this case, the transition densities for the diffusion process admit the representation given in equation (1.120), which we write here as

$$f_{\Delta_i}(y_{t_i}|y_{t_{i-1}}, \theta) = g_i(y_{t_i}, y_{t_{i-1}}, \theta) h_i(y_{t_i}, y_{t_{i-1}}, \theta), \tag{2.8}$$

where

$$g_i(y_{t_i}, y_{t_{i-1}}, \theta) = \mathrm{N}(y_{t_i}|y_{t_{i-1}}, \Delta_i) \exp\left\{ A_\theta(y_{t_i}) - A_\theta(y_{t_{i-1}}) - \Delta_i \left[l(\theta) + r(\theta) \right] \right\}; \tag{2.9}$$

$$h_i(y_{t_i}, y_{t_{i-1}}, \theta) = \mathbb{E}_{\mathbb{W}^{y_{t_{i-1}}}} \left[\exp\left\{ r(\theta) \int_{t_{i-1}}^{t_i} \left[1 - \varphi_\theta(y_s)\right] \mathrm{d}s \right\} \middle| y_{t_i} \right]. \tag{2.10}$$

The self-similarity of Brownian Motion ensures the Weiner measure $\mathbb{W}^{y_{t_{i-1}}}$ is well defined on the set $\mathcal{C}_{[t_{i-1},t_i]}$, as the measure induced on $\mathcal{C}_{[0,\Delta_i]}$ by a Brownian motion W started at $W_0 = y_{t_{i-1}}$.

Notice that the Markov property of the diffusion process guarantees that, for each $i = 1, \ldots, n$, the conditional density for y_{t_i} given the previous $y_{t_0}, \ldots, y_{t_{i-1}}$, is given by $f_{\Delta_i}(y_{t_i}|y_{t_{i-1}}, \theta)$. Therefore, equation (2.7) corresponds with expression (4), the starting point for the latent variable expansion presented in the Introduction of the thesis, for a general Bayesian model. Moreover, if we define an infinite dimensional variable $\lambda_i = \{y_s : 0 < s < \Delta_i\} \in \mathcal{C}_{(0,\Delta_i)}$, the complete path between the two consecutive observations y_{i-1}, y_i, with reference measure ν induced by the Brownian motion conditional on $W_0 = y_{t_{i-1}}$ and $W_{\Delta_i} = y_{t_i}$. In other words, a Brownian bridge measure. It then becomes evident that equation (2.10) has the same form of expression (7), namely

$$h_i(y_i, y_{i-1}, \theta) = \int \exp\{r(\theta) b_i(y_{t_i} y_{t_{i-1}}, \theta, \lambda)\} \mathrm{d}\nu(\lambda), \tag{2.11}$$

and

$$b_i(y_{t_i} y_{t_{i-1}}, \theta, \lambda) = b_i(y_{t_i \leq s \leq t_{i-1}}, \theta) = \int_{t_{i-1}}^{t_i} [1 - \varphi_\theta(y_s)] \mathrm{d}s. \tag{2.12}$$

The latent model construction of Beskos *et al.* (2006b) proceeds from here by introducing an auxiliary Poisson process to aid in the estimation of the intractable integral h_i. We argue that the latent variable k can be alternatively be derived from the known series expansion for the exponential function,

$$\exp(rb) = \sum_{k=0}^{\infty} \frac{(rb)^k}{k!}, \tag{2.13}$$

from which it follows that

$$\begin{aligned} h_i(y_{t_i}, y_{t_{i-1}}, \theta) &= \mathbb{E}_{\mathbb{W}^{y_{t_{i-1}}}} \left[\sum_{k_i=0}^{\infty} \frac{[r(\theta)]^{k_i}}{k_i!} \left(\int_{t_{i-1}}^{t_i} [1 - \varphi_\theta(y_s)] \mathrm{d}s \right)^{k_i} \middle| y_{t_i} \right] \\ &= \sum_{k_i=0}^{\infty} \frac{[r(\theta)]^{k_i}}{k_i!} \mathbb{E}_{\mathbb{W}^{y_{t_{i-1}}}} \left[\left(\int_{t_{i-1}}^{t_i} [1 - \varphi_\theta(y_s)] \mathrm{d}s \right)^{k_i} \middle| y_{t_i} \right]. \end{aligned} \tag{2.14}$$

Thus, we arrive at the general expression

$$h_i(y_{t_i}, y_{t_{i-1}}, \theta) = \sum_{k_i=0}^{\infty} c_{i,k_i}(\theta) h_{i,k_i}(y_{1:i}, \theta), \tag{2.15}$$

which characterizes our method.

In this case, $c_{i,k_i}(\theta) = \big[r(\theta)\big]^{k_i}/k_i!$ and

$$h_{i,k_i}(y_{1:i}, \theta) = h_{i,k_i}(y_{t_i}, y_{t_{i-1}}, \theta) = \mathbb{E}_{\mathbb{W}^{y_{t_{i-1}}}}\left[\left(\int_{t_{i-1}}^{t_i}\big[1-\varphi_\theta(y_s)\big]\mathrm{d}s\right)^{k_i}\Big| y_{t_i}\right]. \tag{2.16}$$

We can then replace the k_i-th power with a product,

$$\begin{aligned} b_i(y_{t_i}, y_{t_{i-1}}, \theta, \lambda) &= \left(\int_{t_{i-1}}^{t_i}\big[1-\varphi_\theta(y_s)\big]\mathrm{d}s\right)^{k_i} = \prod_{l=1}^{k_i}\int_{t_{i-1}}^{t_i}\big[1-\varphi_\theta(y_{\tau_{i,l}})\big]\mathrm{d}(\tau_{i,l}) \\ &= \int_{t_{i-1}}^{t_i}\cdots\int_{t_{i-1}}^{t_i}\prod_{l=1}^{k_i}\big[1-\varphi_\theta(y_{\tau_{i,l}})\big]\mathrm{d}(\tau_{i,1})\ldots\mathrm{d}(\tau_{i,k_i}) = b_i(y_{t_i}, y_{t_{i-1}}, \theta, s_{i,1:k_i}), \end{aligned} \tag{2.17}$$

where $s_{i,l} = (\tau_{i,l}, y_{\tau_{i,l}})$. This expression depends only on the values of the path $y_{\tau_1}, \ldots, y_{\tau_{k_i}}$ and not on the values between them. Therefore, we may write

$$\begin{aligned} h_{i,k_i}(y_i, y_{i-1}, \theta) =& \mathbb{E}_{\mathbb{W}^{y_{t_{i-1}}}}\left[b_i(y_{t_i} y_{t_{i-1}}, \theta, \lambda)\big| y_{t_i}\right] \\ =& \mathbb{E}_{\mathbb{W}^{y_{t_{i-1}}}}\left[b_i(y_{t_i} y_{t_{i-1}}, \theta, s_{i,1:k_i})\big| y_{t_i}\right] \\ =& \int_{\mathbb{S}_i^{k_i}}\prod_{l=1}^{k_i} b_{i,l}(y_i, y_{i-1}, \theta, s_{i,1:k_i})\mathrm{d}\nu(s_{i,1:k_i}), \end{aligned} \tag{2.18}$$

where $\mathbb{S}_i^{k_i} = [t_{i-1}, t_i] \times \mathbb{R}$, with reference measure ν given by the product of the k_i-fold product Lebesgue measure on $[t_{i-1}, t_i]$ and the k_i-dimensional distribution of the Weiner measure $\mathbb{W}^{y_{t_{i-1}}}$ on $\mathcal{C}[t_{i-1}, t_i]$, conditional on $W_{t_i} = y_{t_i}$ for every $i = 1, \ldots, n$. It is, however, more convenient to revert to Lebesgue measure, since the finite-dimensional densities of Brownian motions are known multivariate normal densities. So we write

$$h_{i,k_i}(y_i, y_{i-1}, \theta) = \int_{\mathbb{S}^{k_i}}\prod_{l=1}^{k_i} b_{i,l}(y_i y_{i-1}, \theta, s_{i,1:k_i})\mathrm{d}\nu(s_{i,1:k}), \tag{2.19}$$

where

$$b_{i,l}(y_i, y_{i-1}, \theta, s_{i,1:k_i}) = \mathrm{N}\big(x_{i,(l)} | x_{i,(l-1)}, \tau_{i,(l)} - \tau_{i,(l-1)}\big)\big[1 - \varphi_\theta(x_{i,l})\big], \tag{2.20}$$

depends on the observations only through the convention $x_{i,(0)} = y_{t_i}$. The latent variables $s_{i,l} = (\tau_{i,l}, x_{i,l})$ take values on $\mathbb{S} = [0, \Delta_i] \times \mathbb{R}$ and ν is the corresponding Lebesgue measure. The notation (l) for the subindices in the normal density functions represents a permutation of the $\tau_{i,1:k_i}$ such that $0 < \tau_{i,(l)} < \ldots < \tau_{i,(k_i)}$ and is simply an aid to factorize the multivariate normal density into univariate normal densities. This is the notation we use throughout the remaining of this Chapter. We have replaced $y_{\tau_{i,l}}$ with $x_{i,l}$ to emphasize the fact that we are dealing here with auxiliary variables, as opposed to observations, denoted by y_{t_i}.

Through the above construction, we have arrived at a latent model for discretely observed diffusions in the form of the general latent likelihood (15) anticipated in the Introduction to this thesis, namely

$$f(y_{1:n}, k_{1:n}, s_{1:n,1:\infty} | \theta) = g(y_{1:n}, \theta) \prod_{i=1}^{n} c_{i,k_i}(\theta) \left(\prod_{l=1}^{k_i} b_{i,l}(y_{1:i}, \theta, s_{i,1:k_i}) \right) \left(\prod_{l>k_i} f(s_{i,l}) \right), \tag{2.21}$$

where

$$g(y_{1:n}, \theta) = \exp\big\{ A_\theta(y_t) - A_\theta(y_0) - t\big[l(\theta) + r(\theta)\big] \big\} \prod_{i=1}^{n} \mathrm{N}(y_{t_i} | y_{t_{i-1}}, \Delta_i); \tag{2.22}$$

$$c_{i,k_i}(\theta) = \frac{\big[r(\theta)\big]^{k_i}}{k_i!}; \tag{2.23}$$

the functions $b_{i,l}$ are given by equation (2.20); and $f(s_{i,l})$ denotes any fully known density function on $[t_{i-1}, t_i] \times \mathbb{R}$. In the next section, we present a convenient choice for MCMC simulation.

This is the latent model induced by the latent variable construction proposed by Beskos *et al.* (2006a) for Bayesian inference using their exact simulation algorithm. The original likelihood (2.3) can be recovered from the latent expression, by integrating out all auxiliary variables. However, the latent likelihood expression can be presented using a more compact notation, in which the latent variables are not indexed by i. This alternative representation is better suited for

the alternative MCMC algorithm for Bayesian inference we present in the next Section.

Let $k = \sum_i k_i$, and $s_{1:\infty} = \bigcup_{i,l}\{s_{i,l}\}$. Then, for each $l \geq 1$, $s_l = (\tau_l, x_l)$ takes values in $[0, t] \times \mathbb{R}$. In order to write the double product

$$\prod_{i=1}^{n}\prod_{l=1}^{k_i} \mathrm{N}\big(x_{i,(l)} | x_{i,(l-1)}, \tau_{i,(l)} - \tau_{i,(l-1)}\big) \tag{2.24}$$

as a single product, in terms of the new indices for the latent variables, we need to account for the clustering structure induced by the observation times, which is relevant, since $x_{i,0} = y_{t_{i-1}}$ for every i. We do this by introducing new notation.

Let $\tilde{T}_k = (\tau_l)_{l=1}^{k} \cup (t_i)_{i=1}^{n}$. For each $l = 1 \ldots . k$, let $\tilde{\tau}_{l-1}$ and $\tilde{\tau}_{l+1}$ be the skeleton times or observation times immediately to the left and right of τ_l. In other words

$$\tilde{\tau}_{l-1} = \max\big\{\tilde{\tau} \in \tilde{T}_j : \tau < \tau_l\big\}; \quad \tilde{\tau}_{l+1} = \min\big\{\tilde{\tau} \in \tilde{T}_k : \tau_l < \tau\big\}. \tag{2.25}$$

Denote by $\tilde{x}_l$ the point corresponding to a time $\tilde{\tau}_l$, i.e.

$$\tilde{x}_l = \begin{cases} x_j & \text{if } \tilde{\tau}_l = \tau_j \\ y_i & \text{if } \tilde{\tau}_l = t_i \end{cases} \tag{2.26}$$

and notice that, for the ordered skeleton, the point $\tilde{\tau}_{(l-1)}$ to the left of $\tau_{(l)}$ is either some observation time t_i or the previous skeleton time $\tau_{(l-1)}$. Analogously, $\tilde{\tau}_{(l+1)}$ may be equal to $\tau_{(l+1)}$ or some observation time.

With this new notation, we can rewrite

$$\prod_{i=1}^{n}\prod_{l=1}^{k_i} \mathrm{N}\big(x_{i,(l)} | x_{i,(l-1)}, \tau_{i,(l)} - \tau_{i,(l-1)}\big) = \frac{\prod_{i=1}^{n} k_i!}{k!} \prod_{l=1}^{k} \big(\tilde{x}_{(l)} | \tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big), \tag{2.27}$$

where the factorial terms account for the arbitrary reindexation of the latent variables, and how that affects the ordering of the τ_l values into the different $[t_{i-1}, t_i]$ time intervals between consecutive observations. Thus, we arrive at the compact form of the latent model which we use throughout the rest of this chapter.

$$f(y_{1:n}, k, s_{1:\infty} | \theta) = g(y_{1:n}, \theta) c_k(\theta) \left(\prod_{l=1}^{k} b_l(y_{1:n}, \theta, s_{1:k})\right) \left(\prod_{l>k} f(s_l)\right) \tag{2.28}$$

where $g(y_{1:n}, \theta)$ is defined by equation (2.22), $c_k(\theta) = \big[r(\theta)\big]^k/k!$ and

$$b_l(y_{1:n}, \theta, s_{1:k}) = \big[1 - \varphi_\theta(x_l)\big]\mathrm{N}\big(\tilde{x}_{(l)}|\tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big). \tag{2.29}$$

In the next section, we show how a fully MCMC based Bayesian inference method is possible for this model.

2.2 MCMC Simulation and Posterior Inference

We propose the use of a Gibbs sampling algorithm to produce a sample from the latent model defined by the extended likelihood (2.28) and the prior distribution Π over the parameter space Θ. In other words, we propose a sampling scheme in which each of the latent variables, $k, s_{1:\infty}$ and the parameter θ are updated by drawing samples from their full conditional distributions. In some cases, direct sampling is not possible, therefore we use a hybrid method in which some of the updates are done through a Metropolis-Hastings step.

Recall that each latent variable $s_l = (\tau_l, x_l)$ can be decomposed into a time τ_l and a point x_l, and, given k, the set $s_{1:k}$ is called the skeleton, following the terminology introduced by Beskos *et al.* (2006b). The skeleton points and times are updated separately. Notice that, conditional on k, the values s_l for $l > k$ i.i.d from a known density and, more importantly, they do not appear in the full conditional density expression for the rest of the variables. Therefore, at any step of the algorithm, only a finite number of variables needs to be recorded and updated. In fact, we use the additional variables only to represent the fully extended model space proposed by Godsill's 2001 (Section 1.2.2.3 of the present work) to deal with the sampling of variables with random dimension.

The algorithm begins by initializing the necessary variables. A sensible way to do this is by initially making $k = 0$, so that no other latent variable initialization is needed. The initial value for the parameter θ can be chosen in the usual way, by simply fixing a value or drawing one from the prior distribution Π.

2.2.1 Updating the Skeleton Size, k

The latent model can be interpreted as a family of fully nested models $\{f_k\}_{k\geq 0}$, indexed by k, where

$$f_0(y_{1:n}|\theta) \propto g(y_{1:n}, \theta), \tag{2.30}$$

and for each for each $k \geq 1$

$$f_k(y_{1:n}|s_{1:k}, \theta) \propto g(y_{1:n}, \theta) \left(\prod_{l=1}^{k} b_l(y_{1:n}, \theta, s_{1:k}) \right). \tag{2.31}$$

Therefore, following Godsill (2001), we extend the sampling space to include the complete set of variables $s_{l:\infty}$ and the index k for the model in the MCMC simulation scheme. We update the model index k through a Metropolis-Hastings step with proposal distribution given by

$$q(k', s'_{1:k'}|s_{1:k}, \theta, y_{1:n}) = q(k'|k) q(s'_{1:k'}|s_{1:k}, \theta, y_{1:n}) \left(\prod_{l>k'} f(s_l) \right), \tag{2.32}$$

where

$$q(k'|k) = \begin{cases} p & \text{if } k' = k+1 \\ 1-p & \text{if } k' = k-1 \\ 0 & \text{otherwise,} \end{cases} \tag{2.33}$$

for some $0 < p < 1$, and

$$q(s'_{1:k'}|\theta_{1:k}) = \begin{cases} q(\tau'_{k+1}) q(x'_{k+1}|\tau'_{k+1}, s_{1:k}, \theta, y_{1:n}) \mathbf{1}_{s_{1:k}}(s'_{1:k}) & \text{if } k' = k+1 \\ \mathbf{1}_{s_{1:k'}}(s'_{1:k'}) & \text{if } k' = k-1 \\ 0 & \text{otherwise,} \end{cases} \tag{2.34}$$

In other words, the only possible changes for k are to $k+1$ or $k-1$. If a move down is proposed, the skeleton has to be adjusted by dropping the last point s_k, while the rest remain the same. If a move up is proposed, the skeleton is augmented with a new proposed skeleton point $s_{k+1} = (\tau_{k+1}, x_{k+1})$, drawn from a proposal distribution $q(\tau'_{k+1}) q(x'_{k+1}|\tau'_{k+1} s_{1:k}, \theta, y_{1:n})$. At this point any choice would lead to a valid MCMC chain with the desired equilibrium distribution, as long as the acceptance probability is given by

$$\alpha = \min \left\{ 1, \frac{(1-p)}{p} \frac{f(k+1, s_{1:k+1}|\theta, y_{1:n})}{q(s_{k+1}|s_{1:k}, \theta, y_{1:n}) f(k, s_{1:k}|\theta, y_{1:n})} \right\}, \tag{2.35}$$

when $k' = k+1$; or, when $k' = k-1$, by

$$\alpha = \min\left\{1, \frac{p}{(1-p)} \frac{q(s_k|s_{1:k-1}, \theta, y_{1:n}) f(k-1, s_{1:k-1}|\theta, y_{1:n})}{f(k, s_{1:k}|\theta, y_{1:n})}\right\}. \tag{2.36}$$

However, many expensive calculations can be avoided by a suitable choice of proposal distribution, which we now present.

First, we let $q(\tau_{k+1}) = \mathrm{U}(\cdot|0,T)$, so the new time is generated uniformly over the complete time interval under consideration. Before we determine the proposal distribution for the new skeleton point, x_{k+1}, recall that

$$\begin{aligned} f(k+1, s_{1:k+1}|\theta, y_{1:n}) \propto & \frac{[r(\theta)]^{k+1}}{(k+1)!} \prod_{l=1}^{k+1} b_l(y_{1:n}, \theta, s_{1:k}) \\ = & \frac{[r(\theta)]^{k+1}}{(k+1)!} \prod_{l=1}^{k+1} [1 - \varphi_\theta(x_l)] \mathrm{N}\big(\tilde{x}_{(l)}|\tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big). \end{aligned} \tag{2.37}$$

The product of normal terms in this expression corresponds to the $k+1$-dimensional distribution of a Brownian motion path at times $\tau_{1:k+1}$, conditioned to pass through every observation y_{t_i} at time t_i, and evaluated at $x_{1:k+1}$. Recall that $\tilde{\tau}_k$ and $\tilde{\tau}_{k+2}$ denote the times immediately to the left and right of τ_{k+1}, respectively, so that $\tilde{\tau}_k < \tau_{k+1} < \tilde{\tau}_{k+2}$; while $\tilde{x}_k$ and $\tilde{x}_{k+2}$ denote their associated points. The $(k+1)$-dimensional distribution for the Brownian motion path can be factorized as the product of the k-dimensional distribution at times $\tau_{1:k}$ and the conditional distribution for the state of the process at time τ_{k+1} given all others. Formally,

$$\begin{aligned} \prod_{l=1}^{k+1} \mathrm{N}\big(\tilde{x}_{(l)}|\tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big) = & \mathrm{N}(x_{k+1}|\mu_{k+1}, \sigma^2_{k+1}) \\ & \prod_{l=1}^{k} \mathrm{N}\big(\tilde{x}_{(l)}|\tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big), \end{aligned} \tag{2.38}$$

where the subindex (l) represent the ordering of $\tau_{1:k+1}$ on the left hand side expression, and the ordering of $\tau_{1:k}$ on the right side product. The mean and

variance of the conditional normal distribution for τ_{k+1} are given by

$$\mu_{k+1} = \tilde{x}_k \left(1 - \frac{\tau_{k+1} - \tilde{\tau}_k}{\tilde{\tau}_{k+2} - \tilde{\tau}_k}\right) + \tilde{x}_{k+2} \frac{\tau_{k+1} - \tilde{\tau}_k}{\tilde{\tau}_{k+2} - \tilde{\tau}_k}; \tag{2.39}$$

$$\sigma^2_{k+1} = \frac{(\tau_{k+1} - \tilde{\tau}_k)(\tilde{\tau}_{k+2} - \tau_{k+1})}{\tilde{\tau}_{k+2} - \tilde{\tau}_k}. \tag{2.40}$$

We are now ready to define the proposal distribution

$$q(x_{k+1}|\tau_{k+1}, s_{1:k}, \theta, y_{1:n}) = \mathrm{N}(x_{k+1}|\mu_{k+1}, \sigma^2_{k+1}), \tag{2.41}$$

the use of which simplifies the calculation of the acceptance probability for the Metropolis-Hastings step to

$$\alpha = \min\left\{1, \frac{(1-p)}{p} \frac{T\, r(\theta)}{k+1} \left[1 - \varphi_\theta(x_{k+1})\right]\right\}, \quad \text{when } k' = k+1; \tag{2.42}$$

$$\alpha = \min\left\{1, \frac{p}{(1-p)} \frac{k}{T\, r(\theta)} \left[1 - \varphi_\theta(x_k)\right]^{-1}\right\}, \quad \text{when } k' = k-1; \tag{2.43}$$

2.2.2 Updating the Skeleton Times, $\tau_{1:k}$

For the skeleton times $\tau_{1:k}$, the full conditional distribution is given by

$$f(\tau_{1:k}|x_{1:k}, y_{1:n}) \propto \prod_{l=1}^{k} \mathrm{N}\big(\tilde{x}_{(l)}|\tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big)\mathbf{1}\{0 < \tau_l < T\}. \tag{2.44}$$

It is difficult, from this expression to derive an update scheme for the skeleton times, since each $\tilde{\tau}_{(l-1)}$ may be a skeleton time or an observation time. Therefore, we use again the properties of the multivariate normal distribution to rearrange this product. For each l, we write the k-variate normal density represented by this product, into the $k-1$-variate normal and the univariate conditional for x_l given all the other variables, as we did in the previous section.

Thus, we update each τ_l from the full conditional distribution

$$f(\tau_l|\tau_{-l}, x_{1:k}, y_{1:n}) \propto \mathrm{N}(x_l|\mu_l, \sigma^2_l)\mathbf{1}\{\tilde{\tau}_{l-1} < \tau_l < \tilde{\tau}_{l+1}\}, \tag{2.45}$$

where

$$\mu_l = \tilde{x}_{l-1} \left(1 - \frac{\tau_l - \tilde{\tau}_{l-1}}{\tilde{\tau}_{l+1} - \tilde{\tau}_{l-1}}\right) + \tilde{x}_{l+1} \frac{\tau_l - \tilde{\tau}_{l-1}}{\tilde{\tau}_{l+1} - \tilde{\tau}_{l-1}}; \tag{2.46}$$

$$\sigma^2_l = \frac{(\tau_l - \tilde{\tau}_{l-1})(\tilde{\tau}_{l+1} - \tau_l)}{\tilde{\tau}_{l+1} - \tilde{\tau}_{l-1}}. \tag{2.47}$$

Conditional on everything else, each τ_l appears only in the normal distribution, evaluated at x_l, corresponding to a Brownian bridge connecting the closest points to the right and left of τ_l.

As a function of τ_l, the above expression does not resemble any known density. Therefore, we update each τ_l using a Metropolis-Hastings step, with uniform proposal distribution

$$q(\tau_l|\tau_{-l}) = \mathrm{U}(\tau_l|\tilde{\tau}_{l-1}, \tilde{\tau}_{l+1}). \tag{2.48}$$

The calculation of the acceptance probability requires only the evaluation of a ratio of normal density functions.

2.2.3 Updating the Skeleton Points, $\mathbf{x_{1:k}}$

The full conditional distribution for the skeleton points $x_{1:k}$, is given by

$$f(x_{1:k}|\tau_{1:k}, \theta, y_{1:n}) \propto \prod_{l=1}^{k} \big[1 - \varphi_\theta(x_l)\big] \mathrm{N}\big(\tilde{x}_{(l)}|\tilde{x}_{(l-1)}, \tilde{\tau}_{(l)} - \tilde{\tau}_{(l-1)}\big). \tag{2.49}$$

We can use the same factorization as in the above section to update each x_l from the full conditional distribution

$$f(x_l|x_{-l}, \tau_{1:k}, y_{1:n}) \propto \big[1 - \varphi_\theta(x_l)\big] \mathrm{N}(x_l|\mu_l, \sigma_l^2), \tag{2.50}$$

where μ_l and σ_l^2 are given by (2.46) and (2.47) respectively.

Since $0 < \big[1 - \varphi_\theta(x_l)\big] < 1$, a simple rejection algorithm can be implemented for this update, by generating the new x_l from the normal distribution and accepting it with probability $\big[1 - \varphi_\theta(x_l)\big]$.

Note that, up to this point, the value of the parameter is fixed. Therefore, without the need for a prior, Π or for additional updating steps, this algorithm can be used for exact simulation of a diffusion bridge on a time interval $[0, T]$, with fixed end point y_T. It is enough to consider a single observation at time $\iota_n = T$ in all of the update steps described above.

A diffusion path with free end point can also be simulated in this manner, by adding an extra update step for the end point.

2.2.4 Updating the End Point, y_T

If the algorithm is being used to simulate observations from a diffusion path on $[0, T]$ and the end point is now known, y_T must also be part of the MCMC scheme. In this case, it must be updated from the full conditional distribution

$$f(y_T|k, s_{1:k}, \theta, y_{0:n}) \propto \exp\{A_\theta(y_T)\}\mathrm{N}(y_T|\tilde{x}_m, T - \tilde{\tau}_m), \tag{2.51}$$

where $\tilde{\tau}_m = \max\{\tilde{\tau} \in \tilde{T}\}$ is the maximum of the observation and skeleton times, and $\tilde{x}_m$ is the corresponding observation or skeleton point.

The method used to sample from this distribution depends on the specific shape of the A_θ function. However, the form of the density suggests that a rejection algorithm or a MH step with normal proposal distribution might be a good choice in many cases. In Section 2.3 we present two illustrations for which we use a rejection sampler for the update of the end point y_T.

When the algorithm is being used for posterior simulation, this step may still be useful. If we make $T > t_n$, the time of the last observation, this provides a sample from the predictive distribution at time T. Furthermore, the sample obtained for the skeleton would include the complete interval $[0, T]$, so that observations from the predictive distribution at any time $t \in [t_n, T]$ can be obtained by Brownian Bridge interpolation between the observations, the skeleton points and the final point y_T.

We now proceed with the final update step required for MCMC posterior simulation from the diffusion model.

2.2.5 Updating the Parameter θ.

Observe that, conditional on the latent variables, the parameter is independent of the data, with full conditional density

$$\begin{aligned} f(\theta|k, s_{1:\infty}) = f(\theta|k, x_{1:k}) \propto \Pi(\theta) \exp\left\{A_\theta(y_t) - A_\theta(y_0) - t\big[l(\theta) + r(\theta)\big]\right\} \\ \big[r(\theta)\big]^k \prod_{l=1}^{k} \big[1 - \varphi_\theta(x_l)\big]. \end{aligned} \tag{2.52}$$

Clearly, no general method can be suggested to simulate from this density, since it depends on the shapes of the functions A, r, l and φ. In the next Section, we consider two examples, for which the parameter space Θ is a bounded interval $[a, b] \in \mathbb{R}$. We therefore use a Metropolis-Hastings step with uniform proposal distribution $q = \mathrm{U}(\cdot|a, b)$. Other proposal distributions may be explored, which depend on the conditioning variables $k, x_{1:k}$, but for the concrete examples we study it is not clear what a better choice would be, and the uniform proposal seems to work well.

2.3 Illustrations

In this Section we illustrate our methodology with two concrete examples of real valued diffusion processes. In each case, we generate a sample from the true model and use our algorithm to perform Bayesian inference. We compare our results with those obtained using the original exact simulation algorithm of Beskos *et al.* (2006b).

2.3.1 Example 1: Sine Diffusion

In this example we consider the diffusion process defined by the SDE

$$\mathrm{d}Y_t = \sin(Y_t - \theta)\mathrm{d}t + \mathrm{d}W_t, \tag{2.53}$$

so the drift coefficient is given by

$$\alpha_\theta(y) = \sin(y - \theta). \tag{2.54}$$

When $\theta \in \Theta = [0, 2\pi)$, the SDE has a unique solution Y to which we refer as the sine diffusion.

In this case,

$$A_\theta(y) = \int_0^y \alpha_\theta(x)\mathrm{d}x = -\cos(y - \theta), \tag{2.55}$$

therefore $\int_{\mathbb{R}} A_\theta(y)\mathrm{d}y$ is not defined, which means the process does not have a stationary density.

For each $\theta \in \Theta$, consider the function

$$\xi_\theta(y) = \alpha_\theta^2(y) + \alpha_\theta'(y) = \sin^2(y-\theta) + \cos(y-\theta) \tag{2.56}$$

and notice that $\xi_\theta(y) \geq -1$ for all $y \in \mathbb{R}$ and $\theta \in \Theta$. Therefore, we define

$$l = l(\theta) = \inf_{y \in \mathbb{R}} \left\{ \frac{\xi_\theta(y)}{2} \right\} = -\frac{1}{2}. \tag{2.57}$$

Also, $\xi_y'(y) = \sin(y-\theta)\big[2\cos(y-\theta) - 1\big] = 0$ when $\sin(y-\theta) = 0$, in which case $|\cos(y-\theta)| = 1$; or when $\cos(y-\theta) = 1/2$, in which case $|\sin(y-\theta)| = \sqrt{3}/2$. Furthermore,

$$\xi_\theta''(y) = -2\sin^2(y-\theta) + \cos(y-\theta)\big[2\cos(y-\theta) - 1\big]. \tag{2.58}$$

So, when $\sin(y-\theta) = 0$ and $\cos(y-\theta) = 1$, the function $\xi_\theta(y)$ has an inflexion point; when $\sin(y-\theta) = 0$ and $\cos(y-\theta) = -1$, the function has a local minimum; and it has a local maximum when $\cos(y-\theta) = 1/2$ and $|\sin(y-\theta)| = \sqrt{3}/2$. Therefore the maximum is reached at $\xi_\theta(\pi/3) = 5/4$, and we define

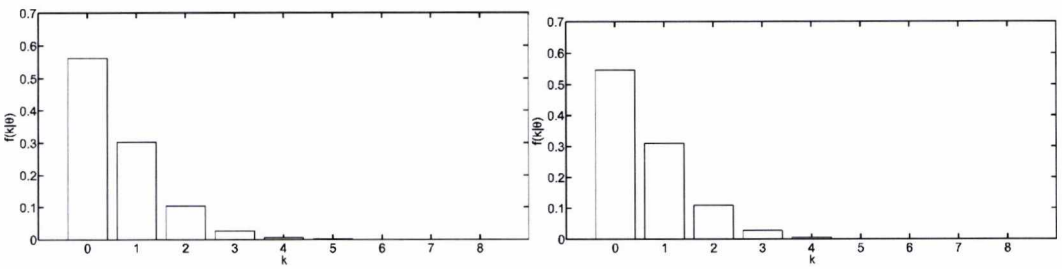
$$r = r(\theta) = \sup_{y \in \mathbb{R}} \left\{ \frac{\xi_\theta(y)}{2} - l \right\} = \frac{9}{8}. \tag{2.59}$$

Finally, the bounded function used in the transition density expression is given by

$$\varphi_\theta(y) = \frac{8}{9} + \frac{4}{9}\cos(y-\theta)\big[1 - \cos(y-\theta)\big]. \tag{2.60}$$

We first illustrate the use of our algorithm for the simulation of diffusion paths. We fix the parameter at a known value $\theta_0 = 2$ and simulate a total of $N = 1,000,000$ skeletons for the diffusion in the time interval $[0, 1]$. We do the same using the retrospective rejection sampler of Beskos *et al.* (2006b) (EA1) and our MCMC alternative, with parameter $p = 1/2$ for the proposal distribution for the MH step update for k (Section 2.2.1). The choice of $T = 1$ is made to keep the time interval for simulation close to the optimal value of $T = 1/r(\theta) = 9/8$ for the EA1. The large Monte Carlo sample is chosen to allow a comparison of the skeleton times and points produced by each algorithm. If the sample is kept smaller, the number of realizations with a large k value would not be large enough for any interesting comparison.

Figure 2.1 shows histograms of the size k of the skeletons produced by each algorithm. As we can see, the skeleton size distributions are similar, with the one corresponding to the EA1 algorithm showing a slightly heavier left tail and the one corresponding to the MCMC algorithm showing a heavier right tail. However, this can be attributed to the fact that the skeletons produced by the MCMC scheme are correlated, unlike those generated by the EA1 algorithm.

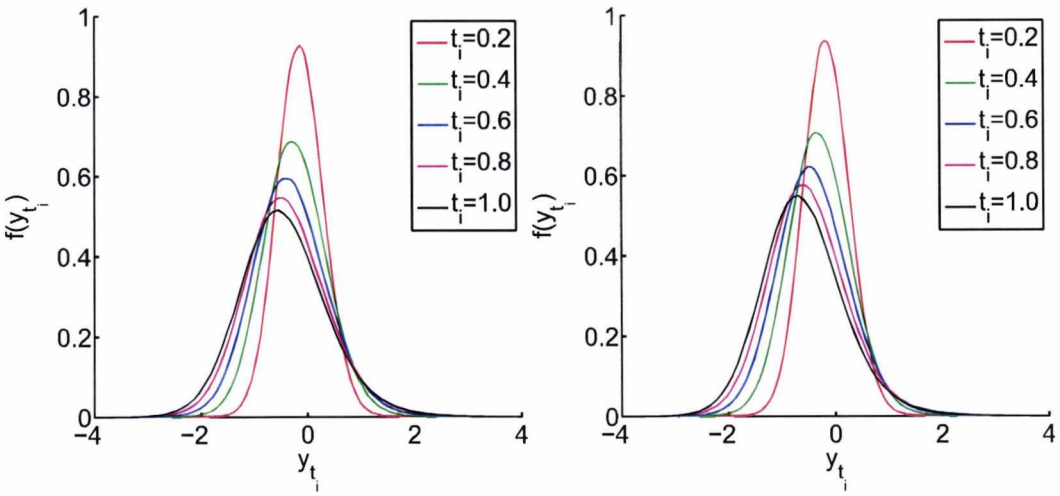


(a) EA1 algorithm (b) MCMC algorithm

Figure 2.1: Histogram of the skeleton size k for the sine diffusion with fixed parameter $\theta_0 = 2$ and initial point $y_0 = 0$, on the time interval $[0, 1]$. The histogram on the left corresponds to the original exact simulation algorithm; the plot on the right corresponds to the MCMC version we propose.

The ultimate goal of the algorithms, when the parameter is fixed and known, is path simulation. We set $l_i \in \{0.2, 0.4, 0.6, 0.8\}$ and simulated the corresponding diffusion points y_{t_i} by Brownian Bridge interpolation between skeleton points, for each of the skeletons obtained from the exact simulation algorithms. This generates, for each i, a sample of size $N = 1,000,000$ for each of the the diffusion points Y_{t_i}. Figure 2.2 shows estimated marginal densities for each one of those points. Once again, we can see the plots are similar, with a smaller variance displayed by the MCMC simulated data, attributable to the correlation in the sample.

The large Monte Carlo sample size of $N = 1,000,000$ allows us to visualize some of the aspects of the skeletons produced by each algorithm. Figure 2.3 illustrates the behaviour of the ordered skeleton times $\tau_{(l)}$, for $l = 1, \ldots, 6$ and their associated points $x_{(l)}$. Once again, we can see the similarity between the plots, with some differences observed for $l = 6$. Notice that only skeleton samples



(a) EA1 algorithm

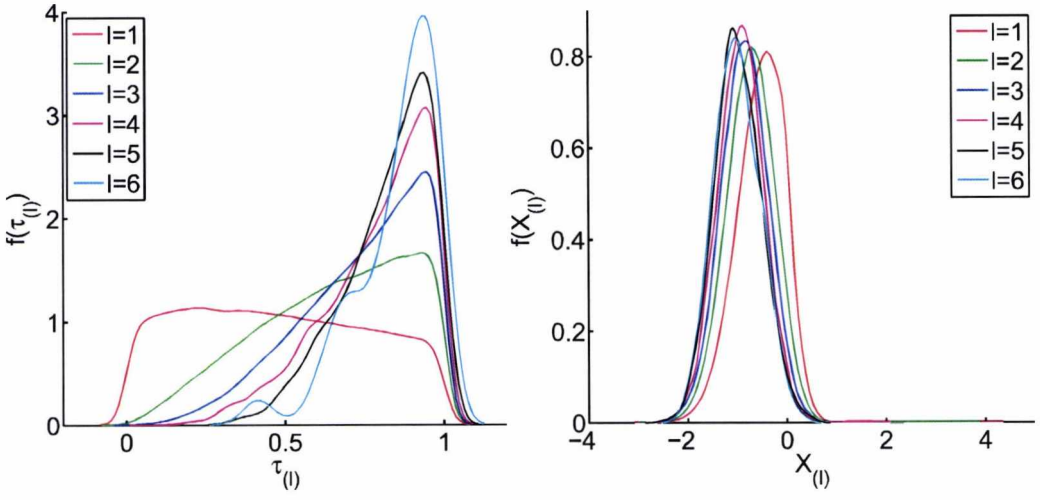
(b) MCMC algorithm

Figure 2.2: Marginal densities of the sine diffusion Y_{t_i} at various times. The plots correspond to smoothed histograms of the data simulated using retrospective rejection sampling (left) and the MCMC approach (right).

with $k \geq 6$ can be used in this case. As we can see from the histograms in Figure 2.1, this is not a common occurrence, so the differences are explained by the small sample sizes.

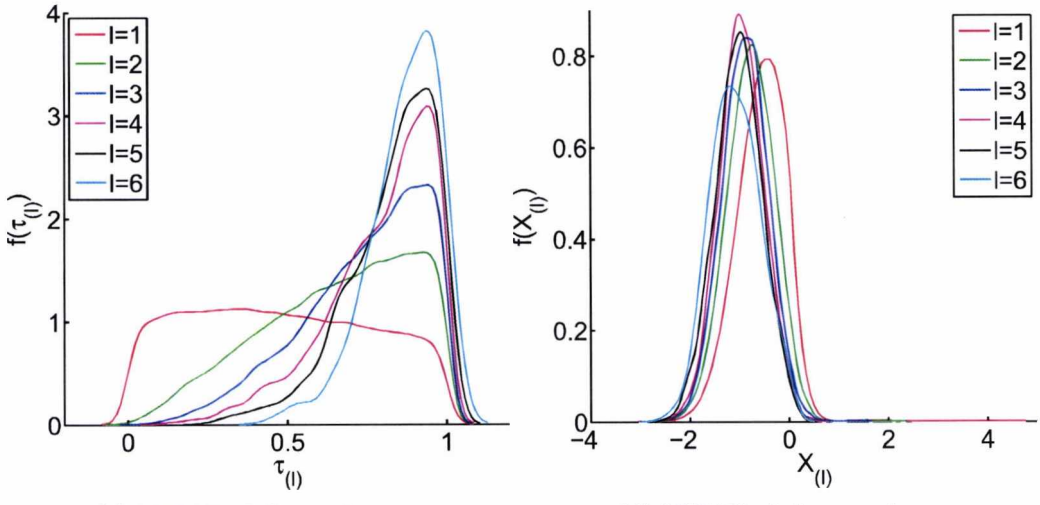
While it seems reasonable to conclude that both algorithms produce equivalent results, it is recommendable to use thinned samples from the MCMC algorithm, in order to reduce the correlation between consecutive states visited by the Markov Chain.

We now proceed to illustrate the use of the MCMC algorithm presented in the previous section, for the purpose of parameter estimation. We produce a sample of what is commonly known as high density data. That is, a high number of observations per time unit. In order to avoid the argument of correlation in the sample induced by the MCMC approach, we produce the data using retrospective rejection sampler. Once again, we fix the true value of the parameter at $\theta = 2$. This time, we generate a single skeleton for the sine diffusion in the time interval $[0, 100]$ and use Brownian bridge interpolation to simulate $10,000$ equally spaced data points, i.e. 100 observations per time unit. Figure 2.4 shows the data and



(a) EA1: skeleton times

(b) EA1: skeleton points



(c) MCMC: skeleton times

(d) MCMC: skeleton points

Figure 2.3: Marginal densities of the first six ordered skeleton times $\tau_{(l)}$ (left) and points $x_{(l)}$ (right), for the sine diffusion. The plots correspond to smoothed histograms of the data simulated using retrospective rejection sampling (above) and the MCMC approach (below).

the skeleton used to produce it.

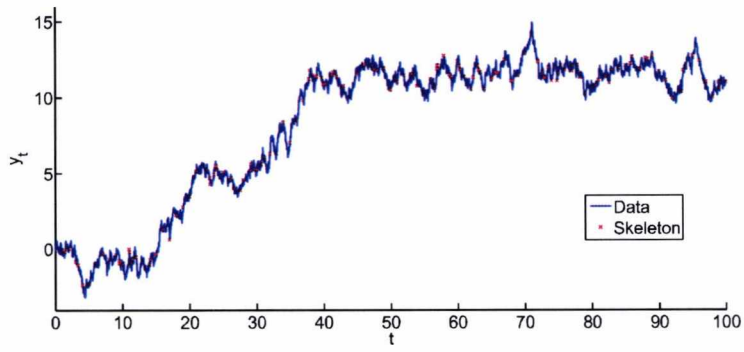
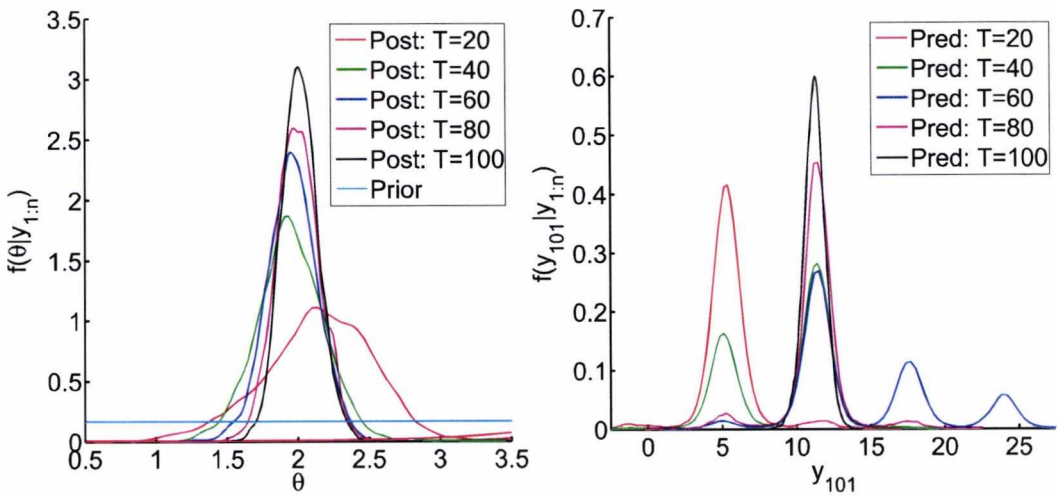


Figure 2.4: 10,000 data points from the sine diffusion in the time interval $[0, 100]$, with parameter $\theta = 2$ and initial point $y_0 = 0$.

We define a uniform prior $\Pi(\theta) = \mathrm{U}(\theta|0, 2\pi)$ on the parameter space and use the MCMC algorithm to produce a sample from the posterior distribution $\Pi^n(\theta)$, for increasing sample sizes. Specifically, we consider the data set consisting of the first $n = 2,000$ data points, in the time interval $[0, 20]$ and produce a posterior sample of size $N = 10,000$ from the MCMC algorithm, with a burning period of 10,000 iterations and a thinning of 1 every 10 iterations for the sample. We repeat the analysis for the time intervals $[0, T]$, $T = 40, 60, 80, 100$, in other words, we increase the sample size by 2,000 points every time.

The estimated posterior densities for the parameter are shown on the left hand side of Figure 2.5. We can see that the posterior mass seems to accumulate around the true value $\theta_0 = 2$ as the sample size n and the limit T of the time interval of observation grow.

The right panel of Figure 2.5 shows the estimated predictive densities for the process y_T at time $T = 101$, for each of the samples. The sine diffusion does not have a stationary density, therefore we don't expect to recover a fixed marginal behaviour. However, as the interval of observations approaches the time of prediction, we can observe the evolution of the predictive distribution. As expected from a regular diffusion process, the variance decreases towards the end, as the point y_{101} is highly correlated to y_{100}, the last data point.



(a) Posterior density

(b) Predictive density

Figure 2.5: Estimated posterior density for the parameter of the sine diffusion (left) and predictive density for the observation at time $T = 101$.

2.3.2 Example 2: Hyperbolic Diffusion

Now, we consider the diffusion process defined by the SDE

$$\mathrm{d}Y_t = \theta \frac{Y_t}{\sqrt{1+Y_t^2}}\mathrm{d}t + \mathrm{d}W_t, \tag{2.61}$$

so the drift coefficient is given by

$$\alpha_\theta(y) = \theta \frac{y}{\sqrt{1+y^2}}. \tag{2.62}$$

We refer to the process Y defined as the weak solution to this SDE, as the Hyperbolic diffusion (see Bibby & Sorensen, 1995). When $\theta < 0$, Y is an ergodic stationary process with invariant density

$$f_\theta(y) \propto \exp\{2A_\theta(y)\}, \tag{2.63}$$

where

$$A_\theta = \int_0^y \alpha_\theta(x)\mathrm{d}x = \theta\sqrt{1+y^2}. \tag{2.64}$$

Observe that $\alpha'_\theta(y) = \theta/(1+y^2)^{3/2}$ and $\alpha^2_\theta(y) = \theta^2 y^2/(1+y^2)$, so we can define

$$\xi_\theta(y) = \alpha^2_\theta(y) + \alpha'_\theta(y) = \frac{\theta^2 y^2}{1+y^2} + \frac{\theta}{(1+y^2)^{3/2}} = \frac{\theta}{1+y^2}\left(\theta y^2 \frac{1}{\sqrt{1+y^2}}\right). \quad (2.65)$$

Therefore

$$\xi'_\theta(y) = \frac{\theta y}{(1+y^2)^2}\left(2\theta - \frac{3}{\sqrt{1+y^2}}\right) \quad (2.66)$$

For $\theta < 0$, we have $2\theta - 3/\sqrt{1+y^2} < 0$, so $\xi'_\theta(y) = 0$ only when $y = 0$. Furthermore, $\xi'_\theta(y) < 0$ when $y < 0$ and $\xi'_\theta(y) > 0$ when $y > 0$, so $\xi_\theta(y)$ has a unique minimum at $\xi_\theta(0) = \theta$. Consequently, we may define

$$l(\theta) = \inf_{y\in\mathbb{R}}\left\{\frac{\xi_\theta(y)}{2}\right\} = \frac{\theta}{2}. \quad (2.67)$$

It is equally straightforward to realize that

$$\lim_{y\to\infty} \xi_\theta(y) = \lim_{y\to-\infty} \xi_\theta(y) = \theta^2, \quad (2.68)$$

so we can take

$$r(\theta) = \sup_{y\in\mathbb{R}}\left\{\frac{\xi_\theta(y)}{2} - l(\theta)\right\} = -\frac{\theta}{2}(1-\theta). \quad (2.69)$$

Finally, we get

$$\varphi_\theta(y) = \frac{1}{r(\theta)}\left(\frac{\alpha^2_\theta(y) + \alpha'_\theta(y)}{2} - l(\theta)\right) \quad (2.70)$$

$$= \frac{1}{(1+y^2)(1-\theta)}\left(1 + y^2(1-\theta) - \frac{1}{\sqrt{1+y^2}}\right). \quad (2.71)$$

Therefore, the Hyperbolic diffusion with parameter space $\Theta \subset (-\infty, 0)$ satisfies all the necessary conditions for the application of the latent variable extension and MCMC method described above.

We fix the true value of the parameter at $\theta_0 = -2$ and generate a single skeleton for the hyperbolic diffusion in the time interval $[0, 24]$. We then generate a sample of $2,400$ equally spaced data points via Brownian bridge interpolation between skeleton points. Once again, we have 100 observations per time unit. Figure 2.6 shows the data and corresponding skeleton.

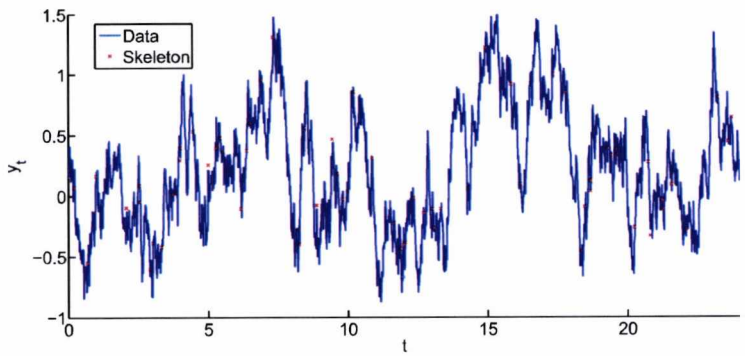


Figure 2.6: 2,400 data points from the hyperbolic diffusion in the time interval $[0.100]$, with parameter $\theta_0 = -2$ and initial point $y_0 = 0$.

We define a uniform prior $\Pi(\theta) = \mathrm{U}(\theta| -11, 0)$ on the parameter space. Once again, we produce a sample from the posterior distribution $\Pi^n(\theta)$, for increasing sample sizes. For this, we use the MCMC algorithm with a Monte Carlo sample size of $N = 10,000$, with a burning period of 10.000 iterations and a thinning of 1 every 10 iterations of the Chain. We consider the data set consisting of the first n data points, in the time interval $[0, T]$, for $T = 3, 6, 12, 24$, and $n = 100T$. Figure 2.7 shows the estimated posterior densities for the parameter.

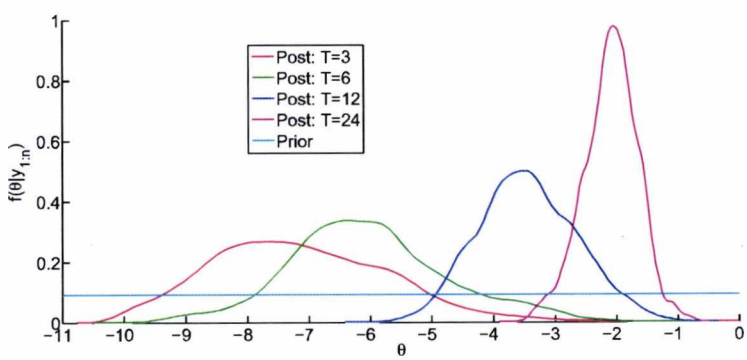
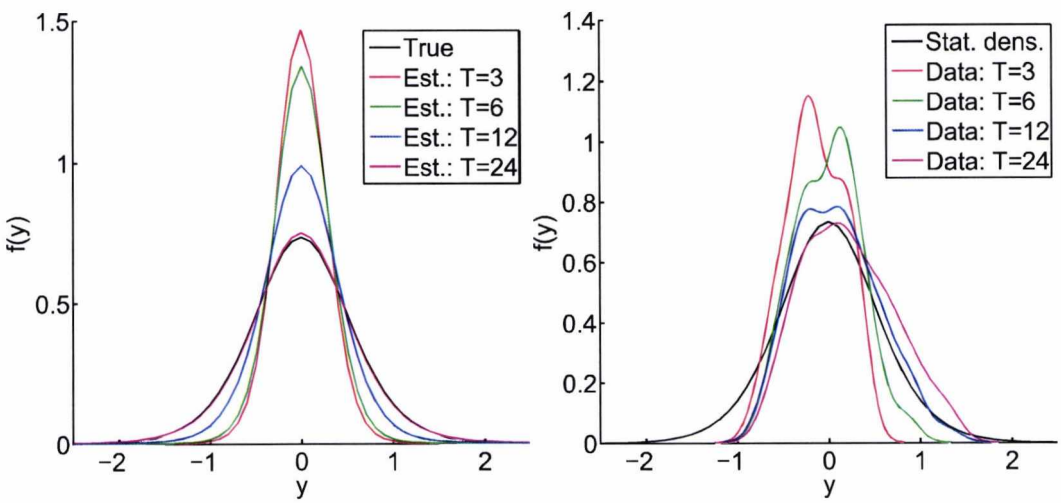


Figure 2.7: Estimated posterior density for the parameter of the hyperbolic diffusion.

As with the sine diffusion, the estimated posterior density seems to accumulate around the true value $\theta_0 = -2$ as the sample size n and the limit T of the time interval of observation grow. This occurs at a faster rate than with the sine

diffusion, a phenomenon that may be related to the stationarity of the hyperbolic diffusion. Such stationarity also raises the question of accurate estimation of the stationary density. On the left hand side of Figure 2.8 we show Monte Carlo estimates of the stationary density for the different sample sizes, as well as the true stationary density. The normalizing constant for the latter is calculated via numerical integration. It can be seen that the estimated density accurately



(a) Stationary density

(b) Data histogram

Figure 2.8: True and estimated stationary density for the hyperbolic diffusion with parameter $\theta_0 = -2$ (left panel). Smoothed histograms for the data at increasing sample sizes on the right panel.

recovers the true stationary density of the process. Notice that the initial point $y_0 = 0$ was not chosen arbitrarily, but as the mode of the stationary density. On the right hand side of Figure 2.8 we present smoothed histograms of the samples with increasing size used for inference throughout. This shows that the sample, produced by the MCMC version of the exact simulation algorithm, adequately reproduces the stationary density of the process, as would be expected.

2.4 Discussion

The retrospective rejection sampler of Beskos *et al.* (2006b) was originally meant for the simulation of diffusion paths at arbitrary points in time. It was then shown to be well suited for maximum likelihood estimation of the model parameters, appearing in the diffusion and drift coefficients. The discussion of its use for Bayesian inference has so far been limited. One of the main advantages of the method, from the simulation point of view, is its exactness, derived from the rejection technique rather than a Markov Chain construction. When using the algorithm for maximum likelihood based inference, a Monte Carlo error must be introduced and, when the focus is on Bayesian inference, the MCMC approach is inevitable for posterior simulation of the parameter. It can be argued that the model still provides advantages with respect to other approximation methods, deriving from the well known properties of MCMC estimation.

The exact simulation algorithm relies on the introduction of a set of latent variables, the skeleton, conditional on which, the parameter is independent of the observations. More importantly, latent variables and parameter are conditionally independent, given the discrete observations, from the unobserved diffusion path between observation points. In this Chapter, we have shown that the latent variable construction is consistent with a more general auxiliary variable method for dealing with intractable likelihoods. We have shown how an MCMC approach for posterior simulation can be implemented, which does not depend on the length of the time interval in which the process is observed. The Markov chain alternative to the original algorithm seems more naturally suited for Bayesian inference with no additional source of error being introduced. Furthermore, the posterior simulation method we propose, allows us to generate posterior samples of the diffusion skeleton beyond the time interval defined by the data. This is an advantage when the emphasis of the analysis is on prediction.

As with the original rejection sampler, the MCMC approach can be used both for path simulation and posterior parameter simulation. The acceptance rate for the rejection sample decays with an increasing time interval size. Therefore, it is recommended that the simulation is carried out by dividing the interval into smaller sets of optimal length and performing the simulation sequentially in each

of this sets, with a dependence structure based on the Markov property. For a large time interval, this may become costly, since the gain in acceptance rate may not compensate the growth of the number of times that this step must be repeated. The MCMC approach, on the other hand, requires a certain amount of iterations before convergence, but it can be performed once over the entire time interval regardless of its size. Future work could include a careful analysis of the convergence properties of the MCMC algorithm and a performance comparison with the rejection sampler. We believe it may be possible to find conditions on the diffusion and time interval for simulation, under which each of the algorithms is more efficient. At this point we can only provide empirical evidence, based on the two examples presented above. For the sine diffusion, a sample size of $10,000$ data points in the time interval $[0, 100]$ made posterior inference using the retrospective sampler too time restrictive to present any results here. For the hyperbolic diffusion, a sample of $2,500$ points within the time interval $[0, 25]$ made even the MCMC approach slow.

At this point all algorithms have been implemented in Matlab (R2012a). Future work would also include the use of more efficient computer languages and a more careful handling of variables to improve computer speed. Then, a sensitivity study of the algorithm to the hyperparameters would be advisable.

Finally, we may consider the extension of the method to a wider family of diffusion processes. First, by replacing the constant diffusion coefficient with a general parametric function, then by removing some of the conditions on both the drift and diffusion coefficients. We believe this could be done through the introduction of further latent variables which would not greatly affect the simulation methods presented here.

Chapter 3

Stationary Time Series Model

In this Chapter we construct a flexible stationary model with nonparametric invariant and transition densities. We believe such construction is a straightforward way to apply the nonparametric mixture idea in the time series context.

The likelihood for the nonparametric model has an intractable component generated by an infinite mixture of parametric functions for which none of the available methods for posterior simulation can be applied. We show that this likelihood is an example of the general case studied in this thesis. Consequently, we provide a latent model extension for which posterior inference is possible using existing techniques for MCMC based inference.

We provide some illustrations, involving transition density estimation for different sets of simulated data. Interestingly, the stationary model can recover the transition density of time homogeneous processes which are not stationary. The complete analysis of this behaviour is beyond the scope of the current work, therefore we only briefly discuss the ability of the model to recover a non stationary transition in terms of the flexibility of the transition densities described by the model.

3.1 The Model

In order to illustrate the main idea behind the construction we propose, we start by considering a very simple parametric first order stationary time series model, the normal AR(1) (1.50) of Section 1.1.2.2.

For fixed $\beta_0, \beta_1 \in \mathbb{R}$ and $\omega > 0$, the transition density for this model is

$$\mathrm{N}(y_i|\beta_0 + \beta_1 y_{i-1}, \omega^2), \tag{3.1}$$

where $\theta = (\beta_0, \beta_1, w)$. If $|\beta_1| < 1$, the stationary density is given by

$$\mathrm{N}(y|\mu, \sigma^2), \tag{3.2}$$

where $\mu = \beta_0/(1 - \beta_1)$ and $\sigma^2 = \omega^2/(1 - \beta_1^2)$.

A common idea in the context of regression is to define a more flexible conditional density as a mixture of parametric densities

$$f(y|x) = \int_\Theta K_\theta(y|x)\mathrm{d}P_x(\theta). \tag{3.3}$$

We discuss this type of models in the next chapter. Here, we focus on the consequences of this type of structure for an autoregressive model, with nonparametric transition density given by $f(y_i|y_{i-1})$ defined above, where the normal autoregressive transition density is a common choice for the parametric kernel.

As mentioned in Section 1.1.2.2, the problem with this type of models is to find conditions on the parametric kernels and mixing distributions which guarantee stationarity, while allowing inference. To resolve this problem, we return to the basic normal AR(1) model. The transition and stationary densities define a joint density

$$\mathrm{N}_2\big((y, x)|(\mu, \mu), \Sigma\big), \tag{3.4}$$

where

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}; \quad \text{for } \rho = \beta_1. \tag{3.5}$$

Note that both of the marginals for this bivariate normal density are identical and equal to the stationary density.

In general, consider a parametric bivariate density $K_\theta(y, x)$ for which the marginals are identical; i.e

$$K_\theta(y) = \int K_\theta(y, x)\,\mathrm{d}\nu(x) \quad \text{and} \quad K_\theta(x) = \int K_\theta(y, x)\,\mathrm{d}\nu(y). \tag{3.6}$$

Clearly, a Markov process with transition density $K_y(y_n|y_{n-1})$, has a stationary density given by the marginal $K_\theta(y)$, since

$$K_\theta(y_n) = \int K_\theta(y_n|y_{n-1})\,K_\theta(y_{n-1})\,\mathrm{d}\nu(y_{n-1}). \tag{3.7}$$

As with all simple parametric models, the dynamics of this process is easily overwhelmed by real data. We propose a nonparametric version of this model, applying the nonparametric mixture construction directly over the bivariate density $K_\theta(y, x)$, thus ensuring that the overall stationarity of the model is preserved.

We begin with the stick breaking representation for nonparametric mixture models presented in Section 1.1.1.2, where each component K_θ is a parametric density over the product space $\mathbb{Y} \times \mathbb{Y}$. In other words, for every y and x in $\mathbb{Y}$, we construct a joint density,

$$f_P(y, x) = \sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y, x). \tag{3.8}$$

The prior $\Pi(P)$ is given by a stick-breaking process with base measure P_0 and parameters (α_j, ζ_j) for the Beta distribution defining prior for the weights. In other words,

$$P | w_{1:\infty}, \theta_{1:\infty} = \sum_{j=1}^{\infty} w_j \, \delta_{\theta_j}; \tag{3.9}$$

and conditional on $v_{1:\infty}$,

$$w_1 = v_1 \quad \text{and} \quad w_j = v_j \prod_{j'<j} (1 - v_{j'}), \tag{3.10}$$

so the prior is defined for the $v_{1:\infty}$ and the $\theta_{1:\infty}$ independently as

$$\Pi(\theta_{1:\infty}) = \prod_{j=1}^{\infty} P_0(\mathrm{d}\theta_j); \tag{3.11}$$

$$\Pi(v_{1:\infty}) = \prod_{j=1}^{\infty} \mathrm{Be}(v_j | \alpha_j, \zeta_j). \tag{3.12}$$

Following the same principle observed in the parametric case, we define a transition density as the conditional distribution for this joint, i.e.

$$f_P(y_n | y_{n-1}) = \frac{\sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y_n, y_{n-1})}{\sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y_{n-1})} \tag{3.13}$$

As before, this transition then defines a stationary Markov process with invariant density given by the marginal

$$f_P(y) = \sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y). \tag{3.14}$$

Observe that the transition mechanism can be re-expressed as

$$f_P(y_n|y_{n-1}) = \sum_{j=1}^{\infty} w_j(y_{n-1})\, K_{\theta_j}(y_n|y_{n-1}), \tag{3.15}$$

where

$$w_j(y) = \frac{w_j\, K_{\theta_j}(y)}{\sum_{j'=1}^{\infty} w_{j'}\, K_{\theta_{j'}}(y)}. \tag{3.16}$$

Therefore we have constructed a model for which both the transition and the stationary densities are defined as nonparametric mixtures.

Looking at equation (3.15) it is tempting to think of this model as a transition density mixture, in the spirit of (1.60). However, we do not propose a mixture of conditional distributions, but a mixture of bivariate ones; the nonparametric nature of the transition density is just a desirable consequence. In doing so, no additional conditions need to be verified to guarantee the existence of the stationary density. Any choice of a stationary parametric kernel, when combined with its corresponding stationary density to produce a joint over which the mixture is defined, results in a nonparametric stationary model.

The dependent weights (3.16) have an interpretation in terms of the region of applicability of each parametric model K_θ within the state space $\mathbb{Y}$. We discuss this interpretation in Chapter 4, in the context of nonparametric regression models.

So far, we have only defined what Martínez-Ovando & Walker (2011) refer to as a benchmark model. The construction is simple and the stationarity and flexibility are given by it, so no additional conditions need to be verified. This model, however, has not been used in the previous literature, as it has been considered to be practically intractable, due to the infinite mixture appearing in the denominator. Only a finite version of this model has been studied by Müller *et al.* (1997), who define a finite mixture of autoregressive AR(1) models, directly for the transition density and do not discuss conditions for stationarity of the process.

In the next Section, we apply the general methodology developed in the introduction to this thesis, for a particular choice of joint kernels. We construct

a tractable latent model, therefore enabling posterior inference for the nonparametric time series model, with normalized weights and a normal density kernel.

3.2 The Latent Model

The likelihood function for the nonparametric autoregressive model with normalized weights, given a sample $y_{0:n} = (y_0, \ldots, y_n)$ is the product

$$f_P(y_{1:n}) = \prod_{i=1}^{n} f_P(y_i|y_{i-1}) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y_i, y_{i-1})}{\sum_{j'=1}^{\infty} w_{j'} \, K_{\theta_{j'}}(y_{i-1})}, \tag{3.17}$$

and a stick-breaking prior Π is placed on the probability measure

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}. \tag{3.18}$$

We are assuming the first observation y_0 is fixed, but this is only in order to simplify notation, and is not an important assumption. We could equally assume that the first observation arises from the stationary density of the time series model, by including an additional factor

$$\sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y_0) \tag{3.19}$$

in the likelihood expression.

Due to the nature of the denominator of the likelihood expression (3.17) we have an intractable component. Our aim is to show how to undertake Bayesian inference for this model using well designed latent variables which result in a viable latent model, as anticipated in the Introduction.

To make this concrete, we adopt a particular parametric model based on the normal distribution. That is,

$$K_\theta(y, x) = \mathrm{N}_2\big((y, x)|(\mu, \mu), \Sigma\big),$$

where

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

for some $-1 < \rho < 1$. Hence, $\theta = (\mu, \sigma^2, \rho)$. In this case, the transition mechanism, modelled as the conditional density, is given by

$$K_\theta(y|x) = \mathrm{N}\big(y|\mu + \rho(x-\mu), (1-\rho^2)\sigma^2\big).$$

And the stationary density is simply $K_\theta(y) = \mathrm{N}(y|\mu, \sigma^2)$. Therefore, we are considering a nonparametric mixture of normal AR(1) models, but the joint mixture construction together with the choice of parametrization for the parametric kernels guarantee, as explained in the previous Section, the stationarity of the resulting time series model.

In order to illustrate the ideas while keeping the notation simple, we consider mixtures over means, i.e. the σ^2 and ρ are fixed across mixture components. Consequently, in what follows, we use

$$K_{\theta_j}(y|x) = \mathrm{N}\big(y|\mu_j + \rho(x-\mu_j), (1-\rho^2)\sigma^2\big),$$

$$K_{\theta_j}(y) = \mathrm{N}(y|\mu_j, \sigma^2).$$

As we have done in the Introduction and Chapter 2, we focus, for each conditional density $f(y_i|y_{1:i-1}) = f(y_i|y_{i-1})$ individually, and observe that it can be factorized as the product of a tractable and an intractable function

$$f(y_i|y_{i-1}) = g_i(y_i, y_{i-1}, w_{1:\infty}, \theta_{1:\infty}) h_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty}). \tag{3.20}$$

where

$$g_i(y_i, y_{i-1}, w_{1:\infty}, \theta_{1:\infty}) = \sigma \sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y_i, y_{i-1}) \tag{3.21}$$

is tractable in the sense that it is a standard nonparametric mixture model for which MCMC methods are available (see Section 1.2.2.2). On the other hand, for a given y_{i-1},

$$h_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty}) = \frac{1}{\sum_{j=1}^{\infty} w_j \, \exp\left\{-\frac{1}{2}(y_{i-1}-\mu_j)^2/\sigma^2\right\}}, \tag{3.22}$$

can be seen as an intractable normalizing constant for the density on y_i defined by g_i.

The denominator in this expression is bounded by 1, and hence it is possible to use the identity

$$\sum_{k=0}^{\infty}(1-b)^k = b^{-1} \quad \text{for any } 0 < b < 1 \tag{3.23}$$

to write

$$h_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty}) = \sum_{k_i=0}^{\infty}\left[1 - b_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty})\right]^{k_i}, \tag{3.24}$$

where

$$b_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty}) = \sum_{j=1}^{\infty} w_j \exp\left\{-\frac{1}{2}(y_{i-1} - \mu_j)^2/\sigma^2\right\}. \tag{3.25}$$

Then k_i is introduced as a latent variable, arriving at the expression

$$f(y_i, k_i|y_{i-1}) = g_i(y_i, y_{i-1}, w_{1:\infty}, \theta_{1:\infty})\left[1 - b_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty})\right]^{k_i}. \tag{3.26}$$

The original transition density (3.20) can be recovered by marginalization with respect to k, but in the latent expression, the intractable component has been moved from the denominator to the numerator. At this point, posterior inference via MCMC methods would still require the sampling of the infinite dimensional parameters $w_{1:\infty}$ and $\theta_{1:\infty}$, so further manipulation is required.

As we mentioned before, the tractable function g_i can be dealt with using standard techniques for nonparametric mixture models with a stick-breaking representation. We use the slice sampling ideas of Kalli *et al.* (2011) presented in Section 1.2.2.2. Concretely, we introduce a latent variable d_i which acts as an index for the specific component from which y_i is generated, conditional on y_{i-1}, thus, we write

$$g_i(y_i, y_{i-1}, d, w_d, \theta_d) = \sigma w_d \, K_{\theta_d}(y_i, y_{i-1}). \tag{3.27}$$

We deal with the intractable component in a similar manner, by first realizing that

$$\begin{aligned} h_{i,k_i} &= \left[1 - b_i(y_{i-1}, w_{1:\infty}, \theta_{1:\infty})\right]^{k_i} \\ &= \prod_{l=1}^{k_i}\left(\sum_{j=1}^{\infty} w_j\left[1 - \exp\left\{-\frac{1}{2}(y_{i-1} - \mu_j)^2/\sigma^2\right\}\right]\right) \\ &= \sum_{D_{i,1}=1}^{\infty}\cdots\sum_{D_{i,k_i}=1}^{\infty}\prod_{l=1}^{k_i} w_{D_{i,l}}\left[1 - \exp\left\{-\frac{1}{2}(y_{i-1} - \mu_{D_{i,l}})^2/\sigma^2\right\}\right] \end{aligned} \tag{3.28}$$

Then, we introduce the indices $D_{i,1:k_i}$ as latent variables, defining the latent expression

$$b_{i,l}(y_{i-1}, w_{D_{i,l}}, \theta_{D_{i,l}}, D_{i,l}) = w_{D_{i,l}} \left[1 - \exp\left\{ -\frac{1}{2} (y_{i-1} - \mu_{D_{i,l}})^2 / \sigma^2 \right\} \right] \tag{3.29}$$

Notice that equation (3.28) coincides with the general latent expression (13) in the Introduction. The latent variables $s_{i,l} = D_{i,l}$ for each $i = 1, \ldots, n$ and $l = 1, \ldots, k_i$, take values on $\mathbb{S} = \mathbb{N}$, with reference measure ν given by the counting measure.

Finally, the full latent expression for the likelihood of the time series model is given by

$$\begin{aligned} f_P(y_{1:n}, d_{1:n}, k_{1:n}, D_{1:n,1:k_i}) &= \prod_{i=1}^{n} g_i(y_i, y_{i-1}, d, w_d, \theta_d) \prod_{l=1}^{k_i} b_{i,l}(y_{i-1}, w_{D_{i,l}}, \theta_{D_{i,l}}, D_{i,l}) \\ &= \sigma^n \prod_{i=1}^{n} w_{d_i} \, \mathrm{N}_2\big((y_i, y_{i-1}) | (\mu_{d_i}, \mu_{d_i}), \Sigma\big) \\ &\quad \prod_{l=1}^{k_i} w_{D_{l,i}} \left[1 - \exp\left\{ -\frac{1}{2} (y_{i-1} - \mu_{D_{l,i}})^2 / \sigma^2 \right\} \right], \end{aligned}$$

from which the original likelihood is recovered by adding over the $d_{1:n}, k_{1:n}$ and $D_{1:n,1:k_i}$. The introduction of this latent variables makes posterior simulation for the $(\mu_j), (w_j), \sigma^2$ and ρ possible via MCMC, through the usual slice sampling method. In order to deal with the variable size of the sampling space induced by the dependence of each $D_{i,1:k_i}$ on k_i, we extend the model further by adding an infinite sequence $D_{i,l>k_i}$ of latent variables which interact with the latent model through fully known densities, in the manner of Godsill's 2001 general algorithm presented in Section 1.2.2.3.

3.3 Posterior Inference via MCMC

The Bayesian model is completed by defining a prior on the mixing measure P; effectively, on the σ^2, ρ and the $w_{1:\infty}$, $\mu_{1:\infty}$. We use a stick–breaking process prior, so for independently distributed $\mathrm{Be}(\alpha_j, \zeta_j)$ variables, $(v_j)_{j=1}^{\infty}$, for some $\alpha_j, \zeta_j > 0$,

we let

$$w_1 = v_1, \quad \text{and for } j > 1, \quad w_j = v_j \prod_{j'<j}(1 - v'_j). \tag{3.30}$$

Concretely, we illustrate the methods using a Dirichlet Process prior, making $\alpha_j = 1$ and $\zeta_j = \zeta$. Alternatively, to show that the same method can be applied for other stick-breaking constructions, we use the Geometric stick-breaking prior presented in Section 1.1.1. Recall that, in this case, the weights are defined as

$$w_j = v(1 - v)^{j-1}; \quad v \sim \text{Be}(\alpha, \zeta). \tag{3.31}$$

In both cases, the base function for the prior is given, independently for each of the $\mu_{1:\infty}$, σ^2 and ρ, as follows. We let the $(\mu_j)_{j\geq 1}$ be independent and identically distributed from a Normal distribution $\text{N}(\cdot|m, t^{-1})$. For $\tau = \sigma^{-2}$ we use a Gamma prior $\text{Ga}(a, c)$. Finally, we define a discrete uniform prior for ρ on some discrete set $R \subset (-1, +1)$. Alternative base measures may be used and inference would still be possible. However, this particular choice simplifies the calculations, and as we show in the illustrations, they are sufficient for the purpose of estimation, at least in the set of examples we present.

Together with the latent model, this prior provides a joint posterior density for all the variables which need to be sampled for posterior estimation, i.e. the parameters $(\sigma, \rho, w_{1:\infty}, \mu_{1:\infty})$ and the latent variables $(d_{1:n}k_{1:n}, D_{1:n,1:k_i})$. We now describe how each of this variables can be sampled through an MCMC scheme, using a general Gibbs sampler structure with Metropolis-Hastings steps when sampling from the full conditional distributions is not possible.

3.3.1 Updating the Indices, $\mathbf{d_i}$ and $\mathbf{D_{i,l}}$

There is still an issue due to the infinite possible values that each of the $D_{i,l}$ and d_i can take. We overcome this, using the slice sampling technique, previously discussed. In order to reduce to a finite set the values from which this variables must be sampled at each step of the MCMC algorithm, we extend the joint model further, through the introduction of indicator variables,

$$\mathbf{1}\left(\nu_i < e^{-\xi d_i}\right) \quad \text{and} \quad \mathbf{1}\left(\nu_{i,l} < e^{-\xi D_{i,l}}\right),$$

for some $\xi > 0$. Hence, the full conditional distributions for the latent indices are given by

$$\mathbb{P}(d_i = j|\cdots) \propto w_j \, e^{\xi j} \, K_{\theta_j}(y_i, y_{i-1}) \, \mathbf{1}\{1 \leq j \leq J_i\}$$

and

$$\mathbb{P}(D_{i,l} = j|\cdots) \propto w_j \, e^{\xi j} \left[1 - \exp\left\{-\frac{1}{2}(y_{i-1} - \mu_j)^2/\sigma^2\right\}\right] \mathbf{1}\{1 \leq j \leq J_{i,l}\}$$

where

$$J_i = \lfloor -\xi^{-1} \log \nu_i \rfloor; \quad J_{i,l} = \lfloor -\xi^{-1} \log \nu_{i,l} \rfloor.$$

Notice that the full conditional densities for the variables involved in the MCMC algorithm do not depend on values of the weights and means (μ_j, w_j) for $j > J = \max_{i,l}\{J_i, J_{i,l}\}$. Therefore, at any given iteration, we only need to sample the (μ_j, w_j) for $j = 1, \ldots, J$, thus solving the problem of the infinite number of mixture components.

3.3.2 Updating the Mixture Weights, $\mathbf{w_{1:J}}$

We next describe how to sample the $w_{1:J}$ at each iteration of the MCMC algorithm. As is well known, when a Dirichlet Process prior is used, these can again be calculated as $w_1 = v_1$ and $w_j = v_j \prod_{l<j}(1 - v_l)$ for $j > 1$. This time, the $v_{1:J}$ must be independently sampled from their full conditional distribution, which can be identified as

$$f(v_j|\cdots) = \text{Be}(\alpha_j + n_j + N_j, \zeta_j + n_j^+ + N_j^+),$$

where

$$\begin{aligned} n_j &= \sum_{i=1}^{n} \mathbf{1}(d_i = j); \quad N_j = \sum_{i=1}^{n} \sum_{l=1}^{k_i} \mathbf{1}(D_{i,l} = j); \\ n_j^+ &= \sum_{i=1}^{n} \mathbf{1}(d_i > j); \quad N_j^+ = \sum_{i=1}^{n} \sum_{l=1}^{k_i} \mathbf{1}(D_{i,l} > j). \end{aligned} \tag{3.32}$$

Alternatively, if a Geometric stick-breaking prior is used, the updated weights are calculated as $w_j = v(1-v)^{j-1}$, where v is sampled from the full conditional distribution given by

$$f(v|\cdots) = \text{Be}(\hat{\alpha}, \hat{\zeta}),$$

where

$$\hat{\alpha} = \alpha + 1 + \sum_{i=1}^{n-1}(k_i + 1);$$
$$\hat{\zeta} = \zeta + \sum_{i=1}^{n}(d_i - 1) + \sum_{i=1}^{n-1}\left(\sum_{l=1}^{k_i} D_{i,l} - k_i\right).$$

3.3.3 Updating the Correlation Coefficient, ρ

A discrete prior for the correlation coefficient ρ, results in a discrete full conditional distribution given by

$$\mathbb{P}(\rho = r|\cdots) \propto \pi(r)(1-\rho^2)^{-\frac{n-1}{2}} \exp\left\{-\frac{1}{2}\tau \sum_{i=1}^{n} \hat{\mu}_i' \Sigma_r^{-1} \hat{\mu}_i\right\}$$

where

$$\hat{\mu}_i = \begin{pmatrix} y_i - \mu_{d_i} \\ y_{i-1} - \mu_{d_i} \end{pmatrix}, \quad \Sigma_r = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}. \tag{3.33}$$

for every $r \in R$. This can be sampled directly, given a particular choice of R and discrete prior π over it.

3.3.4 Updating the Precision Term, $\tau = \sigma^{-2}$

Before updating the τ, it is convenient to introduce some additional latent variables $u_{1:n,1:k_i}$, which allow us to substitute the terms

$$\prod_{l=1}^{k_i}\left[1 - \exp\left\{-\frac{1}{2}(y_{i-1} - \mu_{D_{i,l}})^2/\sigma^2\right\}\right]$$

in the latent likelihood expression, with truncation terms

$$\prod_{l=1}^{k_i} \mathbf{1}\left[u_{i,l} < 1 - \exp\left\{-\frac{1}{2}(y_{i-1} - \mu_{D_{i,l}})^2/\sigma^2\right\}\right].$$

Recall that $\tau = \sigma^{-2}$ is assigned a $\text{Ga}(\tau|a, c)$ prior, which is the conjugate prior for the precision of the Normal density kernel. Therefore, the full conditional distribution for τ is a truncated Gamma distribution,

$$f(\tau|\cdots) \propto \text{Ga}(\tau|\hat{a}, \hat{c})\,\mathbf{1}(\tau > T),$$

where

$$\hat{a} = a + n/2;$$
$$\hat{c} = c + \frac{1}{2}\sum_{i=1}^{n} \hat{\mu}_i' \Sigma_\rho^{-1} \hat{\mu}_i;$$
$$T = \max\left\{\frac{-2\log(1-u_{i.l})}{(y_i - \mu_{D_{i,l}})^2} : i = 1,\ldots,n;\ l = 1,\ldots,k_i\right\};$$

and the $\hat{\mu}_i$, Σ_ρ are defined as in expression 3.33.

Numerous sampling routines may be used to sample from this truncated distribution. We use the latent variable based MCMC method of Damien & Walker (2001), presented in Section 1.2.2.1.

3.3.5 Updating the Kernel Means, $\mu_{1:J}$

The sampling of the $\mu_{1:J}$ is also not problematic. For each j, the prior for μ_j is $\text{N}(\cdot|m, t^{-1})$, therefore the full conditional distribution, given the rest of the variables is a truncated Normal

$$f(\mu_j|\cdots) \propto \text{N}(\mu_j|m_j, t_j^{-1})\mathbf{1}\left\{\mu_j \in \cap_{i=1}^{n} A_{j,i}\right\}$$

where

$$m_j = \frac{1}{t_j}\left[mt + \frac{\tau}{1-\rho}\sum_{d_i=j}(y_i + y_{i-1})\right];$$
$$t_j = t + \frac{2\tau n_j}{1+\rho};$$

and the truncation is defined by the sets $A_{j,i} = (-\infty, y_i - a_{j,i}) \cup (y_i + a_{j,i}, \infty)$,

$$a_{j,i} = \max_l \left\{ \sqrt{-2\tau^{-1}\log(1-u_{i,l})} : D_{i,l} = j \right\}, \tag{3.34}$$

with the convention $\max\{\emptyset\} \equiv \infty$, $\min\{\emptyset\} \equiv -\infty$.

Once again, we use the latent variable based MCMC method of Section 1.2.2.1 to sample from this truncated distribution, but alternative approaches can be found in the literature.

3.3.6 Updating the Latent Model Dimension, k

Since the dimension of the sampling space changes with k_i, we use ideas involving reversible jump MCMC (Godsill, 2001; Green, 1995), as explained in Section 1.2.2.3.

For each $i = 1, \ldots, n$, with probability $0 < p < 1$, we propose a move from k_i to $k_i + 1$ and accept it with probability

$$\min\left\{1, \frac{1-p}{p}\left[1 - \exp\left\{-\frac{1}{2}\tau(y_i - \mu_{D_{i,k_i+1}})^2\right\}\right]\right\}.$$

Clearly, the evaluation of this expression requires the sampling of the additional D_{i,k_i+1}. We take $D_{i,k_i+1} = j$ with probability w_j.

Whenever a move of this type is not proposed, and if $k_i > 0$, we accept a move to $k_i - 1$ with probability

$$\min\left\{1, \frac{p}{1-p}\left[1 - \exp\left\{-\frac{1}{2}\tau(y_i - \mu_{D_{i,k_i}})^2\right\}\right]^{-1}\right\},$$

in which case, the last latent variable D_{i,k_i} is dropped.

We have shown it is possible to perform posterior inference for the stationary time series mixture model we propose. Now we illustrate this in practice with some examples in the next Section.

3.4 Illustrations

In this Section we present some examples that illustrate the usefulness of the time series model presented in Section 3.1, focusing on statistical properties related to prediction.

We present four examples, all of them involving simulated data. In the first example, data is simulated from the stationary model with a fixed known number of fully known mixture components. In the second example, the data corresponds to the discretely observed hyperbolic diffusion discussed in Section 2.3.2. In other words, it is data generated by a stationary process which is not stated in terms of a nonparametric mixture, but still falls within the general definition of the time series model studied in this chapter. In these two examples, we use posterior simulation to recover the transition and stationary densities, the latter corresponding with the data histogram for a large enough sample.

The last two examples are somewhat more interesting, since they are generated from processes for which the stationary density does not exist. Not surprisingly, in this case, our model fails to recover the shape of the data histogram. Nevertheless, both examples have fixed time homogeneous transition densities and we are able to estimate them using the nonparametric stationary mixture model specified in Section 3.1.

Therefore, the set of examples is chosen to illustrate how our model can be used for transition and invariant density estimation simultaneously, when the stationary density exists, and is still useful for transition density estimation, even when the data is not generated by a stationary process.

3.4.1 Example 1: Stationary Mixture Model

We generate a sample of size $n = 1000$ from the stationary mixture model with normalized weights described in this Chapter, with three mixture components and true parameters $\mu_0 = (-1, 0, 3)'$, $w_0 = (0.1, 0.4, 0.5)'$, $\sigma_0^2 = 1$ and $\rho_0 = 0.8$. The data is shown in Figure 3.1.

The prior for the mixing probability P, described above, requires the specification of some hyperparameters. We take a discrete uniform prior for ρ on the set

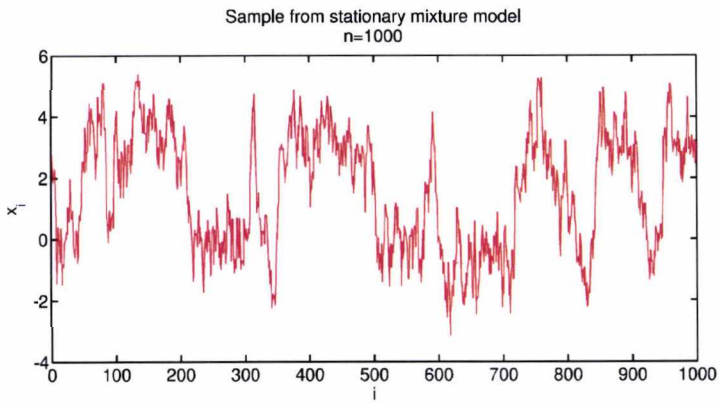


Figure 3.1: Sample of size $n = 1000$ simulated from the stationary mixture model with three mixture components and true parameters $\mu_0 = (-1, 0, 3)'$, $w_0 = (0.1, 0.4, 0.5)'$, $\sigma_0^2 = 1$ and $\rho_0 = 0.8$.

$R = \{r/200 : r = 1, \ldots, 200\}$. The rest of the parameters for the base measure f_0 are

$$\begin{aligned} m = 0, \quad t = 1/4 \quad &\text{for the } \mu_j; \\ a = 1/2, \quad c = 1 \quad &\text{for the } \tau, \end{aligned}$$

and we take $\alpha_j = 1$, $\zeta_j = 1$ for all j, corresponding to a Dirichlet process prior with unit mass parameter.

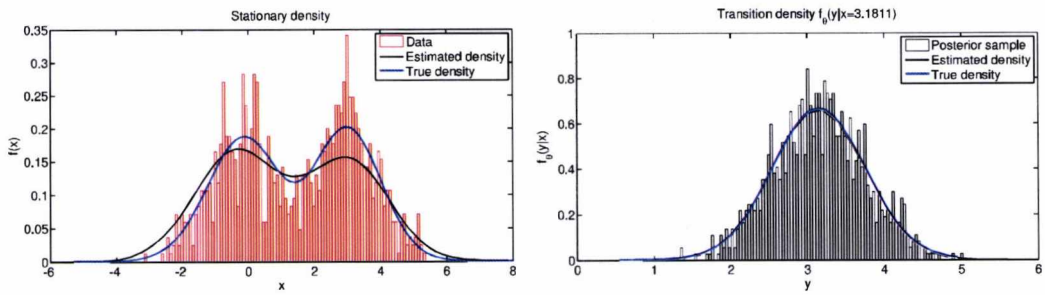
Posterior inference in this case can be carried out both for the stationary and the transition densities. Results are shown on the right and left panels of Figure 3.2

The estimated densities correspond to a Monte Carlo average of the posterior sample produced by the Markov Chain scheme for the latent model. We use a Monte Carlo sample size of $N = 1000$ after a burn in period of 9000 iterations.

Notice that the estimate for the transition density corresponds with the predictive density for Y_{n+1} given the sample, i.e.

$$f_n(y|y_n) = \int f(y|y_n) \mathrm{d}\Pi^n(f).$$

As might be expected, the transition density is recovered by the model better than the stationary density. This can be attributed to the fact that each new



(a) Stationary density (b) Transition density

Figure 3.2: Histogram of the data, with estimated and true stationary densities (left) for a sample from the stationary mixture model. On the right, the true transition density with the estimated density and the histogram of a sample from the predictive.

data point provides more information about the transition mechanism, while the information about the invariant measure is disturbed by the dependence between data points. However, given that the sample size is relatively small for this type of analysis, we believe the estimates to be satisfactory.

3.4.2 Example 2: Stationary Diffusion

We now consider the discretely observed diffusion process introduced in Section 2.3.2. Recall, this is the process $Y = \{Y_t : t \geq 0\}$ defined as a weak solution to the SDE

$$\mathrm{d}Y_t = -\theta \frac{Y_t}{\sqrt{1+Y_t^2}}\mathrm{d}t + \mathrm{d}W_t$$

For $\theta < 0$, we know this is a stationary process, with invariant density

$$f(y) \propto \exp\left\{2\theta\sqrt{1+y^2}\right\}.$$

A sample of size $n = 1000$ of observations, $y_{1:n}$ at times $t_i = i;$, is generated using the exact simulation algorithm of Beskos *et al.* (2006b), from the Hyperbolic diffusion with true parameter $\theta_0 = -2$. The data is shown in Figure 3.3.

The SDE provides a parametric model. However a nonparametric model should be flexible enough to recover the dynamics of the process generating the

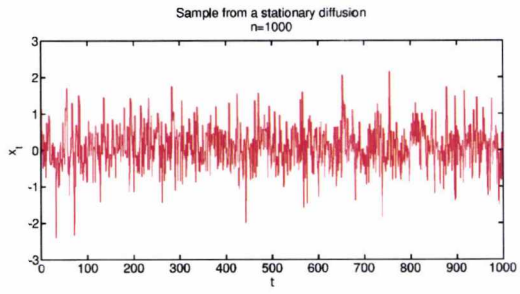
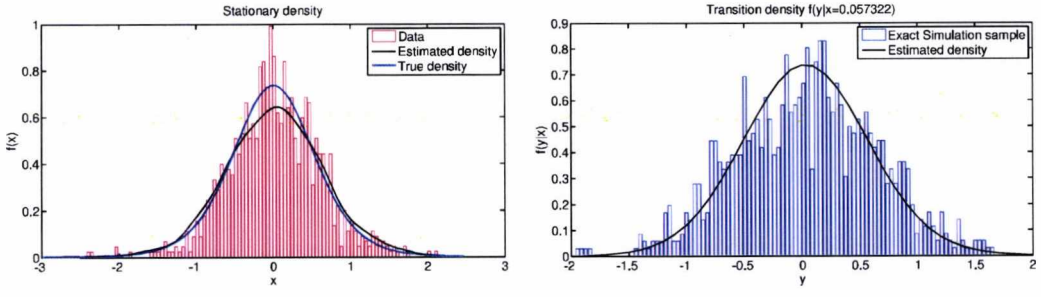


Figure 3.3: Sample of size $n = 1000$ from a discretely observed stationary diffusion process

data. To illustrate this, we do posterior inference using the stationary time series mixture model described in this Chapter, with the following prior specifications. The parameters of the prior density for each of the mixture kernel means are chosen as $m = \bar{y}_n$ and $t = 1/s^2$, the sample mean and precision respectively. As in the previous example the prior for the mixture precision is a Gamma distribution with parameters $a = 1/2$ and $c = 1$ and we use a Dirichlet process prior with unit mass parameter. The correlation coefficient is kept fixed at $\rho = 1$ for illustrative purposes.



(a) Stationary density

(b) Transition density

Figure 3.4: Histogram of the data, with estimated and true stationary densities (left) for a sample from the hyperbolic diffusion. On the right, the estimated transition density and the histogram of a sample generated from the true conditional, via exact simulation.

Again, posterior inference in carried out for the stationary and the transition

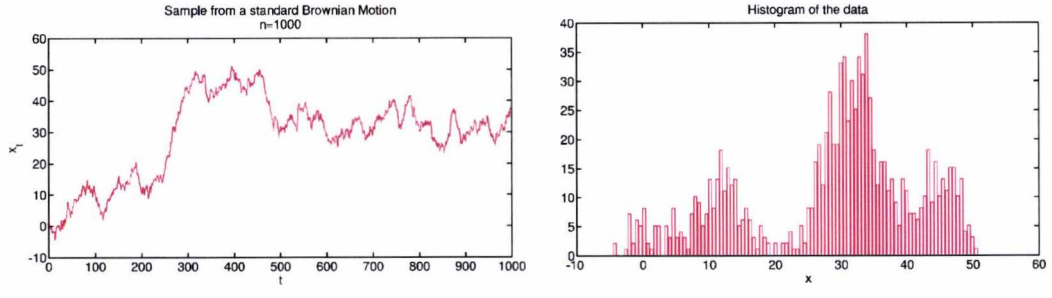
densities, through posterior simulation for the latent model via MCMC, with a Monte Carlo sample size of $N = 1000$ after a burn in period of 9000 iterations. Results are shown on the right and left panels of Figure 3.4, respectively.

The the normalizing constant for the true stationary density is calculated by numerical integration. The estimated transition density is compared to a histogram of a sample of size 1000 points generated from the true diffusion transition, via exact simulation. Both the stationary and the transition densities can be seen to be accurately recovered by the model.

3.4.3 Example 3: Standard Brownian Motion

Standard Brownian motion is a typical example of a non stationary process. For discrete observations at times $t_i = i$, the transition density is known and given by $f(y_i|y_{i-1}) = \mathrm{N}(y_i|y_{i-1}, 1)$, the standard normal distribution centred at y_{i-1}.

Figure 3.5 shows a sample of $n = 1000$ observations, $y_{1:n}$, at times $t_i = i$ from a standard Brownian Motion path, and the corresponding histogram. The latter shows the irregular behaviour of the data, since the marginal distribution of each y_i changes with i.

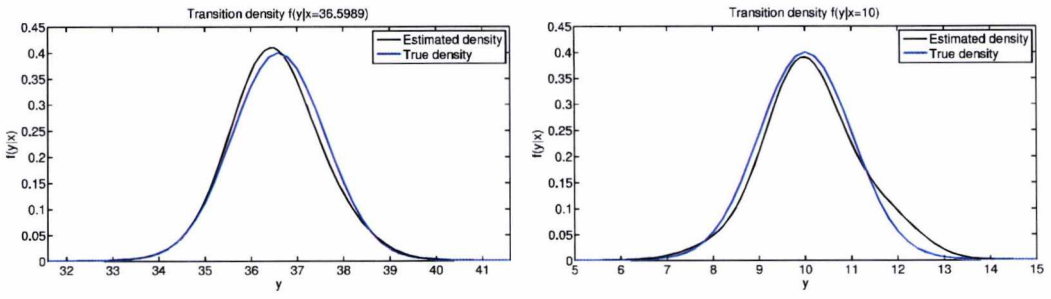


(a) Data (b) Histogram

Figure 3.5: $n = 1000$ equally spaced points from a standard Brownian motion path (left) and the corresponding histogram (right).

The mixture model we propose for time series is stationary. However, for any fixed sample size, it is flexible enough to capture the dynamics of the data, in the sense that we may use the model to estimate the transition density. Figure

3.6 shows estimates of the transition density $f(y|x)$ for two different values of x. The plot on the left hand side corresponds to the predictive density, i.e.



(a) Transition density for $x = x_n$

(b) Transition density for $x = 10$

Figure 3.6: Estimated transition densities $f(y|x)$ given a sample of $n = 1000$ data points from a discretely observed Brownian Motion path. On the left, $x = 36.6$ is the last data point; on the right $x = 10$.

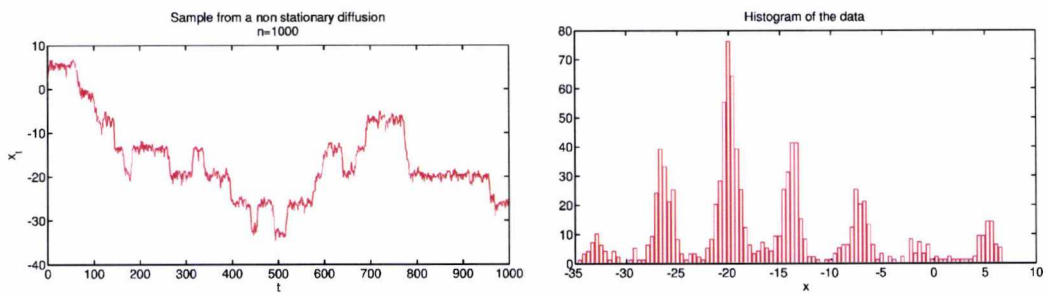
the estimated conditional density given the last observation, $\mathbb{E}[f(y|y_n)|y_{1:n}]$. The sample size is relatively small for this type of problems, yet the model can recover the transition density shape. The plot on the right corresponds to the estimated transition given $x = 10$, which from the data histogram, we know is in a region of the state space not frequently visited by this particular path. This accounts for the heavy right tale of the estimated density with respect to the true one. Overall, we can conclude that the model recovers the transition mechanism generating the data.

3.4.4 Example 4: Non-Stationary Diffusion

Finally, we consider the discretely observed diffusion process introduced in Section 2.3.1, i.e. a stochastic process $Y = \{Y_t : t \geq 0\}$ defined as a weak solution to the SDE

$$\mathrm{d}Y_t = \sin(Y_t - \theta)\mathrm{d}t + \mathrm{d}W_t.$$

A sample of size $n = 1000$ of observations, $y_{1:n}$ at times $t_i = i$, is generated using the exact simulation algorithm of Beskos *et al.* (2006b), with true parameter $\theta_0 = 2$. Figure 3.7 shows the data and corresponding histogram.

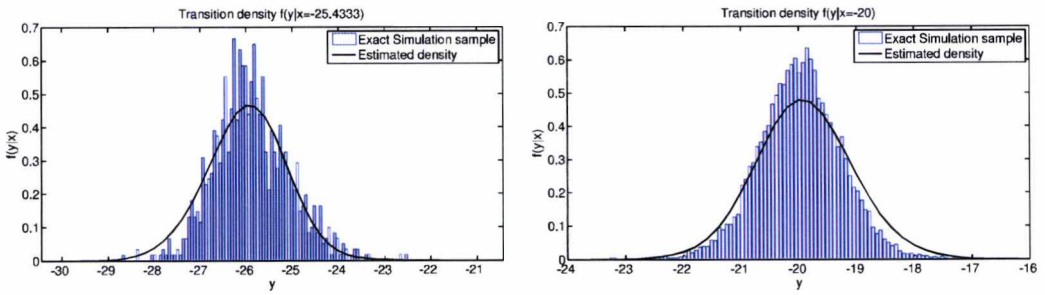


(a) Data (b) Histogram

Figure 3.7: $n = 1000$ equally spaced points from the sine diffusion with true parameter $\theta = 2$ (left) and the corresponding histogram (right).

Posterior inference in carried out for the transition density, through posterior simulation, via the MCMC algorithm for the latent model presented in this Chapter. Once again, the Monte Carlo sample size is $N = 1000$ after a burn in period of 9000 iterations. The parameters of the prior density for each of the mixture kernel means are chosen as $m = \bar{y}_n$ and $l = 1/s^2$, the sample mean and precision respectively. As in the previous example, the prior for the mixture precision is a Gamma distribution with parameters $a = 1/2$ and $c = 1$, and we use a Dirichlet process prior with unit mass parameter; the correlation coefficient is kept fixed at $\rho = 1$.

Figure 3.8 shows the estimated transition density $f(y|x)$ given the last observation, $x = y_n$ (left panel) and given $x = -20$ (right panel). The true transition density for this data is unknown, but the estimates are compared against histograms of samples of size 1000, generated from the true model via exact simulation. Given the irregularity of the data, exhibited in the histogram of Figure 3.7, and the relatively small sample size, the heavier tails of the estimated densities with respect to the exact simulated samples is justified. Overall, the transition density estimates can be considered accurate.



(a) Data (b) Histogram

Figure 3.8: Estimated transition densities $f(y|x)$ given sample of $n = 1000$ data points from a discretely observed sine diffusion. On the left, $x = y_n = -25.4$; on the right $x = -20$.

3.5 Discussion

We have presented a stationary Markov model for which both the transition and stationary densities are nonparametric. The construction is based on an infinite mixture of joint parametric kernels $k_\theta(y, x)$ for which both marginals are identical. The stationary density for the process is then given as the infinite mixture of such marginals, and the transition density is the corresponding conditional density, given by the ratio between the joint and the marginal mixtures. The infinite sum in the denominator can be seen an intractable normalizing constant for the conditional density, or transition. We then extend the model by the introduction of latent variables, based on a series expansion for the normalizing constant; and some existing auxiliary variable methods in the context of inference for mixture models, and model selection with unknown parameter space dimension. The stationary model falls within the general class of intractable models studied in this thesis and the latent model is a particular example of the auxiliary variable scheme we propose for MCMC posterior inference with no approximation error.

We have illustrated the use of the stationary nonparametric model for posterior estimation of the transition and stationary densities when the data is generated by some true but arbitrary stationary process. In this case, a fixed true joint density for pairs of observations is available, and the model is able to recover

it. At the same time, the stationary density is estimated and estimation of the transition density follows as a ratio of the two.

When the data is generated by a non stationary but time homogeneous process, the model is still able to estimate the transition density, as we have empirically shown through some examples. In this case, there are no fixed marginal and joint densities to replicate, so the numerator and the denominator in the transition density expression do not have a direct interpretation. It is a known fact that a ratio can remain constant even when the numerator and denominator change. An analogous phenomenon explains the capacity of a stationary model to replicate a non stationary transition mechanism. Future work will involve the study of the properties of this model, as well as possible interpretations of the numerator and denominator expressions defining the transition density, when the data is not stationary.

We have demonstrated the latent model construction and MCMC algorithm for a particular choice of parametric joint kernel, the bivariate Gaussian density. However, other kernel choices are available. Furthermore, the condition requiring both marginal densities to be equal is only needed to guarantee the stationarity of the mixture Markov model. Arbitrary joint kernels can be used to construct general autoregressive models if stationarity is not an issue. This includes the definition of multivariate time series models.

We have focused here on stationary Markov models. However, a higher order Markov dependence structure may be obtained if the nonparametric mixture is defined over multivariate kernels. If the joint kernel includes $m + 1$ random variables, an order m Markov transition can be defined as the ratio between the joint mixture over the m-variate marginal from which the $(m + 1)$-th variable has been integrated out. Future work would include the study of this type of autoregressive models and their properties.

Chapter 4

Nonparametric Regression Model

The contents of this chapter constitute the body of Antoniano-Villalobos *et al.* (2012).

As mentioned in Section 1.1.3, Bayesian nonparametric mixture models for regression have become a subject of intense research activity. This is due to the flexibility that models of this type can achieve, while retaining useful statistical properties. For these models to be truly flexible, it is necessary to construct covariate dependent weights which can be guaranteed to add up to one at each point of the covariate space.

Formally, we recall the nonparametric regression mixture model given by

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j), \tag{4.1}$$

and a prior specification on the particles $\{\theta_j\}_{j=1}^{\infty}$ and covariate-dependent weights $\{w_j(x)\}_{j=1}^{\infty}$, which must satisfy the constraint

$$\sum_{j=1}^{\infty} w_j(x) = 1 \text{ a.s. for all } x \in \mathbb{X}. \tag{4.2}$$

The usual way to satisfy this condition is the stick-breaking definition (1.66) of MacEachern's 1999 dependent Dirichlet processes. This is in fact, the only approach for which exact posterior sampling methods are available. Effectively, posterior simulation is made possible by imposing restrictions on the structure of

the dependent weights. The construction poses challenges in terms of the various choices that need to be made for functional shapes and hyper-parameters, when defining the $\{v_j(x)\}_{j=1}^{\infty}$. The difficulties are amplified by the lack of interpretation of the quantities involved. Moreover, combining continuous and discrete covariates in a useful fashion is far from straightforward.

On the other hand, condition (4.2) is easily satisfied by defining normalized weights, a version of which is given by

$$w_j(x) = \frac{w_j K(x|\psi_j)}{\sum_{j'=1}^{\infty} w'_j K(x|\psi'_j)}, \tag{4.3}$$

where the denominator must be finite a.s. We argue that this construction has a natural interpretation in the Bayesian setting, which allows for a simple choice of the kernel and hyper–priors involved. Moreover, it is shown to be applicable to both continuous and discrete covariates.

It is to be noted that the infinite sum in the denominator of (4.3) introduces an intractable normalizing constant for which no posterior simulation methods are available to date. Only finite versions of this type of model have been introduced in the literature (see e.g. Adams *et al.*, 2009; Møller *et al.*, 2006; Murray *et al.*, 2006; Pettitt *et al.*, 2003), since simulation methods are available only for the finite case. However, the construction bares a resemblance to the time series model of Chapter 3. Therefore, in a similar fashion, we present it as an example of our general approach to dealing with intractable components of an infinite-dimensional nature and present the latent structure that allows posterior inference through the use of MCMC methods.

4.1 The Model

The aim in this Section is to motivate the normalization approach, as an alternative to the stick-breaking construction of the covariate-dependent weights $w_j(x)$. The idea is to insist that a parametric regression model, used as a component for a mixture, must incorporate information about its range of applicability, or where it holds, within the covariate space. This concept is unrelated to the way in

which the covariates are provided in practice, be it in a random or deterministic fashion.

For concreteness, let $K(y|x,\theta)$ be the normal linear model $\mathrm{N}(y|X\beta,\sigma^2)$, where $X=(1,x)$ and $\theta=(\beta,\sigma^2)$. Bayesian nonparametric modelling can be conceived as the placement of parametric components throughout suitable spaces. Hence, we need to think about how to construct and define a component for a regression model. Our first claim is that $K(y|x,\theta)$ is not in itself sufficient to adequately define a component. The reason being that, as it stands, it would suggest that each and every parametric model in the mixture model is equally valid throughout the complete covariate space $\mathbb{X}$. This is not a sensible or realistic working assumption.

In most applications where a nonparametric regression model is sought, specific parametric components are only assumed to behave locally. Hence, for each j, we need to specify a region $\mathbb{C}_j \subset \mathbb{X}$ within which the parametric model $K(y|x,\theta_j)$ holds. The sets $(\mathbb{C}_j)$ need not constitute a partition, thus allowing for regions of the covariate space $\mathbb{X}$ within which more than one parametric model has an effect. Each region $\mathbb{C}_j$ is, of course, unknown to the experimenter, hence, in a Bayesian framework, the uncertainty about it should be incorporated into the overall model. This is achieved by specifying, for every j and every $A \subseteq \mathbb{X}$, the probability that model j applies within the set A.

With this in mind, we introduce the notion of $\Pi(A|j)$, the prior probability that regression model j applies within the set A. This naturally leads to the idea of a parametric density function $p(x|j)=K(x|\psi_j)$ for which

$$\Pi(A|j)=\int_A p(x|j)\,\mathrm{d}x.$$

Note that the conditional densities $p(x|j)$ are not related to whether the covariates are picked by an expert or sampled from some distribution, which itself could be known or unknown. They only indicate where, in $\mathbb{X}$ the experimenter believes the regression model j provides a good description of the data. In other words, if asked to provide a candidate covariate value for which component j applied, the experimenter would provide such a value by sampling it from $p(x|j)=K(x|\psi_j)$, since $K(x|\psi_j)$ is modelling where the j-th component, namely $K(\cdot|x,\theta_j)$, applies.

It is not clear at this point, how this probability setting can be incorporated into the nonparametric regression model. To proceed, we consider $w_j = \Pi(j)$, the prior probability, according to the experimenter, that an observation is generated by the parametric regression model j. Hence, $\Pi(A|j)\,\Pi(j)$ defines a joint measure on $\mathbb{X} \times \mathbb{N}$, corresponding to the probability that an observation is generated from the parametric model j and that this model provides indeed a good fit for the data within the set A of covariate values.

It is important to clarify that the density

$$p(x) = \sum_{j=1}^{\infty} \Pi(j)\, p(x|j)$$

does not correspond to the distribution from which the covariates are sampled, if indeed they are sampled; it simply represents the distribution of where the combined regression models hold. Therefore,

$$\Pi(A) = \int_A p(x)\, \mathrm{d}x$$

is the prior probability that the complete nonparametric regression model applies, for an arbitrary observation within the set A of covariates.

Once $\Pi(j) = w_j$ and $p(x|j) = K(x|\psi_j)$ have been defined, we can provide a form for $\Pi(j|x)$ based on an application of Bayes Theorem; namely $\Pi(j|x) \propto \Pi(j) \times p(x|j)$.

Recalling the nonparametric mixture model

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j),$$

each covariate dependent weight $w_j(x)$ represents the probability that the parametric model j applies to an observation with covariate value x. In other words, $w_j(x) = \Pi(j|x)$. Therefore, putting things together and incorporating the normalizing constant, we arrive at the normalized expression for the covariate dependent weights, i.e.

$$w_j(x) = \frac{w_j\, K(x|\psi_j)}{\sum_{j'=1}^{\infty} w_{j'}\, K(x|\psi_{j'})}.$$

where $0 \leq w_j \leq 1$ for all j and $\sum_{j=1}^{\infty} w_j = 1$.

We believe this representation for the covariate dependent weights is easy to interpret, in terms of the prior probabilities of observations coming from regression model j combined with prior specifications as to where regression model j applies. Furthermore, $K(x|\psi_j)$ can be modelled via a standard family of density functions, such as the normal, if x is a continuous covariate. In this case, the interpretation would be that there is some central location $\mu_j \in \mathbb{X}$ where regression model j applies best, and a parameter τ_j describing the rate at which the applicability of the model decays around μ_j. On the other hand, if x is discrete, then a standard distribution on discrete spaces can be used, such as the Bernoulli or its generalization, the categorical distribution. Even if x is a combination of both discrete and continuous covariates, it is still possible to specify a joint density by combining both discrete and continuous distributions. This is demonstrated in the next Section.

4.2 The Latent Model

Given a sample $(y_{1:n}, x_{1:n}) = \{(y_1, x_1), \ldots, (y_n, x_n)\}$, the likelihood function for the MDP model with normalized weights is given by

$$f_P(y_{1:n}|x_{1:n}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_i)\, K(y_i|x_i, \theta_j) \right),$$

with covariate dependent weights defined in equation (4.1).

Alternatively, for every $i = 1, \ldots, n$, we may write

$$f_P(y_i|x_i) = \frac{\sum_{j=1}^{\infty} w_j\, K(x|\psi_j) K(y_i|x_i, \theta_j)}{\sum_{j'=1}^{\infty} w_{j'}\, K(x|\psi_{j'})}.$$

The expression in the denominator can be seen as an intractable normalizing constant generated by the infinite sequences of weights, $w_{1:\infty}$ and kernel parameters $\psi_{1:\infty}$, $\theta_{1:\infty}$. Thus, the model falls within the scope of this thesis, and in the present Section we provide the latent variable extension that allows us to undertake Bayesian inference via MCMC posterior simulation. As mentioned in the Introduction, we rely for this, on the series expansion

$$\sum_{k=0}^{\infty} (1-b)^k = b^{-1}, \text{ for } 0 < b < 1, \tag{4.4}$$

as the key for incorporating auxiliary variables to the likelihood expression, thus obtaining a viable latent model.

In order to illustrate ideas with a simplified notation, we start by considering posterior estimation with a single data point. We assume the first q elements of x represent discrete covariates, each x_m taking values in $\{0, \ldots, G_m\}$, for $m = 1 \ldots, q$; the last p elements of x represent continuous covariates. In this case, we have

$$K(y|x, \theta_j) = \mathrm{N}(y|X\beta_j, \sigma_j^2),$$
$$K(x|\psi_j) = \prod_{m=1}^{q} \mathrm{Cat}(x_m|\rho_{j,m}) \prod_{m=1}^{p} \mathrm{N}(x_{m+q}|\mu_{j,m}, \tau_m^{-1}).$$

where $\theta_j = (\beta_j, \sigma_j^2)$, $\psi_j = (\rho_j, \mu_j, \tau)$, $X = (1, x)$ and $\mathrm{Cat}(\cdot|\rho_m)$ represents the categorical distribution, i.e.

$$\mathrm{Cat}(x_m|\rho_m) = \prod_{g=0}^{G_m} \rho_{m,g}^{\mathbf{1}(x_m=g)}.$$

Again, to simplify the expression, we are using $\tau_j \equiv \tau$ for all j. This is not a strong restriction, and it may be removed by making some realistic assumptions on τ_j.

The likelihood for this model may be written as

$$f_P(y|x) = \frac{1}{b(x, w_{1:\infty}, \psi_{1:\infty})} \sum_{j=1}^{\infty} w_j K(x|\psi_j)\, K(y|x, \theta_j), \tag{4.5}$$

where

$$b(x, w_{1:\infty}, \psi_{1:\infty}) = \sum_{j=1}^{\infty} w_j\, K(x|\psi_j),$$
$$K(x|\psi_j) = \prod_{m=1}^{q+p} K(x_m, |\psi_{j,m}),$$

and

$$K(x_m|\psi_{j,m}) = \begin{cases} \prod_{g=0}^{G_m} \rho_{m,g}^{\mathbf{1}(x_m=g)} & m = 1, \ldots, q \\ \exp\{-\frac{1}{2}\tau_{m-q}(x_m - \mu_{j,m-q})^2\} & m = q+1, \ldots, q+p. \end{cases}$$

Notice that we have redefined the kernel function $K(x;\psi_j)$ by cancelling the precision term τ from the normal density, which appears both in the numerator and the denominator of the normalized weights expression. In this way, we guarantee that $0 < b(x, w_{1:\infty}, \psi_{1:\infty}) < 1$ for all $x \in \mathbb{X}$, and sequences $w_{1:\infty}$ and $\psi_{1:\infty}$. We can, therefore, apply the series expansion (4.4) to write

$$\frac{1}{b(x, w_{1:\infty}, \psi_{1:\infty})} = \sum_{k=0}^{\infty} \left[1 - \sum_{j=1}^{\infty} w_j \, K(x|\psi_j)\right]^k = \sum_{k=0}^{\infty} \left[\sum_{j=1}^{\infty} w_j \left[1 - K(x|\psi_j)\right]\right]^k.$$

Then, we introduce k as a latent variable, and obtain the latent model

$$f_P(y, k|x) = \sum_{j=1}^{\infty} w_j K(x|\psi_j) \, K(y|x, \theta_j) \left[\sum_{j=1}^{\infty} w_j \left[1 - K(x|\psi_j)\right]\right]^k.$$

After moving the infinite sum from the denominator to the numerator. we can now deal with the mixture in the usual way (see for example Kalli *et al.*, 2011). As for the time series model of Chapter 3, we first introduce a latent variable d to indicate the mixture component to which a given observation is associated, thus obtaining

$$f_P(y, k, d|x) = w_d K(x|\psi_d) \, K(y|x, \theta_d) \left[\sum_{j=1}^{\infty} w_j \left[1 - K(x|\psi_j)\right]\right]^k.$$

For the remaining sum, we have the exponent k to consider. Therefore, we introduce k latent variables, $D_1, \ldots, D_k$, arriving at the latent model

$$f_P(y, k, d, D_{1:k}|x) = w_d K(x|\psi_d) \, K(y|x, \theta_d) \prod_{l=1}^{k} w_{D_l} \left[1 - K(x|\psi_{D_l})\right].$$

It is easy to check that the original likelihood (4.5) is recovered by marginalizing over the d, k and $D_{1:k}$.

For a sample of size $n \geq 1$ we simply need n copies of the latent variables. Therefore, the full latent model is given by

$$f_P(y_{1:n}, k_{1:n}, d_{1:n}, D_{1:n,1:k_i}|x_{1:n}) = \prod_{i=1}^{n} w_{d_i} K(x_i|\psi_{d_i}) K(y_i|x_i, \theta_{d_i}) \prod_{l=1}^{k_i} w_{D_{i,l}} \left[1 - K(x_i|\psi_{D_{i,l}})\right].$$

Inference can be achieved via posterior simulation, using the slice sampling method of Kalli *et al.* (2011) to deal with the infinite possible values that the $d_{1:n}$ and $D_{1:n,1:k_i}$ can take.

Once again, the original likelihood

$$f_\theta(y_{1:n}|x_{1:n}) = \prod_{i=1}^{n}\left(\sum_{j=1}^{\infty} w(x_i;\psi_j)\, K(y_i|x_i,\theta_j)\right).$$

can be easily recovered by marginalizing over the $d_{1:n}$, $k_{1:n}$, and $D_{1:n,1:k_i}$. However, the introduction of this latent variables makes Bayesian inference possible, via posterior simulation of the weights $w_{1:\infty}$, and kernel parameters $\theta_{1:\infty}$, $\psi_{1:\infty}$, as we show in the next Section.

4.3 Posterior Inference via MCMC

A prior for P, defined by a prior specification for the weights $w_{1:\infty}$, and parameters $\theta_{1:\infty}$ and $\psi_{1:\infty}$, completes the Bayesian model.

Our focus is on Stick-Breaking priors (Section 1.1.1.1), and we define the base measure P_0 through its associated density f_0, given by the product of the following components,

$$f_0(\beta_j,\sigma_j^2) = \mathrm{N}\big(\beta_j|\beta_0,\sigma^2\Sigma^{-1}\big)\mathrm{Ga}\big(1/\sigma^2|\tilde{a},\tilde{c}\big);$$
$$f_0(\mu_j,\tau) = \prod_{m=1}^{p}\mathrm{N}\big(\mu_{j,m}|\mu_{0,m},(\tau_m s_m)^{-1}\big)\mathrm{Ga}\big(1/\tau_m|a_m,c_m\big);$$
$$f_0(\rho_j) = \prod_{m=1}^{q}\mathrm{Dir}(\rho_{j,m}|\gamma_m).$$

Together with the joint latent model, this provides a joint density for all the variables which need to be sampled for posterior estimation, i.e. the (w_j,θ_j,ψ_j), $j=1,\ldots,\infty$ and the $(k_i,d_i,D_{l,i})$, $i=1,\ldots,n$; $l=1,\ldots,k_i$.

4.3.1 Updating the Indices, $\mathbf{d_i}$ and $\mathbf{D_{i,l}}$

We deal with the infinite sample space for the indices $d_{1:n}$ and $D_{1:n,1:k_i}$ as we did in section 3.3.1, by using the slice sampling technique of Kalli *et al.* (2011).

Accordingly, in order to reduce the choices represented by $(d_i, D_{i,l})$ to a finite set, we introduce new latent variables, $(\nu_i, \nu_{i,l})$, which interact with the model through the following indicating functions

$$\mathbf{1}\left(\nu_i < e^{-\xi d_i}\right) \quad \text{and} \quad \mathbf{1}\left(\nu_{i,l} < e^{-\xi D_{i,l}}\right), \tag{4.6}$$

for some $\xi > 0$. Hence, the full conditional distributions for the index variables are given by

$$\begin{aligned}
\mathbb{P}(d_i = j|\cdots) &\propto w_j\, e^{\xi j}\, K(x_i|\psi_j) K(y_i|x_i, \theta_j)\, \mathbf{1}\{1 \leq j \leq J_i\}; \\
\mathbb{P}(D_{i,l} = j|\cdots) &\propto w_j\, e^{\xi j} \left[1 - K(x_i|\psi_j)\right] \mathbf{1}\{1 \leq j \leq J_{i,l}\},
\end{aligned}$$

where $J_i = \lfloor -\xi^{-1} \log \nu_i \rfloor$; $\quad J_{l,i} = \lfloor -\xi^{-1} \log \nu_{l,i} \rfloor$.

At any given iteration, the full conditional densities for the variables involved in the MCMC algorithm do not depend on values beyond $J = \max_{l,i}\{J_i, J_{l,i}\}$, so we only need to sample a finite number of weights and kernel parameters.

4.3.2 Updating the Mixture Weights, $\mathbf{w_{1:J}}$

The $w_{1:J}$ can be updated at each iteration of the MCMC algorithm in the usual way, that is, by making $w_1 = v_1$ and, for $j > 1$, $w_j = v_j \prod_{j'<j}(1 - v_{j'})$. The $v_{1:J}$ must be independently sampled from the corresponding full conditionals, which can easily be identified as

$$f(v_j|\cdots) = \text{Be}(\alpha_j + n_j + N_j, \zeta_j + n_j^+ + N_j^+),$$

where the n_j, N_j, n_j and N_j are given by expression (3.32).

4.3.3 Updating the Regression Kernel Parameters $\theta_{1:J}$

The variables involved in the linear regression kernel, that is, the $\beta_{1:J}$ and $\sigma^2_{1:J}$ are updated in the standard way, well known in the context of Bayesian regression.

We sample independently for each j, from the full conditional density

$$
\begin{aligned}
f(\beta_j, \sigma_j^2 | \cdots) &= \mathrm{N}(\beta_j | \hat{\beta}_j, \sigma_j^2 \hat{\Sigma}_j^{-1}) \mathrm{Ga}(1/\sigma_j^2 | \tilde{a}_j, \tilde{c}_j), \\
\text{where} \qquad \hat{\beta}_j &= \hat{\Sigma}_j^{-1} (\Sigma \beta_0 + \underline{X}_j' \underline{y}_j); \\
\hat{\Sigma}_j &= \Sigma + \underline{X}_j' \underline{X}_j; \\
\tilde{a}_j &= \tilde{a} + n_j/2; \\
\tilde{c}_j &= \tilde{c} + (\underline{y}_j - \underline{X}_j \beta_0)' W_j (\underline{y}_j - \underline{X}_j \beta_0)/2; \\
W_j &= I_j - \underline{X}_j \hat{\Sigma}_j^{-1} \underline{X}_j'.
\end{aligned}
$$

Here, $\underline{X}_j$ denotes the $n_j \times (1+p+1)$ matrix, with rows given by $X_i = (1, x_i')$ for $d_i = j$; $\underline{y}_j$ is defined analogously; and I_j denotes the identity matrix of size n_j.

4.3.4 Updating the Covariate Kernel Parameters $\psi_{1:J}$

Similar to what we did for the time series model in Section 3.3.4, to update the $\psi_{1:J}$ it is convenient to introduce an additional set of latent variables. In order to do so, observe that, for any integer M and vector $(b_1, \ldots, b_M) \in (0,1)^M$, the following identity holds

$$
1 - \prod_{m=1}^{M} b_m = \sum_{u \in \mathbb{U}} \int_{(0,1)^M} \prod_{m=1}^{M} \left[u_m 1\{U_m < b_m\} + (1-u_m) 1\{U_m > b_m\} \right] \mathrm{d}U,
$$

where $U = (U_1, \ldots, U_M)$, $u = (u_1, \ldots, u_M)$ and $\mathbb{U}$ is the set of M-dimensional $\{0,1\}$ vectors of which at least one entry is 0.

We can, therefore, introduce latent variables $(u_{i,l,m}, U_{i,l,m})$, for $i = 1, \ldots, n$, $l = 1, \ldots, k_i$ and $m = 1, \ldots, q+p$, to deal with the terms $\left[1 - \prod_m K(x_{i,m} | \psi_{j,h})\right]$ in the likelihood. The full conditional density for $\psi_{1:J}$ is thus extended to the latent expression

$$
\begin{aligned}
f(\psi_{1:J}, u_{1:n,1:k_i,1:q+p}, U_{1:n,1:k_i,1:q+p} | \cdots) \propto \prod_{j=1}^{J} f_0(\psi_j) \prod_{i=1}^{n} \prod_{m=1}^{q+p} K(x_{i,m} | \psi_{d_i,m}) \\
\prod_{l=1}^{k_i} \left[u_{i,l,m} 1\{U_{i,l,m} < b_{i,l,m}\} + (1-u_{i,l,m}) 1\{U_{i,l,m} > b_{i,l,m}\} \right],
\end{aligned}
$$

where $b_{i,l,m} = K(x_{i,m}|\psi_{D_{i,l,m}})$, from which the original conditional density can be recovered by marginalizing over the $(u_{i,l,m}, U_{i,l,m})$.

The latent variables $(u_{i,l,m}, U_{i,l,m})$ can be sampled from their full conditional density by first observing that they are independent across $i = l, \ldots, n$ and $l = 1, \ldots, k_i$. For each (i, l), $u_{i,l,1:q+p}$, is a $(q + p)$-dimensional vector of zeros and ones with at least one zero entry. There are $2^{p+q} - 1$ such vectors and the update must be done according to the full conditional distribution given by

$$\mathbb{P}(u_{i,l,1:q+p} = u|\cdots) \propto \prod_{m=1}^{q+p} \left[u_m K(x_{i,m}|\psi_{D_{i,l},m}) + (1 - u_m)\left[1 - K(x_{i,m}|\psi_{D_{i,l},m})\right]\right].$$

This is a discrete probability measure with finite support, so the sampling can be done directly.

Conditional on $u_{i,l,1:q+p}$, the latent variables $U_{i,l,1:p+q}$ are independent. Each $U_{i,l,m}$ is uniformly distributed in the interval

$$\left[K(x_{i,m}|\psi_{D_{i,l},m})(1 - u_{i,l,m}),\ K(x_{i,m}|\psi_{D_{i,l},m})^{u_{i,l,m}}\right].$$

Hence, the additional variables do not pose a problem for posterior simulation. Furthermore, the introduction of these new variables transforms the latent term, introduced to deal with the intractable normalizing constant, into a truncation term over the usual posterior density for the nonparametric mixture. Posterior sampling for the $\psi_{1:J}$ is therefore achieved by independently sampling from truncated densities.

We first consider the update of the $\rho_{1:J}$, which is achieved by sampling each $\rho_{j,m}$, for $m = 1, \ldots q$, independently from a truncated Dirichlet distribution,

$$f(\rho_{j,m} \mid \cdots) \propto \mathrm{Dir}(\rho_{j,m} \mid \hat{\gamma}_{j,m})\, \mathbf{1}\,(\rho_{j,m} \in R_{j,m})\,.$$

The truncation region is defined as

$$R_{j,m} = \left\{\rho \in (0,1)^{G_m} : r^-_{j,m,g} < \rho_g < r^+_{j,m,g}, g = 1, \ldots, G_m\right\},$$

where, for $g = 0 \ldots . G_m$,

$$\begin{aligned}
\hat{\gamma}_{j,m,g} &= \gamma_{j,m,g} + \sum_{d_i = j} \mathbf{1}\,(x_{i,m} = g)\,; \\
r^-_{j,m,g} &= \max\left\{U_{i,l,m}\mathbf{1}\,(x_{i,m} = g) : D_{i,l} = j, u_{i,l,m} = 1\right\}, \\
r^+_{j,m,g} &= \min\left\{U_{i,l,m}^{\mathbf{1}\,(x_{i,m}=g)} : D_{i,l} = j, u_{i,l,m} = 0\right\}.
\end{aligned}$$

Next, we consider the $\mu_{1:J}$ and τ. First, we sample each τ_m, for $m = 1, \ldots, p$ independently from a truncated gamma density,

$$f(\tau_m \mid \cdots) \propto \mathrm{Ga}(\tau_h \mid \hat{a}_m, \hat{c}_m)\mathbf{1}\left(\tau_m^- < \tau_m < \tau_m^+\right),$$

where

$$\begin{aligned}
\hat{a}_m &= a_m + J/2, \\
\hat{c}_m &= c_m + \frac{1}{2}\sum_{i=1}^{n}(x_{i,m+q} - \mu_{d_i,m})^2 + \frac{1}{2}s_m\sum_{j=1}^{J}(\mu_{j,m} - \mu_{0,m})^2, \\
\tau_m^- &= \max\left\{\frac{-2\log U_{i,l,m+q}}{(x_{i,m+q} - \mu_{D_{i,l},m})^2} : u_{i,l,m+q} = 0\right\}, \\
\tau_m^+ &= \min\left\{\frac{-2\log U_{i,l,m+q}}{(x_{i,m+q} - \mu_{D_{i,l},m})^2} : u_{i,l,m+q} = 1\right\}.
\end{aligned}$$

Finally, we sample each $\mu_{j,h}$ independently from a truncated normal distribution,

$$f(\mu_{j,m} \mid \cdots) \propto \mathrm{N}\left(\mu_{j,m} \mid \hat{\mu}_{j,m}, (\tau_m \hat{s}_{j,m})^{-1}\right)\mathbf{1}\left\{\mu_{j,m} \in \bigcap_{D_{i,l}=j} A_{i,l,m}\right\},$$

where

$$\begin{aligned}
\hat{s}_{j,m} &= s_m + n_j; \\
\hat{\mu}_{j,m} &= \frac{1}{\hat{s}_{j,m}}\left(s_m\mu_{0,m} + \sum_{d_i=j} x_{i,m+q}\right).
\end{aligned}$$

The truncation is defined through the intervals

$$I_{i,l,m} = \left(x_{i,m+q} - \sqrt{\frac{-2\log U_{i,l,m+q}}{\tau_m}},\ x_{i,m+q} + \sqrt{\frac{-2\log U_{i,l,m+q}}{\tau_m}}\right),$$

letting $A_{i,l,m} = I_{i,l,m}$ when $u_{i,l,m+p} = 1$; and $\mathbb{R} \setminus A_{i,l,m} = I_{i,j,m}$ when $u_{i,l,m+p} = 0$.

4.3.5 Updating the Latent Model Dimension, $\mathbf{k}_{1:n}$

As we have done in the previous chapters for the update of each k_i, we use the ideas of Godsill (2001), presented in Section 1.2.2.3 to deal with the change of

dimension in the sampling space. We start by proposing a move from k_i to $k_i + 1$ with probability $1/2$, and accepting it with probability

$$\min\left\{1, \left[1 - K(x_i|\psi_{D_{i,k_i+1}})\right]\right\}.$$

The evaluation of this expression requires the sampling of the additional index D_{i,k_i+1}, and we choose $D_{i,k_i+1} = j$ with probability w_j.

Similarly, if $k_i > 0$, a move from k_i to $k_i - 1$ is proposed with probability $1/2$, and accepted with probability

$$\min\left\{1, \left[1 - K(x_i|\psi_{D_{i,k_i}})\right]^{-1}\right\}.$$

Thus, we have shown it is possible to perform posterior inference for the nonparametric regression model proposed, via an MCMC scheme applied to the latent model. We have successfully implemented the method in Matlab (R2012a), and present some results in the next Section.

4.4 Illustrations.

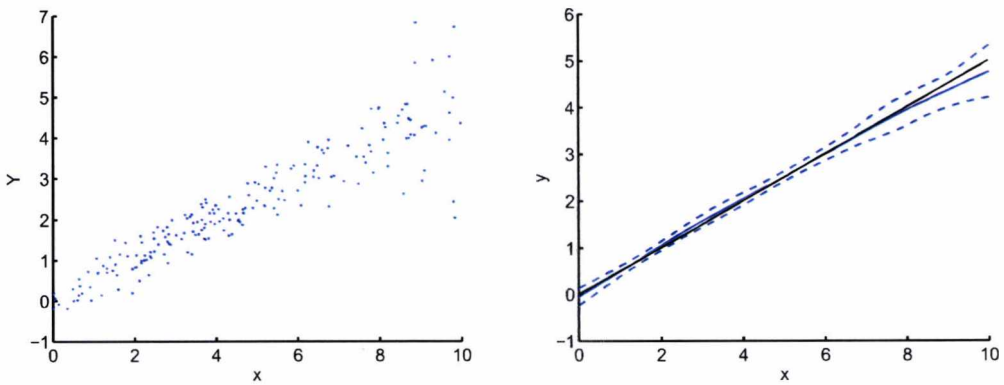
4.4.1 Example 1: Non-Linear Variance

In many situations, the error distribution for the variable y, with respect to the mean regression line, may evolve with x. We consider such a situation in the following example, where $n = 200$ data points (displayed in Figure 4.1) are simulated assuming a linear mean function non-linear increasing variance;

$$\begin{aligned} x_i &\overset{iid}{\sim} \mathrm{U}(x|0, 10), \\ y_i|x_i &\overset{ind}{\sim} \mathrm{N}\left(\frac{1}{2}x_i, \frac{1}{4} + \exp\left(\frac{x_i - 10}{2}\right)\right). \end{aligned}$$

The covariate dependent weights for our model are given by

$$w(x; \theta_j) = \frac{w_j \exp\{-\tau/2(x - \mu_j)^2\}}{\sum_{j'=1}^{\infty} w_{j'} \exp\{-\tau/2(x - \mu_{j'})^2\}}.$$



(a) Simulated data (b) Predicted regression function

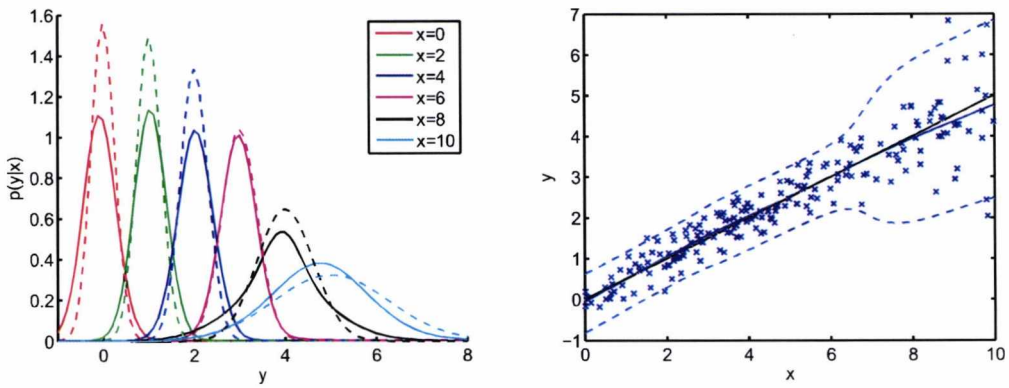
Figure 4.1: The data with y plotted against x on the left. On the right, the predicted regression function for a grid of x values (blue solid line); 95% pointwise credible intervals (blue dashed lines); and the true regression function (in black).

The prior parameters for the weights $(w_j)_{j\geq 1}$ are fixed at $\alpha_j = 1$ and $\zeta_j = 1$, for all j, corresponding to a Dirichlet process. The hyperparameters for the base measure for (θ_j, ψ_j), are selected as

$$\beta_0 = (0, 1/2)'; \quad \Sigma^{-1} = \text{diag}(10, 1/4); \quad \tilde{a} = 1; \quad \tilde{c} = 1;$$
$$\mu_0 = 5; \quad s = 1/4; \quad a = 1; \quad c = 1.$$

We generate a Monte Carlo sample of size 5,000 iterations with a burn in period of 5,000. The initial states assigns one component to each observation, with parameters generated from the prior. The right hand plot in Figure 4.1 depicts the estimated regression function for a grid of covariate values (blue solid line) and 95% pointwise credible intervals (blue dashed lines). The true regression function, shown in black, is a simple linear function, and the model recovers it well.

Predictive densities were estimated for all covariate values in the grid. The plot on the right of Figure 4.2 displays the predictive density estimates $f(y|x)$ for $x = 0, 2, 4, 6, 8, 10$. On the left, we show point and 95% credible interval estimates for y. For high values of the covariate, the variance of the conditional density is slightly overestimated; while for small covariate values, the transition



(a) Predictive densities

(b) Data and prediction

Figure 4.2: The predictive densities, for $x = 0, 2, 4, 6, 8, 10$, with solid lines denoting the prediction and dashed lines denoting the true density, are shown on the left. The right side plot presents the data, the prediction and 95% credible intervals computed from the predictive densities.

density mode is slightly underestimated. However, the general dynamics of the variance function are well captured. Furthermore, the 95% credible intervals for $y|x$ contain the observations and seem to accurately reflect the information present in the data.

Observing Figure 4.1 could lead one to believe that all subjects are assigned to the same component with a high posterior probability. However, there is a more complex aspect to this example; the variance of the error distribution increases with x. In fact, in order to capture this feature, most of the posterior samples of the component allocation, group the data into three clusters. The configuration with the highest posterior probability is depicted in Figure 4.3.

4.4.2 Example 2: Non-Linear Regression Curve

To demonstrate the ability of the model to recover complex regressions functions with the presence of both continuous and discrete covariates, we simulate $n = 200$ data points (depicted on the left of Figure 4.4) through the following formulas,

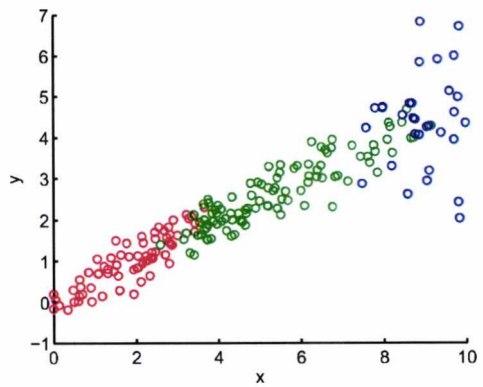


Figure 4.3: The configuration with the highest posterior probability, where the data are coloured by component membership.

$$x_{i,1} \overset{iid}{\sim} \text{Bern}(1/2), \quad x_{i,2} \overset{iid}{\sim} \text{U}(\cdot|-5,5),$$
$$y_i|x_i \overset{ind}{\sim} \text{N}\big([\mathbf{1}\{x_{i,1}=1\} - \mathbf{1}\{x_{i,1}=0\}]\, x_{i,2}^2,\, 1\big).$$

Our model is given by

$$f_P(y|x) = \sum_{j=1}^{\infty} w(x;\theta_j)\text{N}(y|X\beta_j, \sigma_j^2),$$
$$w(x;\theta_j) = \frac{w_j\, \rho_{j,0}^{1\{x_1=0\}}\, \rho_{j,1}^{1\{x_1=1\}} \exp\{-\tau/2(x_2-\mu_j)^2\}}{\sum_{j'=1}^{\infty} w_{j'}\, \rho_{j',0}^{1\{x_1=0\}}\, \rho_{j',1}^{1\{x_1=1\}} \exp\{-\tau/2(x_2-\mu_{j'})^2\}}.$$

We use the prior for the weights $w_{1:\infty}$ and parameters $\theta_{1:\infty}$, $\psi_{1:\infty}$ described in Section 4.3. Specifically, we define a Dirichlet Process prior with unit mass and set the following hyperparameters for the base measure,

$$\beta_0 = (12.5, -25, 0)'; \quad \Sigma^{-1} = \text{diag}(50, 150, 25); \quad \tilde{a} = 1; \quad \tilde{c} = 1;$$
$$\gamma = (1,1)'; \quad \mu_0 = 0; \quad s = 1/4; \quad a = 1; \quad c = 1.$$

Inference is carried out via MCMC posterior simulation, using the latent model representation and the algorithm discussed in the present Chapter. We generate a Monte Carlo sample of size 5,000 iterations after a burn in period of 5,000. The right side of Figure 4.4 depicts the predicted regression function for a grid of

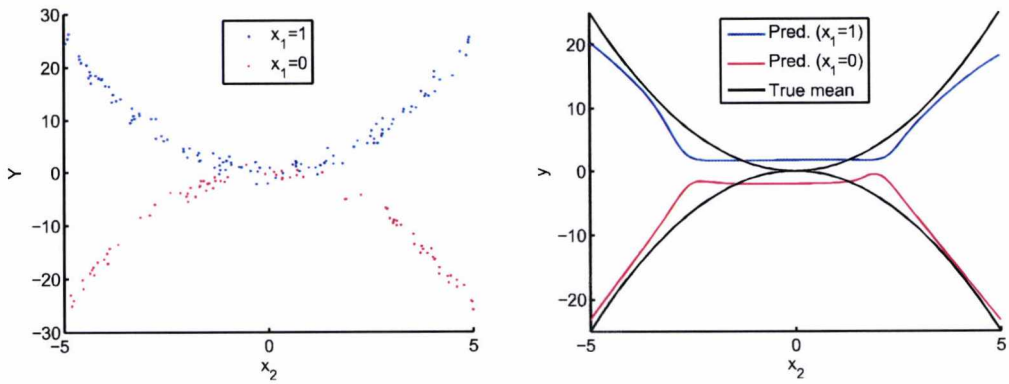
x_2 values with $x_1 = 0$ in blue and $x_1 = 1$ in green. The true regression function is shown in black. Even though the true function has a peculiar shape, the model is able to recover it. This flexibility in estimating the regression function relies heavily on the posterior distribution of the covariate dependent weights. The left panel of Figure 4.5 depicts the the configuration with highest estimated posterior probability, with data points coloured by component membership. The right panel of Figure 4.5 plots a posterior sample of the covariate-dependent weights, given this configuration, as a function of x_2. Solid lines denote the case when $x_1 = 1$ and dashed lines indicate when $x_1 = 0$. It is important to observe that *aposteriori* the weights are able to peak close to one in areas of high applicability of their associated linear regression models, and decay smoothly or sharply, as needed, when the covariates move away from this area.

4.4.3 Example 3: Alzheimer's Disease Study

Alzheimer's disease (AD) is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks (ADEAR, Alzheimer's Disease Education & Referral Center, 2011). Due to its damaging effects and increasing prevalence, it has become a major public health concern, more so amongst populations with increasing life expectancy. Thus, early and differential diagnosis, as well as disease-modifying drugs or therapies are in great need.

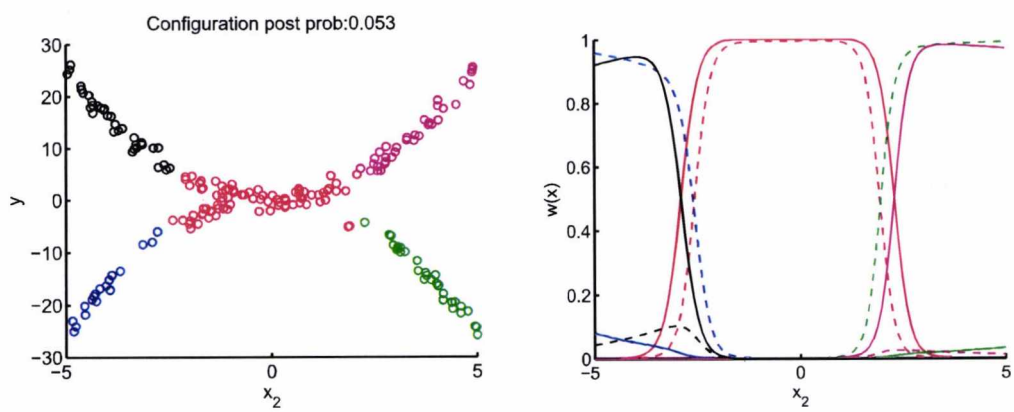
In a clinical trial setting, with the purpose of assessing the effectiveness of any proposed drugs or therapies, accurate tools for diagnosis, disease-staging, and monitoring disease progression are needed. Unfortunately, definite diagnosis requires histopathologic examination of brain tissue, an invasive procedure typically only performed at autopsy.

Non-invasive methods can be used to produce neuroimages and biospecimens which provide evidence of some changes in the brain associated with AD. Moreover, biomarkers based on neuroimaging or biological data may present a higher sensitivity to changes due to drugs or therapies, and over shorter periods of time, making them better suited tools than clinical measures for disease staging and monitoring disease progression in clinical trials.



(a) Simulated data (b) Predicted regression function

Figure 4.4: The left panel depicts the data with y plotted against x_2. The data are coloured by x_1. The right panel depicts the true regression function (black line) for a grid of covariate values; the red and blue lines represent the predicted function for $x_1 = 0$ and $x_1 = 1$ respectively.



(a) Partition with highest prob. (b) Covariate-dependent weights

Figure 4.5: The left panel depicts the partition with the highest posterior probability, where the data are colored by component membership. The right panel depicts the covariate-dependent weights associated to this partition with solid lines representing $w_j(1, x_2)$ and dashed lines representing $w_j(0, x_2)$ for a grid of x_2 values.

However, before biomarkers based on neuroimaging or biological data can be useful in clinical trials, their evolution over time needs to be well understood. Those which change earliest and fastest should be used for diagnosis or as inclusion criteria for the trials; those which change the most in the disease stage of interest should be used for disease monitoring; and all should be combined to assess the disease stage of an individual.

In a recent paper, Jack *et al.* (2010), propose a theoretical model for the evolution of the five most widely studied and well validated biomakers. Their model assumes that biomarkers become abnormal in a time ordered manner, with a sigmoidal path that varies in steepness across biomarkers. Frisoni *et al.* (2010) discuss the model in further detail, focusing on the evolution of biomarkers based on structural Magnetic Resonance Images (sMRI). Recent studies support this theory. Caroli & Frisoni (2010) and Sabuncu *et al.* (2011) assess the fit of parametric sigmoidal curves, and Jack *et al.* (2012) consider a more flexible model based on additive cubic splines with three chosen knot points. This last approach is the most flexible among the three, but they all impose significant restrictions which raise doubts about their conclusions. It is arguably not enough to provide evidence that one fit is better than another when only a limited number of curves can be compared. Ideally, a more flexible model should be able to choose the shape of the regression curve that better fits the data.

The clinical stages of the AD are divided into three phases (Jack *et al.*, 2010); the pre-symptomatic phase, prodromal phase, and the dementia phase. During the pre-symptomatic phase, some AD pathological changes are present, but patients do not exhibit clinical symptoms. This phase may begin possibly 20 years before the onset of clinical symptoms. The pre-prodromal stage of AD is known as mild cognitive impairment (MCI); patients diagnosed with MCI exhibit early symptoms of cognitive impairment, but do not meet the dementia criteria. The final stage of AD is dementia, when patients are officially diagnosed with AD.

Hippocampal volume is one of the best established and most studied biomarkers because of its known association with memory skills and relatively easy identification in sMRI. Jack *et al.* (2010) and Frisoni *et al.* (2010) hypothesized that hippocampal volume evolves sigmoidally over time, with changes starting slightly before the MCI stage and occurring until late in dementia phase. The steepest

changes are supposed to occur shortly after the dementia threshold has been crossed.

To provide validation for this model, we study the evolution of hippocampal volume as a function of age, gender, and disease status. Data was obtained from the Alzheimer's Disease Neuroimaging Initiative database which is publicly accessible at UCLA's Laboratory of Neuroimaging. The ADNI database contains neuroimaging, biological, and clinical data, along with summaries of neuroimages, including the volume of various brain structures. The dataset analysed here consists of the hippocampal volume obtained from the sMRI performed at the first visit, for 736 patients. Of the 736 patients in our study, 159 have been diagnosed with AD, 357 have MCI, and 218 are cognitively normal (CN). Figure 4.6 displays the data.

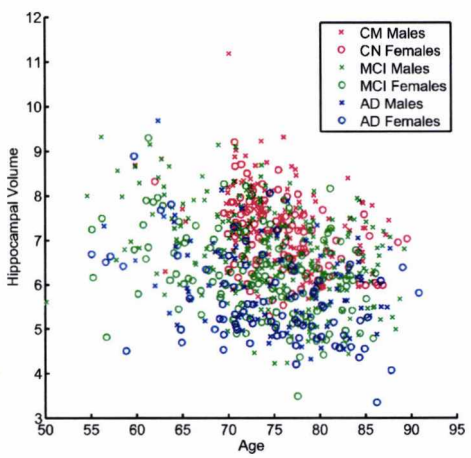


Figure 4.6: Hippocampal volume plotted against age. The data are colored by disease status with circles representing females and crosses representing males.

We consider the model developed in this Chapter, specifically, the infinite Gaussian kernel mixture model with covariate dependent weights given by

$$w(x;\psi_j) = \frac{w_j \prod_{m=1}^{2} \prod_{g=0}^{G_m} \rho_{j,m,g}^{\mathbf{1}(x_m=g)} \exp\{-\tau/2(x_3-\mu_j)^2\}}{\sum_{j'=1}^{\infty} w_{j'} \prod_{m=1}^{2} \prod_{g=0}^{G_m} \rho_{j',m,g}^{\mathbf{1}(x_m=g)} \exp\{-\tau/2(x_3-\mu_{j'})^2\}},$$

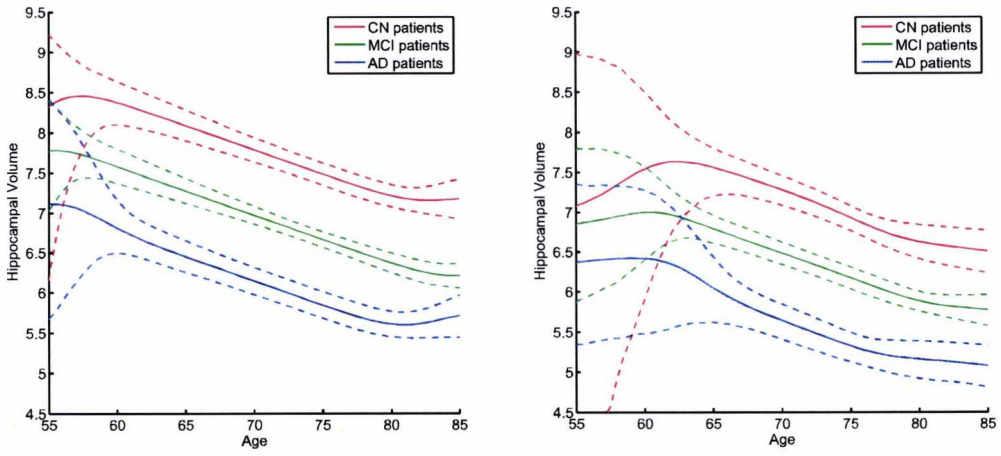
where $G_1 = 1$ (x_1 represents gender) and $G_2 = 2$ (x_2 represents disease status).

Note that here age (x_3) is a real number, measuring the time from birth to exam date, and is therefore treated as a continuous covariate.

The prior distribution for $(w_j)_{j\geq 1}$ and $(\theta_j, \psi_j)_{j\geq 1}$ is described in Section 4.3. We use a Dirichlet Process prior with unit mass parameter and set the hyperparameters for the base measure as

$$\beta_0 = (8, -1, -1, -1/4)'; \quad \Sigma^{-1} = \text{diag}(4, 1/4, 1/4, 1/60); \quad \tilde{a} = 1; \quad \tilde{c} = 1;$$
$$\gamma_1 = (1, 1)'; \quad \gamma_2 = (1, 1, 1)'; \quad \mu_0 = 72.5; \quad s = 1/4; \quad a = 1; \quad c = 1.$$

Inference is carried out via MCMC posterior simulation with a Monte Carlo sample size of $N = 5,000$ iterations after a burn in period of $5,000$.



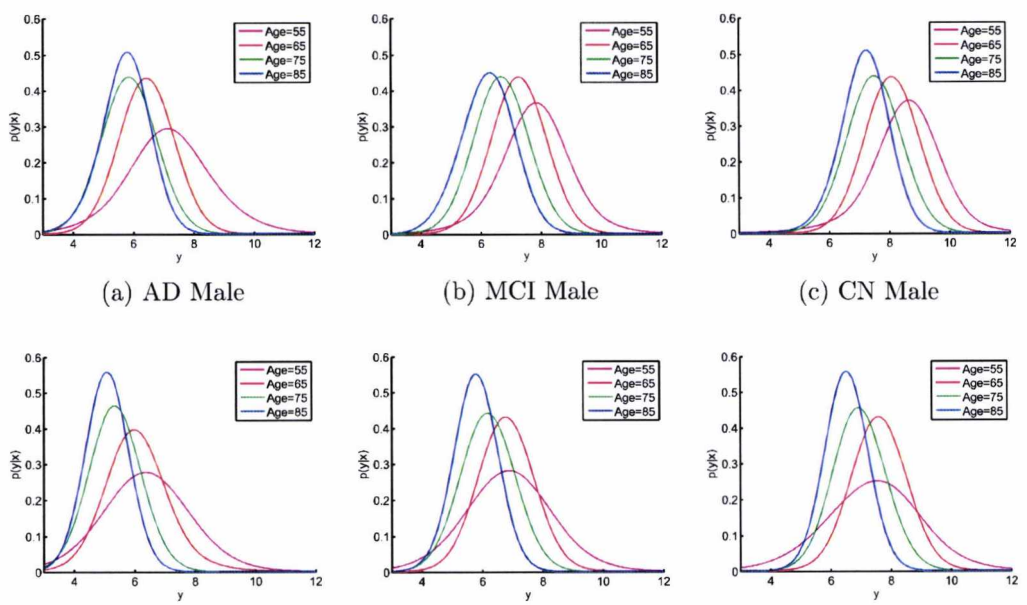
(a) Male patients

(b) Female patients

Figure 4.7: Predicted hippocampal volume as a function of age, disease, and sex. The data are colored by disease status with dashed lines representing 95% pointwise credible intervals around the predictive function.

Figure 4.7 displays the estimated mean regression function for a grid of ages with all possible combinations of disease status and sex. Interestingly, we observe a confirmation of the hypothesized sigmoidal evolution of hippocampal volume with increasing age. Cognitively normal subjects are predicted to have highest values of hippocampal volume at all ages, and MCI patients are predicted to have higher values of hippocampal volume at all ages when compared with AD patients.

This indicates that hippocampal volume may be useful in disease staging during both the MCI and AD phases. Notice that, as expected, females are predicted to have lower values of hippocampal volume, but the decline is predicted to start with a lag of approximately five years when compared to males. We should comment that there is no data for the subgroup of CN females under 60, which reflects on the greater uncertainty in the estimation.



(a) AD Male (b) MCI Male (c) CN Male

(d) AD Female (e) MCI Female (f) CN Female

Figure 4.8: Conditional density estimates for new covariates with ages of 55, 65, 75, and 85 and all combinations of disease status and sex.

Figure 4.8 displays the predictive density estimates given a new set of covariate values, with ages of 55, 65, 75 and 85, and all combinations of disease status and sex. In a clinical trial setting, the preference is for reliable outcome measures, i.e. biomarkers with small variability. In general, we observe that variance decreases with increasing age, indicating that hippocampal volume is more reliable for elderly patients. The difference is more extreme for females as opposed to males. In particular, hippocampal volume is predicted to have a large variability for young females across all disease stages, with the largest for young CN females,

but this may not be reliable due to the lack of data for this group. Instead, for older females, the variance is much smaller for all disease stages. When comparing males across disease status, we notice that young AD patients are predicted to show a large variability compared with young MCI and CN patients, while old MCI patients are predicted to show the largest variability when compared with their CN and AD counterparts.

4.5 Discussion

In this Chapter, we have developed a novel Bayesian nonparametric regression model based on normalized covariate dependent weights. The interpretability of the construction is not its only relevant feature. The empirical analysis in the Illustrations Section shows a great flexibility of the underlying clustering structure induced by the model. Specifically, the posterior distribution of the the weights allows both a sharp or a smooth placement of the mixture components throughout the covariate space, depending on the information contained in the data. This allows the model to recover complex shapes for the regression mean function, as well as nonlinear dependence of the error distribution on the covariates.

We have illustrated the applicability of the model and latent variable construction for combinations of discrete and continuous covariates, but focusing on a univariate continuous response. However, the model can easily be extended for other types of response variables, simply by changing the choice of parametric kernels in the mixture. For example, the simple regression kernel may be replaced by a generalized linear regression model. Future work would involve studying the properties of more general models. In particular, the Alzheimer's disease study is an interesting application that could benefit from a more general setting.

Notes and acknowledgements for ADNI data

Data used in Section 4.4.3 were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI provided data but did not participate in analysis or writing of this. A complete listing of ADNI investigators can be found

at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $ 60 million, 5-year public- private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org

Data collection and sharing for the Alzheimer's disease study was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research

& Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at theUniversity of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

The ADNI database contains many datasets including a file containing the left and right hippocampal volume and exam date for the sMRI (USCDVOL.csv), demographic information including date of birth and sex (PTDEMOG.csv), and diagnostic information (ARM.csv). Preprocessing of the data involved the merger of these datasets. In addition, total hippocampal volume was calculated as the sum of the left and right hippocampal volume, and age at exam date was calculated in fractions of years based on the date of birth and the date of exam. Variable selection and preprocessing of the data was done by Sara Wade.

Chapter 5

The Power Likelihood

The progress of Bayesian nonparametric methods (see Hjort *et al.*, 2010) has led to a surge in theory involving posterior consistency. In Section 1.3 we present some of the existing results regarding Bayesian consistency. While infinite dimensional Bayesian models provide flexible models, useful in many real life situations, they may fail to be strongly consistent, as demonstrated by the famous counterexample of Barron *et al.* (1999), described in Section 1.3.3.

Sufficient conditions for strong Bayesian consistency are, in general, not easily verified. However, as pointed out by Walker & Hjort (2001), the use of a power likelihood guarantees consistency with only the Kullback–Leibler support property. That is, for any $\alpha \in (0,1)$,

$$Q_n(A_\varepsilon) = \frac{R_n^{1-\alpha}(f)\,\Pi(\mathrm{d}f)}{\int_\Omega R_n^{1-\alpha}(f)\,\Pi(\mathrm{d}f)}$$

is such that

$$Q_n(A_\varepsilon) \to 0 \quad \text{a.s.}$$

for all $\varepsilon > 0$ and $A_\varepsilon = \{f : H(f, f_0) > \varepsilon\}$, a set of density functions bounded away from the true density in the Hellinger sense.

A popular model in Bayesian nonparametrics is the mixture of Dirichlet process model, introduced by Lo (1984) and based on the Dirichlet process of Ferguson (1973) (see section 1.1.1.2). We consider here a version of this model, for which the prior generates random density functions of the type

$$f_P(y) = \int K_\theta(y)\,\mathrm{d}P(\theta).$$

In particular, for illustrative purposes, we take $K_\theta = \mathrm{N}(y|\mu, \sigma^2)$, the normal density kernel, with $\theta = (\mu, \sigma^2)$; and P is a random distribution function taken from a Dirichlet process prior. However, the results can be extended to more general stick-breaking priors and parametric kernels.

Bayesian inference for this model, via MCMC methods, is now routine (see e.g. Escobar, 1988; Kalli *et al.*, 2011; MacEachern & Müller, 1998; Neal, 2000). Yet it is not clear that inference can be performed when using a power likelihood. In this case, we need to estimate

$$Q_n(\mathrm{d}f) \propto \Pi(\mathrm{d}f) \prod_{i=1}^{n} \left\{ \int K_\theta(\cdot)\, \mathrm{d}P(\theta) \right\}^{1-\alpha}.$$

The aim of this Chapter is to demonstrate how the general latent variable extension presented in this thesis can be applied to this power likelihood model, thus enabling posterior inference through MCMC simulation. Ideas and results are taken from Antoniano-Villalobos & Walker (2012b).

We do not claim here that inference based on a power likelihood would perform better than the correct Bayesian posterior inference. In fact, the quantity Q_n does not have a clear interpretation other than that of an approximate model (when α is small) which is consistent. The motivation to use the power likelihood is to implement an updating procedure which guarantees consistency for a particular Π without needing to check non–trivial conditions. Additionally, by using different values of $\alpha > 0$, and comparing the results with those obtained with $\alpha = 0$, the validity of the inference obtained for a given sample can be assessed empirically. This idea is explained later in this Chapter.

Furthermore, while our focus here is on consistency, this is not the only motivation for the use of a power likelihood. In general, raising the likelihood to some power smaller than 1 has a smoothing effect, which can be useful in some situations. See for example Friel & Pettitt (2008) for the use of power likelihood in the context of simulated annealing algorithms or Ibrahim & Chen (2000), also concerned with rising likelihoods to powers.

5.1 The Latent Model

As in the previous Chapters, our approach relies on the use of latent variables to define a latent model, which is marginally equivalent to the use of the power likelihood for the model of interest.

We wish to base inference on the power likelihood

$$f_P^{1-\alpha}(y_1 : n) = \prod_{i=1}^{n} f_P^{1-\alpha}(y_i).$$

There is no direct use for this expression, so we may use the stick–breaking representation for P,

$$P = \sum_{j=1}^{\infty} w_j \, \delta_{\theta_j},$$

to obtain an equivalent expression in which the $(1-\alpha)$ power is applied to objects bounded by 1,

$$f^{1-\alpha}(y_{1:n}|w_{1:\infty}, \mu_{1:\infty}, \sigma^2) = \sigma^{-n(1-\alpha)} \prod_{i=1}^{n} \left[\sum_{j=1}^{\infty} w_j \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu_j)^2 \right\} \right]^{1-\alpha}.$$

Here, the $(w_j)_{j\geq 1}$ are based on a sequence of independent and identically distributed $v_j \sim \text{Be}(1, \zeta)$ for some $\zeta > 0$. The $(\mu_j)_{j\geq 1}$ are independent and identically distributed from some distribution P_0 (see Sethuraman, 1994, or Section 1.1.1 for more details). Notice that we are considering a mixture over the means of the normal Kernels only, to keep notation simple. Therefore, the prior for the single variance parameter σ^2 is chosen independently.

We could remove the $(1-\alpha)$ directly using a power series expansion, since, for any $0 < b < 1$,

$$b^{1-\alpha} = \sum_{k=0}^{\infty} (-1)^k a_k (1-b)^k,$$

for some positive sequence $(a_k)_{k\geq 0}$. However, this is not convenient, as the resulting negative terms would invalidate the mixture model representation. On the other hand, we see that

$$b^{-\alpha} = \sum_{k=0}^{\infty} c_k (1-b)^k,$$

where the $(c_k)_{k\geq o}$ are all positive. In fact, $c_0 = 1$, $c_1 = \alpha$ and for $k > 1$,

$$c_k = \frac{\alpha^{(k)}}{k!} = \frac{\alpha(\alpha+1)\dots(\alpha+k-1)}{k!}.$$

Therefore, we can rewrite the power likelihood as

$$f_P^{1-\alpha}(y_{1:n}) = \prod_{i=1}^{n} f_P(y_i) \times f_P^{-\alpha}(y_i).$$

which is equivalent to

$$\sigma^{n\alpha} \prod_{i=1}^{n} f_P(y_i) \sum_{k_i=0}^{\infty} c_{k_i} \left[1 - \sum_{j=1}^{\infty} w_j \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right\}\right]^{k_i}.$$

As we have done in previous Chapters, we then consider $k_{1:n} = (k_1, \dots, k_n)$ as a latent variable and rearrange terms to obtain

$$\tilde{f}_P(y_{1:n}, k_{1:n}) \propto \sigma^{n\alpha} \prod_{i=1}^{n} f_P(x_i) c_{k_i} \left[1 - \sum_{j=1}^{\infty} w_j \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right\}\right]^{k_i}.$$

We have changed the notation to emphasize that, even though $f^{1-\alpha}$ is not a density, the latent expression $\tilde{f}$ is, thus the proportionality sign. This latent likelihood remains a complicated expression, but we can now introduce latent variables $D_{1:n,1:k_i} = (D_{i,l} : i = 1, \dots, n;\ l = 1, \dots, k_i)$ to substitute the term

$$\left[1 - \sum_{j=1}^{\infty} w_j \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right\}\right]^{k_i}.$$

by the latent expression

$$\prod_{l=1}^{k_i} w_{D_{i,l}} \left[1 - \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,l}})^2\right\}\right].$$

From which the desired term can be recovered by summing out the $D_{1:n,1:k_i}$ over the positive integers. Therefore, we now have the latent model

$$\tilde{f}_P(y_{1:n}, k_{1:n}, D_{1:n,1:k_i}) \propto \sigma^{n\alpha} \prod_{i=1}^{n} f_P(y_i)\ c_{k_i} \prod_{l=1}^{k_i} w_{D_{i,l}} \left[1 - \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,l}})^2\right\}\right].$$

Furthermore, the term $f_P(y_i)$ can be dealt with in the usual way, which involves introducing latent variables $d_{1:n} = (d_1, \ldots, d_n)$. and replacing $f_P(x_i)$ by the latent term

$$\sigma^{-1} w_{d_i} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{d_i})^2\right\}.$$

Hence, we arrive at the full latent model, given by

$$\tilde{f}_P(y_{1:n}, k_{1:n}, D_{1:n,1:k_i}) \propto \sigma^{-n(1-\alpha)} \prod_{i=1}^{n} c_{k_i} w_{d_i} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{d_i})^2\right\} \prod_{l=1}^{k_i} w_{D_{i,l}} \left[1 - \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,l}})^2\right\}\right].$$

It is easy to verify that summing over the latent variables, ($d_{1:n}$, $k_{i:n}$ and $D_{1:n,1:k_i}$, returns the $(1-\alpha)$ power likelihood, with P uniquely characterized by the mixture weights $(w_j)_{j\geq 1}$, the mixture means $(\mu_j)_{j\geq 1}$ and the variance term σ^2.

At this point, we are essentially ready to undertake inference for the power likelihood, via MCMC.

5.2 Posterior Inference via MCMC

The joint latent model is complemented by the prior for the mixing probability measure P. Together, they provide all the variables which need to be sampled for posterior estimation, i.e. the latent variables, ($d_{1:n}$, $k_{i:n}$ and $D_{1:n,1:k_i}$, the mixture weights and means, $(w_j)_{j\geq 1}$, the mixture means $(\mu_j)_{j\geq 1}$ and the variance term σ^2. This is achieved using an MCMC scheme with a Gibbs Sampler structure and Metropolis-Hastings steps.

5.2.1 Updating the Indices, d_i and $D_{i,l}$

There is still an issue due to the infinite state space of the indexing variables $d_{1:n}$ and $D_{1:n,1:k_i}$. We can deal with this in the same way we have done before, following Kalli *et al.* (2011). In order to reduce the state space for this variables to a finite set, we can introduce further auxiliary variables $\nu_{1:n}$ which interact with the $d_{1:n}$ in the joint model through the indicating functions

$$1\left(\nu_i < e^{-\xi d_i}\right),$$

for some $\xi > 0$. We can do the same for the $D_{1:n,1:k_i}$ by introducing

$$\mathbf{1}\left(\nu_{i,l} < e^{-\xi D_{i,l}}\right).$$

These variables then allow for finite choices and the easy sampling of the index variables. Hence, at each iteration of the MCMC algorithm, we need to sample each index from its full conditional distribution, given by

$$\mathbb{P}(d_i = j|\cdots) \propto w_j\, e^{\xi j} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right\} \mathbf{1}\,(1 \le j \le J_i);$$

$$\mathbb{P}(D_{i,l} = j|\cdots) \propto w_j\, e^{\xi j}\left[1 - \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right\}\right] \mathbf{1}\,(1 \le j \le J_{l,i}),$$

where $J_i = \lfloor -\xi^{-1} \log v_i \rfloor$ and $J_{l,i} = \lfloor -\xi^{-1} \log v_{l,i} \rfloor$. These values of $(J_i, J_{l,i})$ then tell us exactly how many of the mixture weights and means we need to sample. That is, at any given iteration of the MCMC algorithm, a sampler with the correct target distribution would only need to sample these for $j = 1, \ldots, J$, where $J = \max_{l,i}\{J_i, J_{l,i}\}$, since none of the variables involved in the update step depend on the values beyond J.

5.2.2 Updating the Mixture Weights, $\mathbf{w_{1:J}}$

We next describe how to sample the mixture weights $w_{1:J}$. As is well known these can be constructed from the independent beta distributed variables, $v_{1:J}$, making $w_1 = v_1$ and $w_j = v_j \prod_{j'<j}(1 - v_{j'})$ for $1 < j \le J$. The full conditional distribution for each v_j can easily be identified as

$$\tilde{f}(v_j|\cdots) = \mathrm{Be}(1 + n_j + N_j, \zeta + n_j^+ + N_j^+)$$

where

$$n_j = \sum_{i=1}^{n} \mathbf{1}\,(d_i = j); \quad N_j = \sum_{i=1}^{n}\sum_{l=1}^{k_i} \mathbf{1}\,(D_{i,l} = j);$$

$$n_j^+ = \sum_{i=1}^{n} \mathbf{1}\,(d_i > j); \quad N_j^+ = \sum_{i=1}^{n}\sum_{l=1}^{k_i} \mathbf{1}\,(D_{i,l} > j).$$

5.2.3 Updating the Kernel Variance, σ^2

Before proceeding, it is convenient to introduce new auxiliary variables $u_{i,l}$, for $i = 1, \ldots, n$ and $l = 1, \ldots, k_i$, which transform each product

$$\prod_{l=1}^{k_i} \left[1 - \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,l}})^2 \right\} \right]$$

into a truncation term,

$$\prod_{l=1}^{k_i} 1\left(u_{i,l} < 1 - \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,l}})^2 \right\} \right).$$

It is more convenient to work with the precision, $\tau = \sigma^{-2}$ and the full conditional distribution for τ is then given by

$$\tilde{f}(\tau | \cdots) \propto \Pi(\tau)\, \tau^{n(1-\alpha)/2} \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu_{d_i})^2 \right\} \mathbf{1}(\tau > T)$$

where

$$T = \max\left\{ \frac{-2\log(1-u_{i,l})}{(y_i - \mu_{D_{i,l}})^2} : i = 1, \ldots, n;\ l = 1, \ldots, k_i \right\}.$$

Hence, if $\Pi(\tau)$ is a Gamma distribution, then the conditional is a truncated Gamma distribution, for which numerous sampling routines are available. In particular, we use the latent variable approach described in Section 1.2.2.1.

5.2.4 Updating the Kernel Means, $\mu_{1:J}$

The sampling of the means $\mu_{1:J}$ for the Gaussian kernels is also not problematic. For each j, we must sample from the full conditional distribution

$$\tilde{f}(\mu_j | \cdots) \propto \Pi(\mu_j) \exp\left\{ -\frac{\tau}{2} \sum_{d_i = j} (y_i - \mu_j)^2 \right\} \mathbf{1}\left(\mu_j \in \cap_{i=1}^{n} A_{j,i}\right)$$

where $A_{j,i} = (-\infty, y_i - a_{j,i}] \cup [y_i + a_{j,i}, \infty)$,

$$a_{j,i} = \max\left\{ \sqrt{-2\tau^{-1}\log(1-u_{i,l})} : D_{l,i} = j;\ l = 1, \ldots, k_i \right\}$$

and $A_{j,i} = (-\infty, \infty)$ if $D_{l,i} \neq j$ for every l. So, if the prior is a Normal distribution, then the conditional here will be a truncated Normal distribution from which, once again, we may sample using, for example, the latent variable approach explained in Section 1.2.2.1.

5.2.5 Updating the Latent Model Dimension, $\mathbf{k_{1:n}}$

Finally, we need to describe how to update each k_i. Since the dimension of the sampling space changes with k_i, we use the Metropolis-Hasting approach of (Godsill, 2001). We propose a move from k_i to either $k_i + 1$ or $k_i - 1$, with probability $1/2$ each. The move from k_i to $k_i + 1$ is accepted with probability

$$\min\left\{1, \frac{c_{k_i+1}}{c_{k_i}}\left[1 - \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,k_i+1}})^2\right\}\right]\right\}.$$

On the other hand, the move from k_i to $k_i - 1$ is accepted with probability

$$\min\left\{1, \frac{c_{k_i-1}}{c_{k_i}}\left[1 - \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_{D_{i,k_i}})^2\right\}\right]^{-1}\right\}.$$

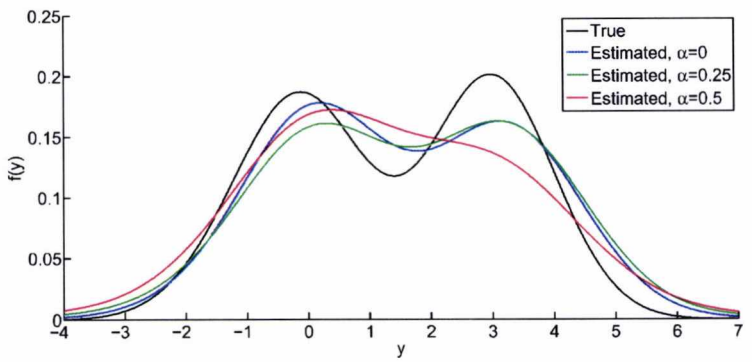
For the move upwards we need to allocate a value to D_{i,k_i+1}. Hence, we take $D_{i,k_i+1} = j$ with probability w_j. This can be implemented straightforwardly, paying special attention to the case when $k_i = 0$, for which we can only propose the move to $k_i + 1$ (and not to $k_i - 1$).

5.3 Illustrations

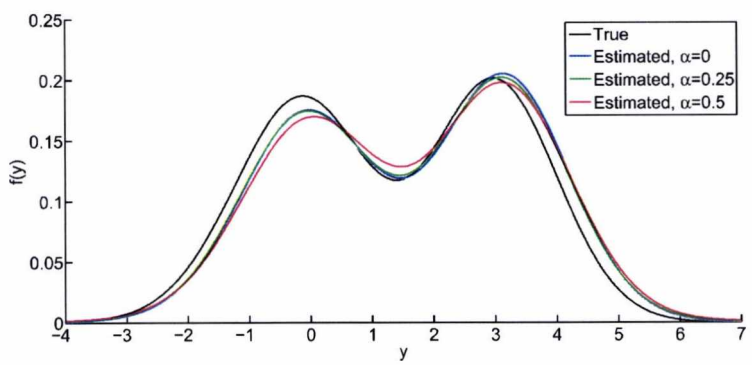
In this section we present some examples. The first one, where the data is simulated from a known density, illustrates the behavior of density estimates based on the power likelihood with different values of α, as the sample size increases. The second example involves density estimation for a real data set. In both cases, the MDP model used is known to be consistent. The third and last example shows how the density estimate obtained using the "true" likelihood ($\alpha = 0$) for an inconsistent model, diverges from those obtained using $\alpha > 0$ which are known to be consistent.

5.3.1 Example 1: Consistent Model

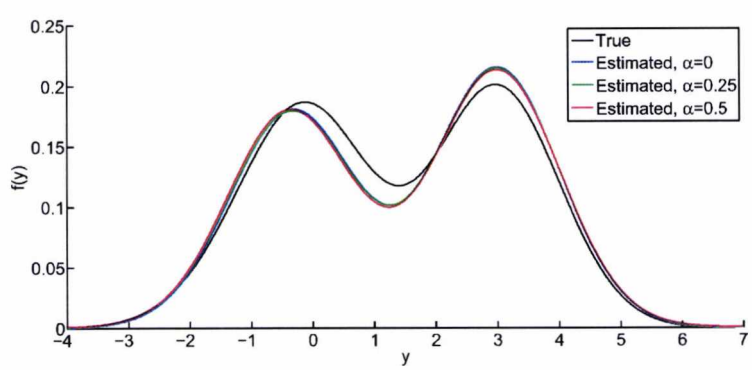
We consider a basic simulation set up. Observations are generated from a bimodal distribution, defined as a mixture of three normal components with means $\mu_1 = -1$, $\mu_2 = 0$ and $\mu_3 = 3$, with common variance $\sigma^2 = 1$, and weights $w_1 = 0.1$,



(a) $n = 10$



(b) $n = 100$



(c) $n = 1000$

Figure 5.1: Estimated predictive density based on $(1 - \alpha)$ power likelihood, for data simulated from the MDP model with three components, and increasing sample sizes.

$w_2 = 0.4$ and $w_3 = 0.5$, respectively. To describe the settings for the model and algorithm, we took $\xi = 0.1$ and the prior for the kernel means $(\mu_j)_{j\geq 1}$ was taken to be Normal with mean $m = 1.2$ (roughly in the mid-range of the data) and variance $t^{-1} = 10$. The purpose of this example is to illustrate the effect of increasing sample sizes on the density estimates obtained using the power likelihood with different values of α, when the model is consistent. Therefore, in order to eliminate any additional noise, we fixed the variance of the mixture components at the true value $\sigma^2 = 1$.

We estimated the posterior density for sample sizes of $n = 10, n = 100$ and $n = 1000$ observations, using the power likelihood, with $\alpha = 0.25, 0.5$ and $\alpha = 0$ (the "true" Likelihood). Each time, we used a Monte Carlo sample size of $N = 5,000$, after a burn–in period of $10,000$ iterations. Figure 5.1 shows the true density f_0 from which the data was generated, and MCMC estimates of the predictive density.

When $n = 10$, we can clearly see the smoothing effect of using $\alpha > 0$. However, since the model is consistent, as the sample size increases, all the estimated densities eventually merge, as the posterior accumulates around the true density.

5.3.2 Example 2: Real Data

Here we consider the galaxy data which consist of the velocities of 82 distant galaxies diverging from our own galaxy. Once again, we took $\xi = 0.1$ and the prior for the $(\mu_j)_{j\geq 1}$ was taken to be Normal with mean equal to the mid-range of the data and variance equal to the range. The prior for $\tau = 1/\sigma^2$ was chosen to be standard exponential and we defined a hyper–prior for ζ which is Ga$(0.5, 0.1)$.

Figure 5.2 shows a histogram of the data and the estimated predictive density for $\alpha = 0, 2/3000, 4/3000$ and $1/300$. It can be seen that, for values of α close to 0, the power likelihood density estimate approaches the density estimated using the "true" consistent model.

The first examples have served to illustrate that the use of the power likelihood does not highlight a discrepancy between the $\alpha = 0$ "true" model and those with $\alpha > 0$, which are known to be consistent. We now present an example where,

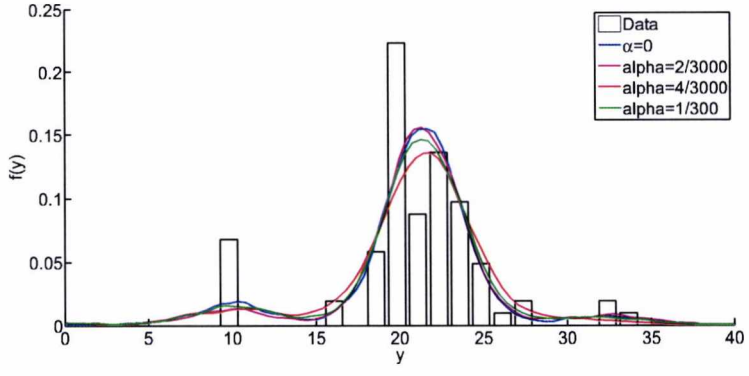


Figure 5.2: Galaxy Data: Estimated predictive density based on $(1-\alpha)$ power likelihood.

for a known inconsistent model, the $\alpha > 0$ models clearly highlight a discrepancy with the $\alpha = 0$ model and hence raise an issue as to the consistency of the "true" model.

5.3.3 Example 3: Inconsistent Model

The results found in the literature present conditions for consistency which are sufficient only. Therefore, in many cases, even when consistency for a model cannot be established, this does not imply it is inconsistent. Hence, if a model is chosen which does not satisfy these sufficient conditions, there would be some interest in diagnosing a possible case of inconsistency.

We study here the interesting example constructed by Barron *et al.* (1999) to show that posterior inconsistency can occur when nonparametric densities are involved. The inconsistent model is described in section 1.3.3. Recall that the idea is to construct a prior which assigns equal probability to a set $\mathcal{F}_0$ of continuous densities and a set $\mathcal{F}_*$ of piecewise constant densities. The role of the first set is to ensure the Kullback-Leibler property is satisfied. while the second ensures posterior probability does not accumulate almost surely on arbitrarily small Hellinger neighborhoods of the true density.

Both the prior and the posterior for this model are non parametric mixtures over the space of densities $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_*$. Therefore, a posterior sample can be

obtained using slice sampling techniques, introducing a latent variable to index the mixture component from which the density is sampled (see Kalli *et al.*, 2011, for details). When the latent variable takes the value 0, we obtain $f = f_\theta$ by sampling the parameter from the corresponding posterior density. In this case, the Hellinger distance to the true density is given by

$$H(f_\theta, f_0) = \sqrt{1 - \exp(-\theta/4)}.$$

When the latent variable takes a value $N > 0$ we know $f \in \mathcal{F}_N$ and the Hellinger distance, $H(f, f_0) = \sqrt{2 - \sqrt{2}}$ is constant. So we may calculate an MCMC estimate of the Hellinger distance between f_0 and a realization f from the posterior Π^n. Our results are shown in Figure 5.3. The horizontal axis corresponds to the sample size n, while the vertical axis shows the MCMC estimate of the Hellinger distance between f_0 and the predictive density, for different choices of α. We see that, for small n and small α, the behavior of the estimate is similar to that of $\alpha = 0$. However, for n large enough, all estimates obtained using $\alpha > 0$ approach the true density f_0, as expected from the consistency property. The estimated distance for $\alpha = 0$, on the other hand, remains constant for large n, since the "true" model is inconsistent.

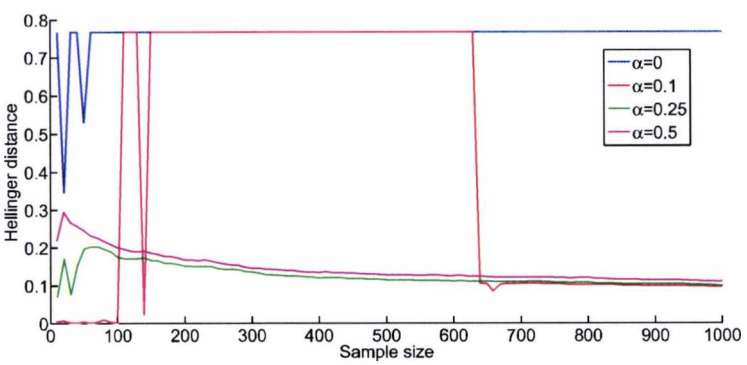


Figure 5.3: Inconsistent model: Estimated Hellinger distance between the true density f_0 and the estimated predictive density based on the $(1 - \alpha)$ power likelihood, for increasing sample size

Hence, in this example, if the model were not known to be inconsistent, the plots of posterior predictive distributions obtained using the true likelihood

($\alpha = 0$) would look different from those obtained with choices of $\alpha > 0$. This discrepancy would or should lead the practitioner to question the appropriateness of the model.

5.4 Discussion

With the mixture models now developing at a pace, both to multivariate versions and regression models, it is becoming harder to establish conditions for consistency. On the other hand there is no work to be done in this direction if one is willing to work with an $\alpha > 0$, however small.

While using a mixture model with a power less than 1 for the likelihood solves the problem of consistency, it brings up the issue of how to do Bayesian inference via MCMC. Furthermore, other motivations can be found for basing the inference on a smoothed version of the likelihood, obtained by raising it to some power smaller than one.

In this chapter, we have demonstrated how power likelihood based inference can be done for a nonparametric mixture model. The trick is to see the power likelihood as

$$R_n(f) \times R_n^{-\alpha}(f)$$

rather than

$$R_n^{1-\alpha}(f)$$

and to use a power series expansion for $b^{-\alpha}$, valid for any $0 < b < 1$, which is guaranteed to have positive weights. We have shown how the likelihood for a MDP model can be appropriately manipulated to ensure we obtain a quantity bounded by 1. There are other examples for which the principle holds. We can consider, for instance, an exponential model where $K_\theta(y) = \theta \exp(-y\theta)$. In this case, we can use

$$K_\theta(y) \propto \frac{\theta \exp(-y\theta)}{y^{-1}e^{-1}},$$

which is, again, bounded above by 1. All of this are examples of the general latent variable approach which is the driving theme of this thesis.

For a consistent model and small α there is no difference between using the likelihood raised to the power $(1 - \alpha)$ when compared with the "true" likelihood

($\alpha = 0$). Yet, for a choice of $0 < \alpha < 1$, the model is proved to be consistent (Walker and Hjort, 2001), therefore, for large enough n, the estimate should move away from the estimate produced by using $\alpha = 0$ if the model is inconsistent.

Since results for consistency involve conditions which are sufficient only, there are models for which consistency may be present but not theoretically verifiable. In such cases, we propose the use of the power likelihood for inference or for checking for discrepancies with the true model. Results can then be compared between $\alpha = 0$ and different values of $0 < \alpha < 1$. If density estimates are similar, the estimation produced by the "true" model may be considered adequate. However, if the estimates seem different, this may be considered as a warning sign that the model may not be consistent. Even for inconsistent models, the power likelihood may be used to produce a consistent estimate of the predictive density or to assess the quality of an estimate obtained using the "correct" likelihood, by comparison.

Chapter 6

Consistency for Markov Models

In this chapter we are concerned with the problem of Bayesian consistency for the transition density of time homogeneous Markov processes. To date, this remains somewhat an open problem, due to the lack of suitable metrics with which to work. Current results derive from generalizations of consistency results for the i.i.d. case and additionally require some non-trivial model assumptions. We propose a transformation of the Hellinger distance between join densities which does not define a metric in the space of transition densities, but is enough to define suitable neighbourhoods around the true transition. We derive conditions for posterior consistency which can be applied in general settings and show that, under reasonable assumptions, consistency with respect to such neighbourhoods is strong. In particular, we apply our result for consistency to a general family of nonparametric time series models. Results and illustrations are taken from Antoniano-Villalobos & Walker (2012a).

Consider an ergodic Markov process $Y = \{Y_n\}_{n\geq 0}$, defined on some separable filtered space $(\mathbb{Y}, \mathcal{A}, \{\mathcal{A}_n\}_{n\geq 0})$. Denote by $\mathbb{P}_0$ the true law of the process. Throughout this Chapter, all probability statements will be made with respect to $\mathbb{P}_0$.

Assume the process is time homogeneous and let f_0 be the true transition density for Y, with respect to some reference measure ν. Let ν_0 be the corresponding ergodic measure of the process. That is, for every $A \in \mathcal{A}$

$$\mathbb{P}_0[Y_{n+1} \in A | Y_n] = \int_A f_0(y|Y_n) \mathrm{d}\nu(y),$$

and for every integrable function $h \in L_1(\nu_0)$

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_n) \to \int h(y)\mathrm{d}\nu_0(y) \quad \text{a.s. when } n \to \infty.$$

In particular, if the process has a stationary density f_0, then the integral in the limit is equal to

$$\int h(y) f_0(y)\nu(\mathrm{d}y). \tag{6.1}$$

If f_0 is fixed but unknown. Bayesian inference begins by constructing a prior distribution Π over the class $\mathcal{F}$ of transition densities on $(\mathbb{Y}.\mathcal{A})$ with respect to the reference measure ν.

The predictive density for Y_{n+1}, given a sample $y_{0:n} = \{y_0, \ldots, y_n\}$, is

$$f_n(\cdot|y_n) = \mathbb{E}[f(\cdot|y_n)|y_{0:n}] = \int f(\cdot|y_n)\mathrm{d}\Pi^n(f).$$

where Π^n denotes the posterior probability given by

$$\Pi^n(A) = \frac{\int_A R_n(f)\mathrm{d}\Pi(f)}{\int R_n(f)\mathrm{d}\Pi(f)},$$

and

$$R_n(f) = \prod_{i=1}^{n} \frac{f(Y_i|Y_{i-1})}{f_0(Y_i|Y_{i-1})}$$

is the likelihood ratio. In order to simplify the notation, here and in the following, we assume that Y_0 is either a fixed known value y_0, or has a known initial distribution.

Accurate estimation of the transition density f_0 and the study of posterior consistency are therefore important in the context of prediction for Markov process models.

As in the case of consistency for i.i.d. observations, (Section 1.3), the general Markov process model is said to be consistent for the transition density f_0 if the posterior mass accumulates around f_0 as n increases. More formally, Π^n is consistent at f_0 if for every suitable neighbourhood B of f_0, we have

$$\Pi^n(B^c|Y_0, \ldots . Y_n) \xrightarrow[n\to\infty]{} 0 \quad [\mathbb{P}_0] - \text{a.s.} \tag{6.2}$$

Just as with densities, the concept of consistency for transition densities depends on the definition of the neighbourhoods. This poses a challenge, since the metrics and semimetrics used to define neighbourhoods in density spaces do not adapt to transition or conditional density spaces in a straightforward manner. The literature concerning consistency for Markov processes is therefore limited, due in great part to the difficulty in finding adequate topologies and distances between transition densities. A straightforward generalization of the Kullback-Leibler property for transition densities is possible through use of the the ergodic measure, as we explain in Section 6.2.3 below. However, it is not clear how the Hellinger distance or the L_1 metric can be generalized for the space of transition densities in a way that makes them useful for the search of sufficient strong consistency conditions.

To highlight the problem of extending the Hellinger distance between densities to the space of transition densities, consider, for a fixed $x \in \mathbb{Y}$ and two transition densities f_1 and f_2 in $\mathcal{F}$, the squared Hellinger distance between $f_1(\cdot|x)$ and $f_2(\cdot|x)$, given by

$$\begin{aligned} H^2\big(f_1(\cdot|x), f_2(\cdot|x)\big) &= \frac{1}{2}\int \left(\sqrt{f_1(y|x)} - \sqrt{f_2(y|x)}\right)^2 \mathrm{d}\nu(y) \\ &= 1 - \int \sqrt{f_1(y|x) f_2(y|x)}\, \mathrm{d}\nu(y). \end{aligned} \tag{6.3}$$

As it stands, H can not be used to define a topology on $\mathcal{F}$, as it depends on the current value of x. In order to adapt this and other quantities commonly used for densities, to define neighbourhoods in a space of transition densities, the dependence on x must somehow be eliminated.

A similar problem appears in the study of posterior consistency for regression models, where a distance between densities for the response variable y depends on the value of a covariate x. In this context, Ghosal & Roy (2006) and Choi & Schervish (2007) define a distance between two conditional densities f_1 and f_2 as

$$h(f_1, f_2) = \int H\big(f_1(\cdot|x), f_2(\cdot|x)\big) \mathrm{d}Q(x),$$

where Q is the distribution for the covariate. The definition of an adequate metric in terms of the Hellinger distance is due to the availability of the measure

Q when assuming the covariates are generated stochastically, i.i.d. from Q and independent of the response variable y. In the Markov process case, however, an adequate choice of integrating measure is unclear.

Tang & Ghosal (2007b) propose different ways of defining a topology on a transition density space. The first is based on the notion of distances on the invariant measures associated with each transition and results in a weak topology. Alternative ideas arise from using integrated and maximized distances between conditional densities respectively, resulting in strong types of neighbourhoods in both cases. In the same paper, the authors prove strong consistency in this sense for a specific family of transition densities based on Dirichlet mixtures, by generalizing the sieve and uniformly consistent tests approach to consistency for i.i.d observation (1.3 of the Background chapter). Ghosal & Tang (2006) extend this result to a general family $\mathcal{F}$ of transitions, providing it is compact with respect to the supremum Hellinger distance,

$$H_s(f_1, f_2) = \sup_x H\big(f_1(\cdot|x), f_2(\cdot|x)\big). \tag{6.4}$$

Compactness with respect to H_s is a very strong condition, as a simple example may show. Consider a simple normal AR(1) model and let $\mathcal{F} = \{N(\cdot|\theta x, 1) : \theta \in \Theta \subset \mathbb{R}\}$. The Hellinger distance between transition densities in this case is given by

$$H^2\big(f_\theta(\cdot|x), f_{\theta^*}(\cdot|x)\big) = 1 - \exp\left\{-\frac{1}{8}x^2(\theta - \theta^*)^2\right\}.$$

Therefore, $H_s\big(f_\theta, f_{\theta^*}\big) = 1$ for every $\theta \neq \theta^*$, so the required compactness is achieved only when Θ is finite; a rather restrictive condition for an already limited model.

Constructing an adequate sieve and proving the existence of a set of uniformly consistent tests is difficult in general. Therefore, in order to remove the compactness assumption, Ghosal & Tang (2006) leave this approach and instead generalize the martingale-based result of Walker (2003, 2004) (also presented in section 1.3). By assuming only the separability of $\mathcal{F}$, they are then able to prove consistency with respect to neighbourhoods of the type $\{f : \tilde{d}(f, f_0) < \epsilon\}$, where

$$\tilde{d}(f, f_0) = \inf_x H^2\big(f(\cdot|x), f_0(\cdot|x)\big). \tag{6.5}$$

Some families of transition densities can be found for which this type of consistency can be considered strong enough. In general, however, neighbourhoods like this correspond to a weak topology. Once more, we illustrate this through the simple AR(1) example. When $x = 0$, for every $\theta \in \Theta$ we have $f_\theta(\cdot|0) = N(\cdot|0, 1)$, yielding

$$\inf_x H^2\big(f_\theta(\cdot|x), f_{\theta^*}(\cdot|x)\big) = 0 \text{ for any } \theta, \theta^* \in \Theta,$$

and therefore $\tilde{d}$ does separate points in $\mathcal{F}$.

Ghosal & Tang (2006) mention this problem, which extends to the nonlinear autoregressive model $f(y|x) = g\big(y - \psi(x)\big)$, whenever g_0 is a location shift of g. Therefore, non trivial conditions must be imposed on a model if existing results are to be used to guarantee strong consistency.

Our main contribution in this Chapter is the definition of a system of neighbourhoods around the true transition density, f_0, based on a natural adaptation of the Hellinger distance between bivariate densities. Each transition density $f \in \mathcal{F}$ is extended to a family of bivariate densities. The distance between f and f_0 is defined as the smallest distance between sets of extended bivariate densities (to be explained in section 6.1). This, as we shall see, guarantees the definition of strong neighbourhoods around f_0, under reasonable conditions satisfied by general families of nonparametric models. Neighbourhoods are strong in the sense that they separate f_0 from other transition densities in $\mathcal{F}$, under conditions milder that those found in previous literature.

We then find sufficient conditions for consistency by extending the martingale result from Walker (2004), assuming only separability of $\mathcal{F}$ with respect to the supremum Hellinger distance H_s. We illustrate this through the problem of transition density estimation for a family of nonparametric dependent mixture models.

6.1 Strong Neighbourhoods

In this Section, we define an operator $d : \mathcal{F} \times \mathcal{F} \to [0, 1]$ and discuss the suitability of the neighbourhoods $B_\varepsilon = \{f \in \mathcal{F} : d\big(f, f_0\big) < \varepsilon\}$ in the study of posterior consistency for Markov processes.

Consider the set $\overline{\mathcal{F}}$ of bivariate densities on $(\mathbb{Y} \times \mathbb{Y}, \mathcal{A} \otimes \mathcal{A})$. Then, for every pair of functions $\bar{f}_1, \bar{f}_2 \in \overline{\mathcal{F}}$, the squared Hellinger distance between them is given by

$$H^2(\bar{f}_1, \bar{f}_2) = \frac{1}{2}\int \left(\sqrt{\bar{f}_1} - \sqrt{\bar{f}_2}\right)^2 \mathrm{d}(\nu \times \nu) = 1 - \int\int \sqrt{\bar{f}_1 \bar{f}_2}\, \mathrm{d}(\nu \times \nu).$$

For each value $x \in \mathbb{Y}$, and transition density $f \in \mathcal{F}$, a density $\bar{f}$ in $\overline{\mathcal{F}}$ is defined by

$$\bar{f}(z, y|x) = f(z|y)f(y|x) \quad \forall\, (z, y) \in \mathbb{Y} \times \mathbb{Y}.$$

Therefore, for any $f_1, f_2 \in \mathcal{F}$, we can define

$$d(f_1, f_2) = \inf_x d_x(f_1, f_2), \tag{6.6}$$

where d_x denotes the Hellinger distance between the two corresponding bivariate densities in $\overline{\mathcal{F}}$, i.e.

$$\begin{aligned} d_x^2(f_1, f_2) &= H^2\big(\bar{f}_1(\cdot, \cdot|x), \bar{f}_2(\cdot, \cdot|x)\big) \\ &= 1 - \int\int \sqrt{f_1(z|y)f_1(y|x)f_2(z|y)f_2(y|x)}\, \mathrm{d}\nu(z)\mathrm{d}\nu(y). \end{aligned}$$

It is important to emphasize the definition of d in (6.6) differs from that of Ghosal & Tang (2006) (equation 6.5), in that the integral for the Hellinger distance in our definition is taken with respect to the product measure $\nu \times \nu$, and not with respect to ν. In other words, it minimizes the Hellinger distance between bivariate, rather than univariate densities. The effect of this can be best understood by revisiting our example involving the simple normal autoregressive model. If $f_\theta(\cdot|x) = N(\cdot|\theta x, 1)$ as before, the squared Hellinger distance between the univariate functions $f_\theta(\cdot|x)$ and $f_{\theta^*}(\cdot|x)$ is given by equation (6). The infimum, reached when $x = 0$ is $\tilde{d}(f_\theta, f_{\theta^*}) = 0$, even if $\theta \neq \theta^*$.

On the other hand, the squared Hellinger distance between the bivariate functions $\bar{f}_\theta(\cdot, \cdot|x)$ and $\bar{f}_{\theta^*}(\cdot, \cdot|x)$ is

$$d_x^2(f_\theta, f_{\theta^*}) = 1 - \frac{2}{\sqrt{4 + (\theta - \theta^*)^2}} \exp\left\{-\frac{x^2}{2}(\theta - \theta^*)^2\left[1 - \frac{2(1 - \theta\theta^*)}{4 + (\theta - \theta^*)^2}\right]\right\}.$$

therefore,

$$\inf_x d_x^2(f_\theta, f_{\theta^*}) = d_0^2(f_\theta, f_{\theta^*}) = 1 - \frac{2}{\sqrt{4 + (\theta - \theta^*)^2}}.$$

In this case, d actually defines a distance, so it follows that $d(f_\theta, f_{\theta^*}) > 0$ whenever $|\theta - \theta^*| > 0$.

The following lemma gives a condition under which d allows us to distinguish f_0 from any other transition density, even when it does not define a distance on $\mathcal{F}$. In other words, if the condition is satisfied, $B_\varepsilon = \{f \in \mathcal{F} : d(f, f_0) < \varepsilon\}$ defines a strong neighbourhood around f_0, for any $\varepsilon > 0$.

Lemma 4 *Assume $\mathcal{F}$ is such that for every $f \neq f_0$ in $\mathcal{F}$,*

$$\inf_x \int H^2\big(f(\cdot|y), f_0(\cdot|y)\big) f_0(y|x) \mathrm{d}\nu(y) > 0. \tag{6.7}$$

Then, for every $\varepsilon > 0$, $B_\varepsilon = \{f \in \mathcal{F} : d(f, f_0) < \varepsilon\}$ defines a strong neighbourhood around f_0 in the sense that

$$d(f, f_0) = 0 \Leftrightarrow f = f_0.$$

Proof It is clear that $f = f_0 \Rightarrow d(f, f_0) = 0$. Let $f \in \mathcal{F}$ and $x \in \mathbb{Y}$. We first observe that

$$H^2\big(f(\cdot|x), f_0(\cdot|x)\big) = 1 - \mathbb{E}_0\left[\sqrt{\frac{f(y|x)}{f_0(y|x)}} \,\middle|\, x\right],$$

where $\mathbb{E}_0$ denotes the expectation with respect to f_0. We may also observe that

$$d_x^2(f, f_0) = 1 - \mathbb{E}_0\left[\sqrt{\frac{f(z|y)f(y|x)}{f_0(z|y)f_0(y|x)}} \,\middle|\, x\right].$$

Now,

$$\begin{aligned}\mathbb{E}_0\left[\sqrt{\frac{f(z|y)}{f_0(z|y)}}\sqrt{\frac{f(y|x)}{f_0(y|x)}} \,\middle|\, x\right] &= \mathbb{E}_0\left[\mathbb{E}_0\left[\sqrt{\frac{f(z|y)}{f_0(z|y)}} \,\middle|\, y\right]\sqrt{\frac{f(y|x)}{f_0(y|x)}} \,\middle|\, x\right] \\ &= \mathbb{E}_0\left[\Big\{1 - H^2\big(f(\cdot|y), f_0(\cdot|y)\big)\Big\}\sqrt{\frac{f(y|x)}{f_0(y|x)}} \,\middle|\, x\right].\end{aligned}$$

Since $0 \leq 1 - H^2\big(f(\cdot|y), f_0(\cdot|y)\big) \leq 1$, from the Cauchy-Schwartz inequality and $\mathbb{E}_0\left[f(y|x)/f_0(y|x) \,|\, x\right] = 1$, the above expression leads to

$$\mathbb{E}_0\left[\sqrt{\frac{f(z|y)f(y|x)}{f_0(z|y)f_0(y|x)}} \,\middle|\, x\right] \leq \sqrt{1 - \mathbb{E}_0\Big[H^2\big(f(\cdot|y), f_0(\cdot|y)\big) \,|\, x\Big]}$$

and so

$$d_x^2(f, f_0) \geq 1 - \sqrt{1 - \mathbb{E}_0\Big[H^2\big(f(\cdot|y), f_0(\cdot|y)\big) \,\Big|\, x\Big]} \geq 0.$$

Thus,

$$d(f, f_0) = 0 \quad \Rightarrow \quad \inf_x \mathbb{E}_0\Big[H^2\big(f(\cdot|y), f_0(\cdot|y)\big) \,\Big|\, x\Big] = 0. \tag{6.8}$$

Condition (6.7) ensures this only happens when $f = f_0$, completing the proof.

A better understanding of this result derives from the realization that, for condition (6.7) to be violated, it is not enough that $f_0(\cdot|x)$ becomes a mass function when we take the infimum over x; at the same time its support must include only the values $y \in \mathbb{Y}$ for which $f(\cdot|y) \equiv f_0(\cdot|y)$. The following corollary enables the use of neihbourhoods based on d for the study of strong consistency for many Markov models found in the literature and presented as an illustration of our results in section 6.3.

Corollary 1 *Assume that for every transition density $f \in \mathcal{F}$, $f(\cdot|x)$ is a continuous function of x. If there exists a density function g with full support over $(\mathbb{Y}, \mathcal{A})$ such that*

$$\inf_x f_0(y|x) > \beta g(y) \quad \forall\, y \in \mathbb{Y}, \tag{6.9}$$

for some $\beta > 0$, then

$$d(f, f_0) = 0 \Leftrightarrow f = f_0.$$

Proof We only need to prove $d(f, f_0) > 0$ whenever $f \neq f_0$.

Let $f \neq f_0$ in $\mathcal{F}$ and define, for $\varepsilon > 0$,

$$A_\varepsilon = \{y : H^2\big(f(\cdot|y), f_0(\cdot|y)\big) > \varepsilon\}. \tag{6.10}$$

The space of density functions over $(\mathbb{Y}, \mathcal{A})$, to which $f(\cdot|y)$ and $f_0(\cdot|y)$ belong, is separable with respect to the Hellinger distance, so A_ε is a non empty set. It follows form the continuity of the transition densities (with respect to x), the full support of g and condition (6.9), that

$$\inf_x \mathbb{P}_0(A_\varepsilon|x) = \inf_x \int_{A_\varepsilon} f_0(y|x)\mathrm{d}\nu(y) > \beta \int_{A_\varepsilon} g(y)\mathrm{d}\nu(y) = \beta \mathbb{P}_g(A_\varepsilon) > 0.$$

Therefore

$$\begin{aligned}\inf_x \int H^2\big(f(\cdot|y), f_0(\cdot|y)\big) f_0(y|x)\mathrm{d}\nu(y) &> \inf_x \int_{A_\varepsilon} H^2\big(f(\cdot|y), f_0(\cdot|y)\big) f_0(y|x)\mathrm{d}\nu(y) \\ &> \varepsilon \inf_x \int_{A_\varepsilon} f_0(y|x)\mathrm{d}\nu(y) > 0.\end{aligned} \tag{6.11}$$

The result follows from Lemma 4, thus ending the proof.

Notice that

$$d_x^2(f_1, f_2) = H^2\big(f_1(\cdot|x), f_2(\cdot|x)\big) + \int H^2\big(f_1(\cdot|y), f_2(\cdot|y)\big)\sqrt{f_1(y|x)f_2(y|x)}\,\nu(\mathrm{d}y).$$

And using similar arguments to those in the proof of Lemma 4, we can see that

$$d_x^2(f_1, f_2) \leq 2H_s^2(f_1, f_2). \tag{6.12}$$

This inequality between the squared Hellinger distance on bivariate conditional densities and the supremum Hellinger distance on univariate conditional densities is useful for the consistency result in the next Section.

6.2 Posterior Consistency

In this Section we establish the basic notation (following the setup of Ghosal & Tang, 2006; Walker, 2004) and present the main Theorem regarding consistency.

6.2.1 Preliminaries and Notation

Let $y_{1:n} = (y_1, \ldots, y_n)$ denote a sample of size n from $\mathbb{P}_0$ (formally, from the restriction of $\mathbb{P}_0$ to $\mathcal{A}_n$). The likelihood ratio for a transition density $f \in \mathcal{F}$ is denoted by

$$R_n(f) = \prod_{i=1}^{n} \frac{f(y_i|y_{i-1})}{f_0(y_i|y_{i-1})}.$$

Let Π denote a prior on $\mathcal{F}$ and define the integrated likelihood ratio over a measurable subset $A \subset \mathcal{F}$ as

$$L_n = L_{nA} = \int_A R_n(f)\Pi(\mathrm{d}f).$$

The posterior mass assigned to A is then given by

$$\Pi^n(A) = \frac{L_n}{I_n}, \tag{6.13}$$

where $I_n = L_{n\mathcal{F}} = \int R_n(f)\Pi(\mathrm{d}f)$.

Finally, we define the bivariate predictive density, with posterior restricted to the set A as

$$f_{nA}(y, x|y_n) = \int_A f(y|x)f(y|y_n)\mathrm{d}\Pi_A^n(f); \quad (y, x) \in \mathbb{Y} \times \mathbb{Y},$$

where

$$\mathrm{d}\Pi_A^n(f) = \frac{\mathbf{1}(f \in A)\mathrm{d}\Pi^n(f)}{\int_A \mathrm{d}\Pi^n(f)}.$$

Below, we provide sufficient conditions for posterior consistency, following the martingale approach of Walker (2004) for the i.i.d. case. The equivalent, in this case, of the key identity (1.137) is:

$$\frac{L_{n+2}}{L_n} = \frac{f_{nA}(y_{n+2}, y_{n+1}|y_n)}{f_0(y_{n+2}|y_{n+1})f_0(y_{n+1}|y_n)}. \tag{6.14}$$

Notice that in this case, the ratio is defined with a step of size 2, while in the i.i.d. case a size 1 step is sufficient. Furthermore, $\{L_{2n}\}$ is a martingale with respect to $\{\mathcal{A}_{2n}\}$, since $\mathbb{E}_0[L_2|\mathcal{A}_0] = L_0 = \Pi(A)$ and for every $n \geq 1$, $\mathbb{E}_0[L_{2n+2}|\mathcal{A}_{2n}] = L_{2n}\mathbb{E}_0[L_{2n+2}/L_{2n}|\mathcal{A}_{2n}] = L_{2n}$. Analogously, $\{L_{2n+1}\}$ is a martingale with respect to $\{\mathcal{A}_{2n+1}\}$, since $\mathbb{E}_0[L_{2n+3}|\mathcal{A}_{2n+1}] = L_{2n+1}$ for every $n \geq 0$.

The posterior mass assigned to $A \subset \mathcal{F}$, given a sample of size n, is defined by the ratio (6.13) and the different results regarding posterior consistency found in the literature deal with the numerator and the denominator in this expression separately. We do the same here.

6.2.2 The Numerator

The following Lemma regards a general property, essential for the treatment of the numerator in equation (6.13).

Lemma 5 *For each* $n \geq 1$

$$\mathbb{E}_0\Big[\sqrt{L_{n+2}}\,\Big|\,\mathcal{A}_n\Big] \leq \sqrt{L_n}\,\Big[1 - d^2_{y_n}\big(f_{nA}, f_0\big)\Big].$$

Proof Notice that L_n is $[\mathcal{A}_n]$-measurable, so

$$\frac{\mathbb{E}_0\big[\sqrt{L_{n+2}}\,\big|\,\mathcal{A}_n\big]}{\sqrt{L_n}} = \mathbb{E}_0\left[\sqrt{\frac{L_{n+2}}{L_n}}\,\middle|\,\mathcal{A}_n\right].$$

Applying the identity (6.14), and rearranging terms, we obtain

$$\mathbb{E}_0\Big[\sqrt{L_{n+2}}\,\Big|\,\mathcal{A}_n\Big] = \sqrt{L_n}\,\mathbb{E}_0\left[\sqrt{\frac{f_{nA}(y_{n+2}, y_{n+1}|y_n)}{f_0(y_{n+2}|y_{n+1})f_0(y_{n+1}|y_n)}}\,\middle|\,\mathcal{A}_n\right].$$

By applying the definition of d_x for $x = y_n$, we arrive at

$$\mathbb{E}_0\Big[\sqrt{L_{n+2}}\,\Big|\,\mathcal{A}_n\Big] = \sqrt{L_n}\,\Big[1 - d^2_{y_n}\big(f_{nA}, f_0\big)\Big].$$

This completes the proof.

Consider a set A of transition densities. If we assume $\mathcal{F}$ is separable with respect to some distance d^*, then for every $\delta > 0$, we can find a d^*-cover for A of size δ. That is, a collection $\{A_j\}_{j\geq 1}$ such that

$$A \subseteq \bigcup_{j=1}^{\infty} A_j$$

and for each j there exists $f_j \in A$ for which

$$A_j = \{f : d^*(f, f_j) < \delta\}.$$

Lemma 6 *Let* $A_\varepsilon \subset \mathcal{F}$ *be a set of transition densities* d*-bounded away from* f_0*,*

$$A_\varepsilon = \{f \in \mathcal{F} : d\big(f, f_0\big) > \varepsilon\}.$$

Assume $\mathcal{F}$ *is separable with respect to the supremum Hellinger distance,* H_s *and that*

$$\sum_{j=1}^{\infty} \sqrt{\Pi(A_j)} < \infty, \tag{6.15}$$

for some H_s-cover , $\{A_j\}_{j\geq 1}$ for A_ε, of size $\delta < \varepsilon/\sqrt{2}$.

Then, for some $b > 0$

$$\sum_{j=1}^{\infty} \sqrt{L_{nA_j}} < \exp(-nb),$$

$[\mathbb{P}_0]$-a.s. for all n sufficiently large.

Proof Let $\{A_j\}_{j\geq 1}$ be a cover satisfying assumption (6.15). Let $\gamma = \varepsilon - \sqrt{2}\delta > 0$. For simplicity, denote $L_{nj} = L_{nA_j}$ and $f_{nj} = f_{nA_j}$.

Observe that $d(f,g) \leq d_x(f,g) \leq \sqrt{2}H_s(f,g)$, for any two densities $f, g \in \mathcal{F}$ and $x \in \mathbb{Y}$, by the definition of d and equation (6.12). Therefore, for each j,

$$d_{y_n}(f_{nj}, f_0) \geq d_{y_n}(f_j, f_0) - d_{y_n}(f_{nj}, f_j) \geq d(f_j, f_0) - \sqrt{2}H_s(f_{nj}, f_j),$$

and so

$$d_{y_n}(f_{nj}, f_0) > \gamma \tag{6.16}$$

We know from (6.14) that

$$L_{n+2j} = L_{nj} \frac{f_{nj}(y_{n+2}, y_{n+1}|y_n)}{f_0(y_{n+2}|y_{n+1}) f_0(y_{n+1}|y_n)},$$

with $L_{0j} = \Pi(A_j)$ by definition. Taking conditional expectations and applying Lemma 5, we get

$$\mathbb{E}_0\left[\sqrt{L_{n+2j}}\,\middle|\,\mathcal{A}_n\right] \leq \sqrt{L_{nj}}\left\{1 - d^2_{y_n}(f_{nj}, f_0)\right\} < \sqrt{L_{nj}}\,(1-\gamma^2).$$

Now, if we let k be the smallest integer larger than $n/2$, by iterating over k we find

$$\mathbb{E}_0\left[\sqrt{L_{n+2j}}\right] < \sqrt{L_{0j}}\,(1-\gamma^2)^k < \sqrt{\Pi(A_j)}\,(1-\gamma^2)^{(n+2)/2}.$$

Markov's inequality implies that, for any $b > 0$,

$$\mathbb{P}_0\left[\sum_{j=1}^{\infty} \sqrt{L_{nj}} > \exp(-nb)\right] < \exp(nb)\,(1-\gamma^2)^{n/2} \sum_{j=1}^{\infty} \sqrt{\Pi(A_j)}.$$

Finally, taking $b < -\log(1-\gamma^2)/2$, by condition (6.15), we arrive at

$$\sum_{j=1}^{\infty} \sqrt{L_{nj}} < \exp(-nb) \quad [\mathbb{P}_0]\text{-a.s.}$$

for all large n.

If $\mathcal{F}$ is not separable with respect to the supremum Hellinger distance, it is still possible to achieve the result, whenever d defines a distance over $\mathcal{F}$.

Lemma 7 *Let $A_\varepsilon \subset \mathcal{F}$ be a set of transition densities d-bounded away from f_0,*

$$A_\varepsilon = \{f \in \mathcal{F} : d(f, f_0) > \varepsilon\}.$$

Assume the operator d defined by (6.6) is a distance with respect to which $\mathcal{F}$ is separable and

$$\sum_{j=1}^{\infty} \sqrt{\Pi(A_j)} < \infty, \tag{6.17}$$

for some d-cover , $\{A_j\}_{j\geq 1}$ for A_ε, of size $\delta < \varepsilon/\sqrt{2}$.

Then, the result of Lemma 6 still holds.

Proof The inequality (6.16) is derived from the triangle inequality for d and the observation $d(f,g) \leq d_x(f,g)$ for all $f, g \in \mathcal{F}$ and $x \in \mathbb{Y}$, since

$$d_{y_n}(f_{nj}, f_0) \geq d(f_{nj}, f_0) \geq d(f_j, f_0) - d(f_{nj}, f_j) > \gamma. \tag{6.18}$$

The rest of the proof follows as the proof for the previous lemma.

6.2.3 The Denominator

For every $x \in \mathbb{Y}$, the Kullback-Leibler divergence from $f_0(\cdot|x)$ to $f(\cdot|x)$ is given by

$$K(f(\cdot|x), f_0(\cdot|x)) = \int \log\left(\frac{f_0(y|x)}{f(y|x)}\right) f_0(y|x) \mathrm{d}\nu(y). \tag{6.19}$$

Once again, an adequate generalization of the semimetric must be found, to remove the random element x from (6.19), before a Kullback-Leibler property for transition densities can be defined. Since the Kullback-Leibler property regards

the prior. this time it is convenient to define a semimetric on $\mathcal{F}$ by integration of the additional variable, since in the context we use the expression, such variable does not represent an observation. A common practice is to exploit the ergodicity of the process to perform the integration.

The integrated Kullback-Leibler divergence between f_0 and f is defined as

$$K(f, f_0) = \int K\big(f(\cdot|x), f_0(\cdot|x)\big)\mathrm{d}\nu_0(x).$$

In particular, if the stationary density f_0 is well defined, then

$$K(f, f_0) = \mathbb{E}_0\left[K\big(f(\cdot|x), f_0(\cdot|x)\big)\right] = \int K\big(f(\cdot|x), f_0(\cdot|x)\big) f_0(x)\mathrm{d}\nu(x).$$

Lemma 8 *Assume the prior Π has the Kullback-Leibler property at f_0, that is*

$$\Pi\big(\{f : K(f, f_0) < \varepsilon\}\big) > 0 \quad \textit{for all} \quad \varepsilon > 0.$$

Then for every $c > 0$ and sufficiently large n

$$I_n > \exp(-nc) \quad [\mathbb{P}_0]\text{-}a.s.$$

The proof follows from Fatou's lemma and the law of large numbers for ergodic Markov processes (see Ghosal & Tang, 2006; Tang & Ghosal, 2007b).

6.2.4 Posterior Consistency Result

We now have everything we need to present our main result.

Theorem 4 *Let A_ε be a set of transition densities d-bounded away from f_0,*

$$A_\varepsilon = \{f \in \mathcal{F} : d(f, f_0) > \varepsilon\}$$

with d defined by (6.6). Assume Π has the Kullback-Leibler property and

$$\sum_{j=1}^{\infty} \sqrt{\Pi(A_j)} < \infty, \tag{6.20}$$

where one of the following is true:

i) The operator d defines a distance on $\mathcal{F}$ and $\{A_j\}_{j\geq 1}$ is a countable cover for A_ε of d-size $\delta < \varepsilon/\sqrt{2}$;

ii) or $\{\Lambda_j\}_{j\geq 1}$ is a countable cover for Λ_ε of Π_s-size $\delta < \varepsilon/\sqrt{2}$.

Then

$$\Pi^n(A_\varepsilon) \to 0 \quad [\mathbb{P}_0]\text{-}a.s.$$

Proof Let $\{A_j\}$ be the cover satisfying condition (6.20), and denote $L_{nj} = L_{nA_j}$ for simplicity. Then

$$\begin{aligned}\Pi^n(A_\varepsilon) &\leq \sum_{j=1}^{\infty} \Pi^n(A_j) \leq \sum_{j=1}^{\infty} \sqrt{\Pi^n(A_j)} \\ &= \sum_{j=1}^{\infty} \sqrt{L_{nj}/I_n} = I_n^{1/2} \sum_{j=1}^{\infty} \sqrt{L_{nj}}.\end{aligned}$$

Applying Lemmas 8, and 6 or 7 as required, we have $\Pi^n(A_\varepsilon) \leq \exp\{-nb\}/\exp\{-nc\}$ for every $c > 0$ and $b < -\log\left[1+(\varepsilon-\sqrt{2}\delta)^2\right]/2$. Therefore, $\Pi^n(A_\varepsilon) \to 0$ as $n \to \infty$ exponentially fast, $[\mathbb{P}_0]$-a.s.

6.3 Illustrations

6.3.1 Example 1: Normal Autoregressive Model

Recall once more the simple parametric model mentioned in Section 6.1, with transition density given by

$$f_\theta(y_{n+1}|y_n) = N(\cdot|\theta y_n, 1) \quad \theta \in \Theta \subset \mathbb{R}.$$

This corresponds to the normal autoregressive AR(1) model

$$Y_{n+1} = \theta Y_n + \epsilon_n; \quad \epsilon_n \overset{iid}{\sim} \mathrm{N}(\cdot|0,1), \tag{6.21}$$

which is known to be stationary only for $\theta \in (0,1)$. This is one of the simplest and common time series models, yet there is no straightforward result guaranteeing strong consistency for the transition densities that can be applied to it when the stationarity assumption is not satisfied. In particular, Ghosal & van der Vaart

(2007) provide results for consistency, only when the process is stationary, while Ghosal & Tang's 2006 results guarantee strong consistency only when Θ is compact. Other ideas, based on the construction of sieves and uniformly consistent tests for adequate metrics, would require a careful study for each proposed prior.

On the other hand, the separability of $\mathbb{R}$ makes it straightforward to check if a prior Π on Θ satisfies the conditions of Theorem 4 and, as we mentioned before, the operator d defines a metric on the space $\mathcal{F}$ of transition densities for this particular model, even if $\Theta = \mathbb{R}$.

Recall that

$$d(f_\theta, f_{\theta^*}) = 1 - \frac{2}{\sqrt{4 + (\theta - \theta^*)^2}}. \tag{6.22}$$

Therefore, for every $\delta > 0$,

$$|\theta - \theta^*| < \bar{\delta} = 2\sqrt{(1-\delta^2)^{-2} - 1} \quad \Rightarrow \quad d^2(f_\theta, f_{\theta^*}) < \delta, \tag{6.23}$$

so a countable d-cover of size δ for $\mathcal{F}$ can be defined in terms of a cover of size $\bar{\delta}$ for $\mathbb{R}$ in the following way:

$$B_j = \left(j\bar{\delta}, (j+3/2)\bar{\delta}\right) \subset \mathbb{R}; \quad A_j = \{f_\theta : \theta \in B_j\}; \quad j \in \mathbb{Z}.$$

By symmetry, in order to prove

$$\sum_{j=-\infty}^{\infty} \sqrt{\Pi(f_\theta \in A_j)} = \sum_{j=-\infty}^{\infty} \sqrt{\Pi(\theta \in B_j)} < \infty$$

it is enough to show

$$\sum_{j=0}^{\infty} \sqrt{\Pi(\theta \in B_j)} < \infty,$$

which can be easily verified for any particular choice of Π.

6.3.2 Example 2: Nonparametric Mixture Model

Consider a time series model with transition densities given by

$$f(y_{n+1}|y_n) = \int_\Theta K(y_{n+1}|y_n, \theta) \mathrm{d}P_{y_n}(\theta), \tag{6.24}$$

where $K(\cdot|\theta)$ is a parametric density on $\mathbb{Y}$, for every $\theta \in \Theta$ and $\{P_x\}_{x\in\mathbb{Y}}$ is a family of mixing probability measures on Θ. In the most general case, the P_x may be non parametric and the prior Π placed over them is usually some dependent measure valued process. Models of this type are becoming common in the literature; some of them can be found in e.g. Mena & Walker (2007, 2005) and Martínez-Ovando & Walker (2011). In particular, the stationary time series model constructed in Chapter 3 has this general form.

The family $\mathcal{F}$ of transition densities of interest for this type of models is defined by the support of the prior Π.

Assume a sequence of observations $\{y_n\}_{n\geq 0}$ is generated from a time homogeneous Markov process with transition density $f_0 \in \mathcal{F}$. In other words, there is some probability measure P_0 such that, for every n,

$$f_0(y_{n+1}|y_n) = \int_\Theta K(y_{n+1}|y_n, \theta)\mathrm{d}P_0(\theta|y_n). \tag{6.25}$$

Assume that

$$\inf_x K(y|x,\theta) > \beta g(y), \quad \forall\, y \in \mathbb{Y},\, \theta \in \Theta, \tag{6.26}$$

for some $\beta > 0$ and a density function g with full support on $\mathbb{Y}$. Then for every $y, x \in \mathbb{Y}$,

$$f_0(y|x) > \beta \int_\Theta g(y)\mathrm{d}P_0(\theta|x) = \beta g(y). \tag{6.27}$$

If additionally, $f_0(\cdot|x)$ is continuous on x, then the conditions of Corollary 1 are satisfied and the operator d can be used to define strong neighbourhoods around f_0. Under this assumptions, strong consistency follows for any prior Π for which the conditions of Theorem 4 hold. The verification of consistency is therefore reduced to checking conditions on the prior.

Note that condition 6.26 hold whenever the state space $\mathbb{Y}$ and the parameter space Θ are compact. Therefore, the results of this Chapter can be applied to the general stationary model of Chapter 3 to prove strong consistency, for many choices of parametric kernel. The particular Gaussian kernel, however, does not satisfy this condition, so further analysis is required before consistency can be assessed.

6.4 Discussion

In this Chapter, we present a result which guarantees consistency of Bayesian transition density estimates, for general Markov models. We prove that the Kullback-Leibler property, together with a single additional condition on the prior distribution for the model, are sufficient to guarantee consistency, in the sense that the posterior distribution accumulates all its mass around d-neighbourhoods of the true transition density f_0. The opperator d, which we define, does not constitute a distance on the space of transition densities. It does, however, allow the definition of a neighbourhood system around the true transition density generating the data. We provide conditions on which such neighbourhoods are strong, in the sense that $d(f_0, f) > 0$ whenever $f \neq f_0$, therefore allowing us to uniquely identify f_0.

Our consistency result generalizes the martingale approach of Walker (2003, 2004) for i.i.d. observations. It succeeds in providing a single set of sufficient conditions which need to be verified for any model, without the need for constructing sieves and uniformly consistent tests. A previous result, due to Ghosal & Tang (2006), provides sufficient conditions for a somewhat weaker topology. Our main contribution is the definition of the d-neighbourhoods which provide a more useful form of consistency for more general classes of Markov models. Future work can be carried out to more accurately describe the conditions on the family of transition densities in the Kullback-Leibler support of the prior, for which the d operator defines strong neighbourhoods.

The key idea behind our consistency result is to use the Markov dependence structure of the data to construct a system of neighbourhoods, based on the Hellinger distance between the joint densities defined by the two step transitions. This idea can be easily extended to higher dimensions, thus providing consistency results for higher order Markov models. Furthermore, similar ideas could be applicable for regression models, or in general, for the estimation of any conditional density.

Chapter 7

Discussion and future work

The driving theme of this thesis is intractability in Bayesian models. The problem is not new, and throughout the years, many methods have been developed to deal with it. Some of them rely on approximations for the likelihood functions, for the posterior densities of the model parameters, or for point and interval estimates directly. A favoured approach consists of MCMC methods, in which Monte Carlo estimates are produced by posterior sampling through a Markov chain construction, of which the equilibrium distribution coincides with the desired posterior density. In many cases, such constructions are enabled or facilitated by the introduction of auxiliary variables, resulting in latent models for which posterior simulation can be achieved. An emphasis is placed on the definition of exact simulation methods, for which no approximation error is introduced; Monte Carlo error is known to be well behaved, and therefore preferred over a fixed approximation error.

Various ideas can be found in the literature to deal with intractable posterior distributions; many of them can be applied to achieve exact posterior simulation when all random objects are finite dimensional. When infinite-dimensional spaces are involved, things get more complicated. An exact simulation algorithm for diffusion processes was recently developed by Beskos *et al.* (2006b) which enables exact posterior simulation for parametric diffusion models, in which the state space for the diffusion paths is a functional space. Several methods exist which enable posterior inference for Bayesian nonparametric models with a stick-breaking representation (see e.g. Escobar, 1988; Kalli *et al.*, 2011; MacEachern &

Müller, 1998; Neal, 2000; Papaspiliopoulos & Roberts, 2008). These are examples of inference methods developed for intractable models in the presence of infinite-dimensional quantities. Each construction is specific to the model at hand and not applicable in other situations. In the present work, we approach the problem of inference for this type of model with infinitely-generated intractable components. The key is to use a power series representation for the intractable functions and then introduce the exponents as auxiliary variables for a latent model. Once this is done, we are able to profit from existing methods to construct viable MCMC schemes for posterior simulation.

The general idea we propose is applicable in a wide range of situations. We illustrate this through various examples. In Chapter 2, we deal with inference for discretely observe diffusions and show our series expansion latent variable approach to be equivalent to the auxiliary variable scheme implicit in Beskos *et al.*'s exact simulation method. Chapters 3 and 4 propose novel nonparametric time series and regression models, respectively, for which dependent mixture weights are constructed via normalization. Even though the construction is simple, only the finite mixture versions of these models have been used before, since no inference method was available for the infinite versions. In Chapter 5 we illustrate the use of our method for inference based on the power likelihood of nonparametric models. At the end of each chapter, we included a brief discussion on the uses and possible extensions for each model. We now proceed with a more general discussion.

The models presented in Chapters 3 and 4 are clearly related, the first one being an autoregressive version of the second. For each model, however, we chose the parametrization that better suited the study of relevant model properties: stationarity for the time series; mean curve shape for the regression. For illustrative purposes, we chose to develop the time series model in a simple form. However, it should be evident from the regression model construction, that a generalization is relatively straightforward. We can, therefore, consider the definition of multivariate time series, with combinations of discrete and continuous variables; covariate dependent time series; or state-space models, in which time and other covariates can be introduced in the joint mixture model, resulting in complex conditional distribution structures. Each of these generalizations would surely present its

own challenges, but we believe our latent variable and MCMC approach provides a good starting point for the analysis of a new family of dependent and time dependent processes.

An interesting observation arises from the analysis of a discretely observed diffusion path using the stationary model of Chapter 3. Namely, that a flexible time series model can be a good alternative to the use of diffusion models, in situations where the theoretical framework does not suggest a specific form of the diffusion and drift coefficients, or an interpretation of the diffusion model parameters. Since diffusion processes are defined through the infinitesimal characteristics of their paths, which can not be observed in real situations, a dependent nonparametric process may be better suited for statistical inference. Furthermore, applying the ideas presented in Chapter 5, inference could be achieved for a smoothed version of the likelihood, a power-likelihood, thus reducing the effect of noisy observations, and producing consistent estimates.

Finally, in Chapter 6, we present a result for Bayesian consistency of transition density estimates for Markov models. We find a set of sufficient conditions under which a strong type of consistency can be proved, applicable in particular to the general family of Markov models defined in Chapter 3, provided a suitable choice of the parametric kernel is made. Some work is yet to be done, before the result can be applied for more general kernel choices. However, we believe the idea of defining neighbourhoods with respect to distances on extended spaces, this case, the Hellinger distance over joint densities, is worth exploring. We will continue to study this type of quantities, and their potential for defining strong neighbourhoods around conditional densities.

As we have mentioned several times thorough this work, we do not claim to have exhausted here the subjects of estimation for intractable models, or posterior consistency for dependent densities. We have merely set the first stone over which we expect to build and, hopefully, succeed in developing better and stronger results.

References

Adams, R.P., Murray, I. & MacKay, D.J.C. (2009). Nonparametric Bayesian Density Modeling with Gaussian Processes. URL http://arxiv.org/abs/0912.4896. 117

ADEAR, Alzheimer's Disease Education & Referral Center (2011). Alzheimer's Disease Fact Sheet. *NIH Publication*, **11-6423**. 132

Andrieu, C. & Roberts, G.O. (2009). The Pseudo-Marginal Approach for Efficient monte carlo Computations. *The Annals of Statistics*, **37**, 697–725. 44

Antoniak, C.E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, **2**, 1152–1174. 37

Antoniano-Villalobos, I. & Walker, S.G. (2012a). Bayesian Consistency for Markov Models, working paper. 155

Antoniano-Villalobos, I. & Walker, S.G. (2012b). Bayesian Nonparametric Inference for the Power Likelihood. *Journal of Computational and Graphical Statistics*, accepted for publication. 142

Antoniano-Villalobos, I., Wade, S. & Walker, S.G. (2012). A Bayesian Nonparametric Study of Hippocampal Atrophy in Alzheimer's Disease, submited paper. 35, 116

Barndorff-Nielsen, O.E. & Sorensen, M. (1994). A Review of Some Aspects of Asymptotic Likelihood Theory for Stochastic Processes. *International Statistical Review*, **62**, 133–165. 55

BARRIENTOS, A.F., JARA, A. & QUINTANA, F.A. (2012). On the Support of MacEachern's Dependent Dirichlet Processes and Extensions. *Bayesian Analysis*, **7**, 277–310. 38

BARRON, A., SCHERVISH, M.J. & WASSERMAN, L. (1999). The Consistency of Posterior Distributions in Nonparametric Problems. *The Annals of Statistics*, **27**, 536–561. 63, 65, 67, 68, 69, 141, 151

BESAG, J. & GREEN, P.J. (1993). Spatial Statistics and Bayesian Computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 25–37. 42, 46

BESKOS, A. & ROBERTS, G.O. (2005). Exact Simulation of Diffusions. *Annals of Applied Probability*, **15**, 2422–2444. 55

BESKOS, A., PAPASPILIOPOULOS, O. & ROBERTS, G.O. (2006a). Retrospective Exact Simulation of Diffusion Sample Paths With Applications. *Bernoulli*, **12**, 1077–1098. 28, 56, 57, 59, 74

BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G.O. & FEARNHEAD, P. (2006b). Exact and Computationally Efficient Likelihood-Based Estimation for Discretely Observed Diffusion Processes (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **68**, 333–382. 3, 56, 59, 60, 72, 76, 82, 83, 92, 109, 112, 173, 174

BESKOS, A., PAPASPILIOPOULOS, O. & ROBERTS, G. (2008). A Factorisation of Diffusion Measure and Finite Sample Path Constructions. *Methodology and Computing in Applied Probability*, **10**, 85–104. 58

BESKOS, A., PAPASPILIOPOULOS, O. & ROBERTS, G. (2009). Monte Carlo Maximum Likelihood Estimation for Discretely Observed Diffusion Processes. *The Annals of Statistics*, **37**, 223–245. 56, 59

BIBBY, B.M. & SORENSEN, M. (1995). Martingale Estimation Functions for Discretely Observed Diffusion Processes. *Bernoulli*, **1**, 17–39. 55, 88

BLACKWELL, D. & MACQUEEN, J.B. (1973). Ferguson Distributions via Pólya Urn Schemes. *The Annals of Statistics*, **2**, 353–355. 16

CAPPÉ, O., MOULINES, E. & RYDÉN, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics, Springer. 34

CARLIN, B.P. & CHIB, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 473–484. 53

CAROLI, A. & FRISONI, G. (2010). The Dynamics of Alzheimer's Disease Biomarkers in the Alzheimer's Disease Neuroimaging Initiative Cohort. *Neurobiology of Aging*, **31**, 1263–1274. 134

CHIB, S. (1995). Marginal Likelihood From the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321. 53

CHIB, S. & GREENBERG, E. (1998). Analysis of Multivariate Probit Models. *Biometrika*, **85**, 347–361. 53

CHOI, T. & SCHERVISH, M.J. (2007). On Posterior Consistency in Nonparametric Regression Problems. *Journal of Multivariate Analysis*, **98**, 1969–1987. 157

CHUNG, Y. & DUNSON, D.B. (2009). Nonparametric Bayes Conditional Distribution Modeling With Variable Selection. *Journal of the American Statistical Association*, **104**, 1646–1660. 39

CIFARELLI, D. & REGAZZINI, E. (1978). Problemi Statistici Nonparametrici in Condizioni di Scambiabilità Parziale e Impiego di Medie Associative. *Quaderni Istituto di Matematica Finanziaria, Università di Torino*, **12**, 1–36, english translation available at `www.unibocconi.it/wps/allegatiCTP/CR-Scamb-parz[1].20080528.135739.pdf`. 37

CRUZ-MESÍA, QUINTANA, F.A. & MÜLLER, P. (2007). Semiparametric Bayesian Classification With Longitudinal Markers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56**, 119–137. 38

DAMIEN, P. & WALKER, S.G. (2001). Sampling Truncated Normal, Beta, and Gamma Densities. *Journal of Computational and Graphical Statistics*, **10**, 206–215. 48, 49, 105

DAMIEN, P., WAKEFIELD, J. & WALKER, S. (1999). Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 331–344. 47

DE IORIO, M., MÜLLER, P., ROSNER, G.L. & MACEACHERN, S.N. (2004). An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, **99**, 205–215. 38

DE IORIO, M., JOHNSON, W.O., MÜLLER, P. & ROSNER, G.L. (2009). Bayesian Nonparametric Nonproportional Hazards Survival Modeling. *Biometrics*, **65**, 762–771. 38

DENISON, D.G.T., HOLMES, C.C., MALLICK, B.K. & SMITH, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley. 35

DIACONIS, P. & FREEDMAN, D. (1986). On the Consistency of Bayes Estimates. *The Annals of Statistics*, **14**, 1–26. 61, 62

DICICCIO, T.J., KASS, R.E., RAFTERY, A. & WASSERMAN, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, **92**, 903–915. 2, 40

DIMATTEO, I., GENOVESE, D.R. & KASS, R.E. (2001). Bayesian Curve Fitting With Free-Knot Splines. *Biometrika*, **88**, 1055–1071. 35

DOOB, J.L. (1949). Application of the Theory of Martingales. In *Le Calcul des Probabilités et ses Applications*, 23–27, Colloques Internationaux du Centre National de la Recherche Scientifique, Paris. 61

DUNSON, D.B. & PARK, J.H. (2008). Kernel Stick-Breaking Processes. *Biometrika*, **95**, 307–323. 39

DUNSON, D.B. & RODRÍGUEZ, A. (2011). Nonparametric Bayesian Models Through Probit Stick-Breaking Processes. *Bayesian Analysis*, **6**, 145–178. 39

ESCOBAR, M.D. (1988). *Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means*. Ph.D. thesis, Department of Statistics, Yale University, New Haven. 50, 142, 173

ETHIER, S.N. & KURTZ, T.G. (1986). *Markov Processes. Characterization and Convergence.*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York. 26

EVANS, M. & SWARTZ, T. (1995). Methods for Approximating Integrals in Statistics With Special Emphasis on Bayesian Integration Problems. *Statistical Science*, **10**, 254–272. 2, 40

FERGUSON, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 209–230. 15, 16, 141

FRIEL, N. & PETTITT, A.N. (2008). Marginal Likelihood Estimation via Power Posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 589–607. 142

FRISONI, G., FOX, N., JACK, C.J., SCHELTENS, P. & THOMPSON, P. (2010). The Clinical Use of Structural MRI in Alzheimer Disease. *Nature Reviews Neurology*, **6**, 67–77. 134

FUENTES-GARCÍA, R., MENA, R.H. & WALKER, S.G. (2010). A New Bayesian Nonparametric Mixture Model. *Communications in Statistics - Simulation and Computation*, **39**, 669–682. 18

GELFAND, A.E., KOTTAS, A. & MACEACHERN, S.N. (2005). Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing. *Journal of the American Statistical Association*, **100**, 1021–1035. 38

GELMAN, A. & MENG, X.L. (1998). Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Statistical Science*, **13**, 163–185. 44

GHOSAL, S. & ROY, A. (2006). Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression. *The Annals of Statistics*, **34**, 2413–2429. 157

GHOSAL, S. & TANG, Y. (2006). Bayesian Consistency for Markov Processes. *The Indian Journal of Statistics*, **68**, 227–239. 158, 159, 160, 163, 168, 170, 172

GHOSAL, S. & VAN DER VAART, A. (2007). Convergence Rates of Posterior Distributions for Noniid Observations. *The Annals of Statistics*, **35**, 192–223. 169

GHOSAL, S., GHOSH, J.K. & RAMAMOORTHI, R.V. (1999). Posterior Consistency of Dirichlet Mixtures in Density Estimation. *The Annals of Statistics*, **27**, 143–158. 65

GODSILL, S.J. (2001). On the Relationship Between Markov Chain Monte Carlo Methods for Model Uncertainty. *Journal of Computational and Graphical Statistics*, **10**, 230–248. 5, 53, 54, 76, 77, 101, 106, 127, 148

GREEN, P.J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, 711–732. 54, 106

GRIFFIN, J.E. & STEEL, M.F.J. (2006). Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 179–194. 34, 39

GRIFFIN, J.E. & STEEL, M.F.J. (2011). Stick-Breaking Autoregressive Processes. *Journal of Econometrics*, **162**, 383–396. 34

HANNAH, L.A., BLEI, D.M. & POWELL, W.B. (2011). Dirichlet Process Mixtures of Generalized Linear Models. *Journal of Machine Learning Research*, **12**, 1923–1953. 39

HIGDON, D.M. (1998). Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications. *Journal of the American Statistical Association*, **93**, 585–595. 47

HJORT, N.L., HOLMES, C., MÜLLER, P. & WALKER, S.G. (2010). *Bayesian Nonparametrics*. Cambridge University Press. 141

IBRAHIM, J.G. & CHEN, M.H. (2000). Power Prior Distributions for Regression Models. *Statistical Science*, **15**, 46–60. 142

ISHWARAN, H. & JAMES, L.F. (2001). Gibbs Sampling Methods for Stick Breaking Priors. *Journal of the American Statistical Association*, **96**, 161–173. 15

ISHWARAN, H. & ZAREPOUR, M. (2000). Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models. *Biometrika*, **87**, 371–390. 50

JACK, C.J., KNOPMAN, D., JAGUST, W., SHAW, L., P.S., A., WEINER, M., PETERSEN, R. & TROJANOWSKI, J. (2010). Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade. *Lancet Neurology*, **9**, 119–128. 134

JACK, C.J., VEMURI, P., WISTE, H., WEIGAND, S., LESNICK, T., LOWE, V., KANTARCI, K., BERNSTEIN, M., SENJEM, M., GUNTER, J., BOEVE, B., TROJANOWSKI, J., SHAW, L., AISEN, P., WEINER, M., PETERSEN, R. & KNOPMAN, D. (2012). Shapes of the Trajectories of 5 Major Biomarkers of Alzheimer Disease. *Archives of Neurology*, **69**, 856–867. 134

JARA, A., LESAFFRE, E., DE IORIO, M. & QUINTANA, F. (2010). Bayesian Semiparametric Inference for Multivariate Doubly-Interval-Censored Data. *The Annals of Applied Statistics*, **4**, 2126–2149. 38

KALLI, M., GRIFFIN, J.E. & WALKER, S.G. (2011). Slice Sampling Mixture Models. *Statistics and Computing*, **21**, 93–105. 50, 51, 52, 100, 122, 123, 142, 145, 152, 173

KANG, C. & GHOSAL, S. (2009). Clusterwise Regression Using Dirichlet Mixtures. In *Advances in Multivariate Statistical Methods*, 305–325. 39

KARATZAS, I. & SHREVE, S.E. (1991). *Brownian Motion and Stochastic Calculus*, vol. 113 of *Graduate Texts in Mathematics*. Springer, 2nd edn. 29

KELLY, L., PLATEN, E. & SORENSEN, M. (2004). Estimation for Discretely Observed Diffusions Using Transform Functions. *Journal of Applied Probability*, **41**, 99–118. 55

LAMPERTI, J. (1977). *Stochastic Processes. A Survey of the Mathematical Theory*. Springer-Verlag, New York. 26

LENK, P.J. (1991). Towards a Practicable Bayesian Nonparametric Density Estimator. *Biometrika*, **78**, 531–543. 33

LEONARD, T. (1978). Density Estimation, Stochastic Processes and Prior Information. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40**, 113–146. 33

LIJOI, A., MENA, R.H. & PRÜNSTER, I. (2005). Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors. *Journal of the American Statistical Association*, **100**, 1278–1291. 18

LO, A.Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, **12**, 351–357. 20, 141

MACEACHERN, S.N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55, American Statistical Association. 37, 38, 116

MACEACHERN, S.N. (2000). Dependent Dirichlet Processes. Tech. rep., Department of Statistics, Ohio State University. 37, 38

MACEACHERN, S.N. (2001). Decision Theoretic Aspects of Dependent Nonparametric Processes. *Proceedings of Bayesian Methods with Applications to Science, Policy, and Official Statistics (ISBA 2000)*, 351–360. 38

MACEACHERN, S.N. & MÜLLER, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**, 223–238. 50, 142, 173

MARTÍNEZ-OVANDO, J.C. & WALKER, S.G. (2011). Time-Series Modelling, Stationarity and Bayesian Nonparametric Methods. Tech. rep., Banco de México. 33, 34, 97, 171

MENA, R. & WALKER, S.G. (2007). On the Stationary Version of the Generalized Hyperbolic ARCH Model. *Annals of the Institute of Statistical Mathematics*, **59**, 325–348. 171

MENA, R.H. & WALKER, S.G. (2005). Stationary Autoregressive Models via a Bayesian Nonparametric Approach. *Journal of Time Series Analysis*, **26**, 789–805. 32, 33, 34, 171

MØLLER, J., PETTITT, A.N., REEVES, R. & BERTHELSEN, K.K. (2006). An Efficient Markov Chain Monte Carlo Method for Distributions With Intractable Normalising Constants. *Biometrika*, **93**, 451–458. 44, 117

MULIERE, P. & PETRONE, S. (1993). A Bayesian Predictive Approach to Sequential Search for an Optimal Dose: Parametric and Nonparametric Models. *Statistical Methods & Applications*, **2**, 349–364. 37

MÜLLER, P. & QUINTANA, F. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 95–110. 39

MÜLLER, P. & QUINTANA, F. (2010). Random Partition Models With Regression on Covariates. *Journal of Statistical Planning and Inference*, **140**, 2801–2808. 39

MÜLLER, P., ERKANLI, A. & WEST, M. (1996). Bayesian Curve Fitting Using Multivariate Normal Mixtures. *Biometrika*, **83**, 67–79. 34, 39

MÜLLER, P., WEST, M. & MACEACHERN, S. (1997). Bayesian Models for Non-Linear Auto-Regressions. *Journal of Time Series Analysis*, **18**, 593–614. 31, 97

MÜLLER, P., QUINTANA, F. & ROSNER, G. (2004). A Method for Combining Inference Across Related Nonparametric Bayesian Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 735–749. 39

MÜLLER, P., ROSNER, G.L., IORIO, M.D. & MACEACHERN, S. (2005). A Nonparametric Bayesian Model for Inference in Related Longitudinal Studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 611–626. 38

MURRAY, I. & GHAHRAMANI, Z. (2004). Bayesian Learning in Undirected Graphical Models: Approximate MCMC Algorithms. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 392–399, AUAI Press. 44

MURRAY, I., GHAHRAMANI, Z. & MACKAY, D.J.C. (2006). MCMC for Doubly-Intractable Distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, 359–366, AUAI Press. 45, 117

NEAL, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 249–265. 50, 142, 174

NIETO-BARAJAS, L.E., PRÜNSTER, I. & WALKER, S.G. (2004). Normalized Random Measures Driven by Increasing Additive Processes. *The Annals of Statistics*, **32**, 2343–2360. 21

PAPASPILIOPOULOS, O. & ROBERTS, G.O. (2008). Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models. *Biometrika*, **95**, 169–186. 50, 174

PARK, J. & DUNSON, D.B. (2010). Bayesian Generalized Product Partition Model. *Statistica Sinica*, **20**, 1203–1226. 39

PETTITT, A.N., FRIEL, N. & REEVES, R. (2003). Efficient Calculation of the Normalizing Constant of the Autologistic and Related Models on the Cylinder and Lattice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 235–246. 117

PITMAN, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In *Statistics, Probability and Game Theory, Papers in Honor of David Blackwell*, 245–267, Institute of Mathematical Statistics, Hayward, CA. 16

PITT, M.K. & WALKER, S.G. (2005). Constructing Stationary Time Series Models Using Auxiliary Variables With Applications. *Journal of the American Statistical Association*, **100**, 554–564. 32

PITT, M.K., CHATFIELD, C. & WALKER, S.G. (2002). Constructing First Order Stationary Autoregressive Models via Latent Processes. *Scandinavian Journal of Statistics*, **29**, 657–663. 32, 34

RASMUSSEN, C.E. & WILLIAMS, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press. 35

REGAZZINI, E., LIJOI, A. & PRÜNSTER, I. (2003). Distributional Results for Means of Normalized Random Measures With Independent Increments. *The Annals of Statistics*, **31**, 560–585. 21

REICH, B.J. & FUENTES, M. (2007). A Multivariate Semiparametric Bayesian Spatial Modeling Framework for Hurricane Surface Wind Fields. *Annals of Applied Statistics*, **1**, 249–264. 39

REN, L., DU, L., CARIN, L. & DUNSON, D.B. (2011). Logistic Stick-Breaking Process. *Journal of Machine Learning Research*, **12**, 203–239. 39

REVUZ, D. & YOR, M. (1999). *Continuous Martingales and Brownian Motion (Grundlehren der mathematischen Wissenschaften)*. Springer-Verlag. 29

ROBERT, C.P. & CASELLA, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edn. 42

RODRIGUEZ, A. & TER HORST, E. (2008). Bayesian Dynamic Density Estimation. *Bayesian Analysis*, **3**, 339–365. 38

RODRIGUEZ, A., DUNSON, D.B. & GELFAND, A.E. (2008). The Nested Dirichlet Process. *Journal of the American Statistical Association*, **103**, 1131–1154. 39

SABUNCU, M., DESIKAN, R., SEPULCRE, J., YEO, B., LIU, H., SCHMANSKY, N., REUTER, M., WEINER, M., BUCKNER, R., SPERLING, R. & FISCHL,

B. (2011). The Dynamics of Cortical and Hippocampal Atrophy in Alzheimer Disease. *Archives of Neurology*, **68**, 1040–1048. 134

SCHERVISH, M.J. (1995). *Theory of Statistics*. Springer Series in Statistics, Springer. 13

SCHWARTZ, L. (1965). On Bayes Procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **4**, 10–26. 63

SETHURAMAN, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, **4**, 639–650. 15, 143

SHAHBABA, B. & NEAL, R. (2009). Nonlinear Models Using Dirichlet Process Mixtures. *Journal of Machine Learning Research*, **10**, 1829–1850. 39

SMITH, A.F.M. (1991). Computational Methods. *Philosophical Transactions: Physical Sciences and Engineering*, **337**, 369–386. 2

SMITH, A.F.M. & ROBERTS, G.O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 3–23. 40, 42

SORENSEN, H. (2004). Parametric Inference for Diffusion Processes Observed at Discrete Points in Time: a Survey. *International Statistical Review*, **72**, 337–354. 2, 55

SWENDSEN, R.H. & WANG, J.S. (1987). Nonuniversal Critical Dynamics in Monte Carlo Simulations. *Physical Review Letters*, **58**, 86–88. 46

TANG, Y. & GHOSAL, S. (2007a). A Consistent Nonparametric Bayesian Procedure for Eestimating Autoregressive Conditional Densities. *Computational Statistics and Data Analysis*, **51**, 4424–4437. 31

TANG, Y. & GHOSAL, S. (2007b). Posterior Consistency of Dirichlet Mixtures for Estimating a Transition Density. *Journal of Statistical Planning and Inference*, **137**, 1711–1726. 158, 168

TEH, Y.W., JORDAN, M.I., BEAL, M.J. & BLEI, D.M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 1566–1581. 39

TIERNEY, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, **22**, 1701–1728. 42

VAN GAEL, J., SAATCI, Y., TEH, Y.W. & GHAHRAMANI, Z. (2008). Beam Sampling for the Infinite Hidden Markov Model. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, 1088–1095. 34

WALKER, S.G. (2003). On Sufficient Conditions for Bayesian Consistency. *Biometrika*, **90**, 482–488. 66, 158, 172

WALKER, S.G. (2004). New Approaches to Bayesian Consistency. *The Annals of Statistics*, **32**, 2028–2043. 66, 67, 158, 159, 163, 164, 172

WALKER, S.G. (2007). Sampling the Dirichlet Mixture Model With Slices. *Communications in Statistics - Simulation and Computation*, **36**, 45–54. 50

WALKER, S.G. & HJORT, N.L. (2001). On Bayesian Consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 811–821. 67, 141

WILLIAMSON, S., ORBANZ, P. & GHAHRAMANI, Z. (2010). Dependent Indian Buffet Processes. *Journal of Machine Learning Research - Proceedings Track*, **9**, 924–931. 34

ZHU, X., GHAHRAMANI, Z. & LAFFERTY, J. (2005). Time-Sensitive Dirichlet Process Mixture Models. Tech. Rep. CMU-CALD-05-104, Carnegie Mellon University, Pittsburgh. 34