DECLARATION FOR THE PhD THESIS

The undersigned, Di Lucca Maria Anna, PhD Registration Number 1287671

Thesis title: Bayesian Nonparametric Autoregressive Models with Applications

PhD in Statistics

Cycle XXIII

Candidate's tutor Prof. Pietro Muliere

Year of discussion 2012

DECLARES

Under her responsibility:

1. that, according to the President's decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of these hypotheses prove true, all benefits included in this declaration are automatically forfeited from the beginning;

2. that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted;

3. that the Servizio Biblioteca Bocconi will file the thesis in its 'Archivio istituzionale ad accesso aperto' and will permit on-line consultation of the complete text;

4. that in order keep the thesis on file at Biblioteca Bocconi, the University requires that the thesis be delivered by the candidate to Societá NORMADEC (acting on behalf of the University) by online procedure the contents of which must be unalterable and that NORMADEC will indicate in each footnote the

following information:

- thesis: Bayesian Nonparametric Autoregressive Models with Applications;
- by Di Lucca Maria Anna ;
- discussed at Universitá Commerciale Luigi Bocconi - Milano in 2012;
- the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Universitá Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source;

5. that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6. that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties;

7. that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo.

Date January, 31 2012

Signed Maria Anna Di Lucca

Bayesian Nonparametric Autoregressive Models with Applications

Maria Anna Di Lucca

Department of Decision Sciences

Bocconi University

A thesis submitted for the degree of

Doctor of Philosophy

Acknowledgements

Several people have played an essential role for the realization of this thesis. I would like to express my gratitude to Prof. Peter Müller for the main support, with his enthusiasm, his inspiration and his great effort to explain simply and clearly. I considered an honor to work with him.
I sincerly thank you Prof. Pietro Muliere for numerous ideas, methodological suggestions and theoretical support. I really appreciated our daily discussions about life and statistics thoughout all these years.
Thanks are due to Prof. Sonia Petrone for her helping me in making properly connections on and off Bayesian subjects.
I am also greateful to Prof. Ji for introducing me to genetic problem sets.
Special gratitude goes to Prof. Alessandra Guglielmi for providing information on applying Bayesian inferential methods. I have also benefit from many discussions with her. I really liked her irony during our work together.
I am deeply grateful to my teachers of the bachelor degree for personal support throughout all these years of the PhD program.

I would like to express thanks to my parents for supporting me throughout all my studies and feeling always nearby.
Sincere thanks are for my best friend Isabella. I consider her as my youngest sister.

Abstract

Many statistical problem sets are focused on complex phenomena and involve the idea of *dependence*, which is the key word of this thesis. Statistical dependence could be symmetric or asymmetric. Symmetric statistical dependence is, for example, analysis of variance. Asymmetric statistical dependence could be expressed by regression models. Our aim is modeling more levels of dependence and jointly infer on them. Recent statistical literature proves that previous models are not able to consider more complicated levels of dependence. This is one of the reasons of the diffusion of Bayesian methods. We focus on Bayesian nonparametric methods, such that these are able to keep more different levels of dependence, and share more useful information, thank for strong mathematical supports. We will propose models and applications for time series analysis.

Preface

The aim of this thesis is to find and share information about more aspects of time series problem sets. Statistical strenght of dependent prior models is share. For inference, Dirichlet process mixtures procedures achieve more information. We propose a dependent Dirichlet process prior for complex time series. Specifically, we assume a collection of random probability measures such that in some way marginally is a Dirichlet process and across the point masses we introduce dependence as autoregressive fashion-type. Time series analysis is often applied to financial and economical fields, describing phenomena developed over time. One of our relevant applications of our methodological innovations is in genetics, in particular, to a DNA-sequencing dataset.

We analyze density estimation connected to the choice of a dependent Dirichlet process prior as well as applications are to nonparametric autoregressive models. Indeed, we discuss about Bayesian hierarchical models, which are able to describe more levels of dependence. The power of the hierarchy is the introduction of dependence among model parameters. Prior dependence describes jointly more parts of the model and shares information among other parts of the same model. In addition, we introduce dependence among model parameters, using Bayesian hierarchical modelling.

This thesis is organized as follows. We propose five semi-authonomous chapters, which have the dependent Dirichlet process prior as main theme and are independent for some specific characterizations and purposes. Our first Chapter is an essential background about the Dirichlet process. We provide a review of the first definition of the Dirichlet process and its principal properties. We describe some theoretical results and some empirical applications for our purposes. We motivate essential reasons of its large employment. In the second Chapter, we focus on kernel density estimation problems and different methodological aspects for time series analysis. The third Chapter is entirely dedicated to an application for a real dataset, which is already analyzed in the literature using a Bayesian parametric method. In this chapter, we discuss about a Bayesian nonparametric method and we provide a comparison between the two methods for the DNA-sequencing dataset. The fourth Chapter is focused on possible extentions of the autoregressive models with more than one lag and variable weights dependent Dirichlet processes. In the fifth Chapter, we introduce possible further advances of our research.

Contents

Preface iii

Contents v

1 Dirichlet Process and related models 1

1.1 Introduction 1

1.2 Dirichlet Process 2

1.3 Main Properties of the Dirichlet Process 3

1.4 Mixtures of Dirichlet Processes 5

1.5 Applications and developments: Bayesian nonparametric inference for partial exchangeable data 6

1.5.1 Bayesian Nonparametric Linear Model 10

1.6 Dirichlet Process Mixtures 15

1.7 Dependence across a Collection of Random Probability Measures . 18

1.7.1 Dependent Dirichlet Process: definition and main properties 18

1.7.2 ANOVA and Dependent Dirichlet Process prior 22

1.8 Discussion 24

Bibliography 27

2 A Bayesian Nonparametric Autoregressive Model 31

2.1 Introduction 31

2.2 DDP-AR(1) 34
2.2.1 Posterior Distributions 37
2.2.2 Bayesian Inference 38
2.2.3 Model with unknown observations variance 40
2.2.4 Predictive Distribution 40
2.3 Example: Old Faithful Geyser Dataset 42
2.3.1 Co-Clustering 49
2.4 Extensions of the base DDP-AR(1) model 51
2.4.1 Efficient Trajectories 52
2.4.2 Linear and quadratic trajectories 53
2.4.3 Mixture of trajectories 54
2.5 Multivariate DDP-AR(1) 57
2.5.1 Seemingly Unrelated Linear Equations 58
2.5.2 Quadratic Form for the Trajectories 59
2.5.3 Unseparable Linear Equations 62
2.6 Discussion 63

Bibliography **65**

3 A Nonparametric Autoregressive Model for DNA-sequencing **71**
3.1 Introduction 71
3.2 Multivariate Mixture for DNA-Sequencing 79
3.2.1 Parametric Vector Autoregressive Model 79
3.2.2 Nonparametric VAR(1) 80
3.3 Prior Probability Model 84
3.4 Posterior Distributions 85
3.5 Inference on the Posterior Distributions 91
3.6 Final Discussion 96

Bibliography 99

4 A Bayesian Nonparametric Extention for Autoregressive Models 103

4.1 Introduction 104

4.2 DDP-AR(2) 105

4.2.1 Posterior Inference 107

4.3 Varying Weights DDP-AR(2) 109

4.3.1 Posterior Distributions for Varying Weights DDP-AR(2) Model 112

4.3.1.1 A variation for varying weights DDP- AR(2) model . 115

4.4 Example 116

4.5 Model adequacy 122

4.5.1 The influence of a modified DDP-AR(2) Varying Weights . . . 125

4.5.2 Single-p DDP prior for a complete AR(2) sampling model . . 126

4.6 Discussion 130

Bibliography 133

5 Further Research 137

Bibliography 141

Chapter 1

Dirichlet Process and related models

Abstract.

In this chapter we give a brief historical review of the Dirichlet process in Bayesian nonparametric inference, and of prior construction based on the Dirichlet process. In particular, we discuss the mixture of the Dirichlet process prior and the Dirichlet process mixtures, underlying differences and connections. The aim is a brief description of some relevant notions for our further developments, and a brief excursus in the Bayesian nonparametric context. We discuss some of the main reasons of the large employment of the Dirichlet process.

1.1 Introduction

Many models exist for time series problem sets, and most of them use classical inference based on likelihood functions and asymptotic behaviour. In Bayesian nonparametric inference, we need to place a prior on an infinite dimensional space such as the space of probability measures. However, a simple prior distribution, it is not flexible enough to describe complex phenomena. In the last years, several authors proposed dependent nonparametric prior models. Here we explore and motivate our methodological choice and the construction of dependent Dirichlet process prior.

We consider the impact of two alternative Bayesian assumptions: the exchangeability and the partial exchangeability on the data. We describe in the next sections the effects of these two different assumptions and how these influenced on the dependent Dirichlet process definition. For inference, we develop Dirichlet process mixture models. We introduce dependence across mixtures and, inferring on these, we share information, which is our purpose. In the last thirty years, the Dirichlet process played different roles, that we discuss in the next sections.

This chapter is organized as follows. In Section 1.2 we propose the original definition of the Dirichlet process introduced by Ferguson in 1973 and briefly review the main properties of the DP in Section 1.3. The first version of mixture models in Bayesian statistics is presented in Section 1.4. In Section 1.5 we study the inferential impact of the partial exchangeability assumption and in Section 1.6 we discuss about the exchanchangeability assumption in the DP mixture model. In Section 1.7 we illustrate two models based on DP mixtures, which are the main references for our statistical developments in the following chapters.

1.2 Dirichlet Process

The Dirichlet process, briefly DP, was introduced in Bayesian nonparametric inference by Ferguson (1973), and immediatly arose great attention in the literature for its flexibility and deep theoretical properties. In the recent years, its potential in applications has been greatly developed in a wide range of fields, also thanks for the advances in computer science and machine learning.

Ferguson (1973) introduced a class of stochastic processes such that each element can be used as a prior distribution on a measurable space $(\mathcal{X}, \mathcal{X})$. Let G be a stochastic process indexed by elements A of the σ-field of subsets $\mathcal{X}$. Let α be a non-null finite measure (nonnegative and finitely additive) on $(\mathcal{X}, \mathcal{A})$.

Then G is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α, and we write $G \sim DP(\alpha)$, if for every finite measurable partition $(B_1, B_2, \ldots, B_k)$ of $\mathcal{X}$, the distribution of $(G(B_1), G(B_2), \ldots, G(B_k))$ is Dirichlet with parameter $\alpha(B_1), \ldots, \alpha(B_k)$, denoted by $\mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k))$. It could be proved that G is a random probability measure on $(\mathcal{X}, \mathcal{A})$. Its interest in Bayesian statistics is that the DP can play the role of the prior probability measure for inference in nonparametric problems. To this aim, it is important to define a sample from the random probability measure G on $(\mathcal{X}, \mathcal{A})$. Roughly speaking, the sampling model of size n is described as:

$X_1, X_2, \ldots, X_n \mid G$ are independent and identically distributed (i.i.d) according to G, and $G \sim DP(\alpha(\cdot))$.

From the properties of the Dirichlet distribution, it can be easily shown that $X_i \sim E(G(\cdot)) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$; thus, the normalized base measure $\frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$ has the role of prior guess on the unknown distribution of the X_i.

The first attractive property of the DP, as a prior in Bayesian inference with exchangeble data, is that it is conjugate. That is, if $X_i \mid G \overset{iid}{\sim} G$, and $G \sim DP(\alpha(\cdot))$, then it can be shown that the posterior distribution of G is still a DP, with update parameters, namely, $G \mid X_1, X_2, \ldots, X_n \sim DP(\alpha(\cdot) + \sum_{i=1}^{n} \delta_{X_i})$, where δ_x denotes a probability measure degenerate on x. It follows that

$$E(G(A \mid X_1, X_2, \ldots, X_n) = P(X_{n+1} \in A \mid X_1, X_2, \ldots, X_n) = \frac{\alpha(\cdot) + \sum_{i=1}^{n} \delta_{X_i}(A)}{\alpha(n) + n},$$

a weighted average of the prior guess $\frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$ and the empirical distribution.

1.3 Main Properties of the Dirichlet Process

The theoretical properties of the Dirichlet process explain the central role of the Dirichlet process in Bayesian inference. Here we review some of the main properties. As mentioned before, conjugacy of the DP is a crucial property for applications

in Bayesian statistics.
The discrete nature of the Dirichlet process is described by Blackwell (1973). If G is the Dirichlet process on the space of probability measures $(\mathcal{X}, \mathcal{A})$ with parameter α, then G is almost surely (a. s.) discrete. In fact, the discrete nature of the Dirichlet process appears as a drawback in applications to Bayesian inference with continuous variables. We will discuss in the next sections how a 'continuous prior' can be defined, by means of mixtures of kernel densities with a DP mixing distribution.

An important property for our further developments is the DP characterization proved by Sethuraman (1994).

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let $V_1, V_2, \ldots \overset{i.i.d.}{\sim} Beta(1, M)$ be independent of $X_1, X_2, \ldots, \overset{i.i.d.}{\sim} F_0$ and define $\omega_1 = V_1$ and $\omega_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$ for $i \geq 2$. Then for any $B \in \mathcal{B}$

$$G(B) \overset{a.s.}{=} \sum_{i=1}^{\infty} \omega_i \delta_{X_i}(B) \tag{1.1}$$

where $\delta_x(\cdot)$ stands for the probability measure degenerate at x and δ is the dirac measure at point mass X and G is a Dirichlet process with total mass M and base measure F_0, $G \sim DP(M, F_0)$.
The proportions V_i are sequentially broken from the remaining length $\prod_{j=1}^{i-1}(1 - V_j)$ of a unit length stick. If i increases, these weights stochastically decrease, since smaller fractions of sticks remain, and so only a small number of the infinite number of weights have appreciable value. A common approach is to use $G \sim DP(M, F_0)$ as the mixing distribution with a kernel function as we will discuss later in the DP mixtures models.

The so called stick-breaking representation of the DP, given by equation (1.1), was proved by Sethuraman (1994), even if its basis were already contained in the work of Rolin (1993) see e.g. Müller and Quintana (2004) and before, MacCloskey (1965); see e.g. Pitman (1996).

A different discrete representation of the DP was given by Ferguson (1973); where, the weights are ordered in decreasing way, and have a Poisson-Dirichlet distribution. Both constructions underline important aspects of the DP; however, the stick-breaking representation appears simpler, and it is mathematically important for the Pólya urn characterization of the DP. It is a crucial property in the construction of dependent Dirichlet processes that we will describe later in this chapter, and is the basis of our proposal in the thesis.

1.4 Mixtures of Dirichlet Processes

The Dirichlet process can be regarded as a particular case of a more general class of processes introduced by Antoniak (1974). In fact, as remarked by Antoniak, the Dirichlet process is not flexible enough for modeling some real problems, in particular in the bio-assay field. This motivates the following extension. Follow his formal definition.

Definition (Antoniak, 1974) *Let $(\Theta, \mathcal{A})$ be a measurable space, let $(U, \mathcal{B}, \mathcal{H})$ be a probability space, called the index space, and let α be a transition measure on $U \times \mathcal{A}$. We say that G is a mixture of Dirichlet processes on $(\Theta, \mathcal{A})$ with mixing distribution H on the index space $(U \times \mathcal{B})$ and transition measure α, if for all $k = 1, \ldots$ and measurable partition $A_1, A_2, \ldots, A_k$ of Θ we have*

$$\mathcal{P}\{G(A_1) \leq y_1, \ldots, G(A_k) \leq y_k\} = \int_U \mathcal{D}(y_1, \ldots, y_k \mid \alpha(u, A_1), \ldots, \alpha(u, A_k)) dH(u)$$

where $\mathcal{D}(\theta_1, \ldots, \theta_k \mid \alpha_1, \ldots, \alpha_k)$ denotes a Dirichlet distribution function with parameters $(\alpha_1, \ldots, \alpha_k)$.

Roughly speaking, we say that G is a mixture of Dirichlet processes, with mixing distribution H and transition measure α, if:

$$\begin{aligned} G \mid U &\sim DP(\alpha(\cdot \mid u)) \\ U &\sim H. \end{aligned} \tag{1.2}$$

We will write $G \sim \int DP(\alpha(\cdot \mid u))dH(u)$. Antoniak (1974) proved that mixtures of Dirichlet processes, properties analogous to those of the Dirichlet process hold. In particular, a mixture of Dirichlet processes is discrete almost surely. The coniugacy property of the Dirichlet process still holds for mixtures of the Dirichlet processes. Furthermore, a mixture of Dirichlet processes has a stick-breaking representation. Antoniak (1974) elaborated one of the first empirical applications of the mixtures of the Dirichlet processes and he noted that the extension to mixtures of DPs is necessary in many applications, especially for bio-assay problems.

1.5 Applications and developments: Bayesian nonparametric inference for partial exchangeable data

In the previous sections, we briefly reviewed the basic definitions of the DP and of mixture of DPs. The first applications of these processes in Bayesian statistics were focused on Bayesian nonparametric inference for exchangeable data. However, more complex structure of dependence are involved in many applications. The study of nonparametric inference for dependent data is indicated by the general aim of this work. Therefore, we present here some developments for dependent data, starting from the case of partial exchangeability.

Cifarelli and Regazzini (1978) introduced dependence across related random measures, defining a mixture of products of Dirichlet processes prior. Intuitively, this is a first version of dependent Dirichlet processes, in the sense that there is one more level of dependence across random probability measures. This is a brilliant intuition of the use of one more level of dependence in the mixture of Dirichlet processes defined by Antoniak (1974). In the next sections and chapters, we will use a more general construction of dependent Dirichlet processes, introduced by MacEachern (1999; 2001). There is a strong difference between these two definitions: the mixture of products of the Dirichlet processes is based on the partial exchangeability assumption and mathematically it is based on the mixture of Dirichlet processes defined by Antoniak (1974). In the next chapters we will use the exchangeability assumption and the dependence is introduced on the DP mixture model. We will discuss in more details in the next sections of this chapter about the structural differences between these two strong inferential constructions.

Let $(G_1,\dots,G_k)$ a vector of random probability measures. The problem considered by Cifarelli and Regazzini (1978) is to construct a prior for the random vector, such that the G_i are dependent. Informally, they assume that, conditionally on a vector of random variables $(U_1,\dots,U_k)$, the G_i are independent, with G_i having a Dirichlet process prior, indexed by U_i. That is, $G_1,\dots,G_k \mid U_1 = u_1,\dots,U_k = u_k \sim \prod_{i=1}^{k} DP(\alpha(\cdot;u_i))$; and $(U_1,\dots,U_k) \sim \phi$. More formally, the vector $(G_1,G_2,\dots,G_k)$ is a mixture of products of Dirichlet processes if

$$\mathcal{P}\{\bigwedge_{i=1}^{k}\bigwedge_{j=1}^{m_i-1}(G_i(B_{ij}) \leq y_{ij})\} =$$

$$\int_{\Re^k}\prod_{i=1}^{k}\mathcal{D}(y_{i,1}\dots,y_{i,m_i-1} \mid \alpha_i(B_{i,1},u_i),\dots\alpha_i(B_{i,m_i};u_i))d\varphi(\mathbf{u})$$

where $\mathcal{P}$ is the measure of probability on the space $\prod_1^k[0,1]^{\mathcal{B}}, \prod_1^k B\mathcal{F}^{\mathcal{B}}$, $(B_{i,1}\dots,B_{i,m_i})$ for $i = 1,\dots,k$ the generic measurable partition of $\Re$ and $\varphi(\mathbf{u})$ is the cumulative distribution function of the random variable $(U_1,\dots,U_k) = \mathbf{U}$.

Mixtures of products of Dirichlet processes have been used for partially exchangeable data in many applications. Cifarelli (1979) applied them in Bayesian nonparametric analysis of variance (ANOVA). Cifarelli, Muliere and Scarsini (1981) gave an application in Bayesian nonparametric linear regression. Muliere and Scarsini (1983) illustrated two-way ANOVA through mixtures of products of Dirichlet processes.

Partial exchangeability is the first natural extension of exchangeability, introducing a more structured dependence assumption in Bayesian inference. In a nonparametric framework, problems with partially exchangeable observations are thus the basic examples where the need of dependent nonparametric priors arises. Mixtures of products of DPs are a first proposal to address this issue, and ANOVA an important class of problems where they show to be usefully applied. As it is well known, ANOVA is a special case of the linear regression model and it is indirectly connected to autoregressive models that we will develop in the next chapters.

In principle, partial exchangeability summarizes jointly homogeneity within clusters and heterogeneity between clusters of the population. ANOVA problems are a classical example of partially exchangeable data. Lindley (1970 and 1971) studied Bayesian inference for ANOVA problems, using informative and noninformative priors in the linear model.

In a nonparametric approach, one wants to avoid parametric assumptions on the unknown distributions for the different groups. Cifarelli (1979) developed a Bayesian nonparametric ANOVA model using dependent Dirichlet processes.

The idea is sampling within clusters using Dirichlet processes and introducing dependence across clusters through the dependence among the random group-specific distributions expressed in the prior. Thus, a mixture of products of Dirichlet processes on the random vector $(G_1, \ldots, G_k)$, where G_i is the random distribution corresponding to group i, $i = 1, \ldots, k$, is a natural choice. Let U be a random variable with absolutely continuous distribution function, $U \sim H(\cdot)$. Then the prior is expressed hierarchically as

1. $G_1 \mid u \sim DP(\alpha_1(\cdot, u))$, $G_2 \mid u \sim DP(\alpha_2(\cdot, u)) \ldots G_k \mid u \sim DP(\alpha_k(\cdot, u))$, where

$G_1, G_2 \dots G_k$ are independent given u;

2. $(G_1, G_2, \dots, G_k) \sim \int \mu(G_1, G_2, \dots G_k \mid u) dH(u) =$

   $\int DP(\alpha_1(\cdot \mid U) DP(\alpha_2(\cdot \mid u)) \dots DP(\alpha_k(\cdot \mid u) dH(u)$.

Observe that $G_1, G_2 \dots, G_k$ are defined as Dirichlet processes and they are independent given u. In the second stage, the integral is an analytical consequence of the partial exchangeability assumption. As mentioned before, this prior construction represents the first application of the dependence across DP distributed random probability measures. These steps are the essential elements of the mixture of products of DPs prior. They are based on Ferguson definition for the DP, which is used as the de Finetti measure for the exchangeable observations within a cluster (roughly speaking, it models the variance within clusters), and the mixture of DPs of Antoniak (1974), which models the variance between clusters in the ANOVA scheme. Again, the mixture of products of DP's for the analysis of the variance represents a first example of dependent DPs.

If the DPs in the above model are centered on a Gaussian distribution, Cifarelli (1979) calculated the predictive distribution function. Denoting by $Y_1, Y_2 \dots Y_k$ the results of the future observation for each of the k groups:

$$P\{Y_1 \leq y_1, Y_2 \leq y_2 \dots Y_k \leq y_k \mid \mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_k\} =$$

$$= \int_{\Re^k} \prod_{i=1}^{k} \{ \frac{\alpha(\Re)}{\alpha(\Re) + n_i} \phi(\frac{y_i - u_i}{\sigma_A}) + \frac{n_i}{\alpha(\Re) + n_i} F_{i,n_i}(y_i) \} \varphi^{(k)}(\mathbf{u} \mid \mathbf{x}_1, \mathbf{x}_2 \dots, \mathbf{x}_k) du_1 du_2 \dots du_k$$

where for each group we have a weighted mean of the prior and the empirical cumulative distribution functions. Specifically, $\phi(\frac{y_i - u_i}{\sigma_A})$ is the standardized normal cdf; the random cumulative distribution functions, F_{i,n_i}, are selected by the mixture of products of DPs and are not independent in probability. The weights, $\alpha(u_i, \Re) = \alpha(\Re)$ are proportional to the subjective prior and it is completly arbitrary. If the elements of the vector, $(U_1, U_2, \ldots, U_k)$, are normal distributed and exchangeables, but not independent, then the final distribution is $\varphi^{(k)}(\mathbf{u} \mid \mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_k)$.

The generic marginal distribution is

$$P\{Y_i \leq y \mid \mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_k\} = \frac{\alpha(\Re)}{\alpha(\Re) + n_i} \phi(\frac{y - t_i}{\sqrt{\sigma_i^2 + \sigma_A^2}}) + \frac{n_i}{\alpha(\Re) + n_i} F_{i,n_i}(y) \tag{1.3}$$

in equation (1.3) we have the mixture of normal cdf centered on the mean t_i and divided by the variance $(\sigma_i^2 + \sigma_A^2)$ and the empirical cdf. Note that the cdf depends from the whole observations $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_k$ and not only from the i-th population, this is a consequence of the exchangeability assumption on the subgroups of the population $(U_1, U_2, \ldots, U_k)$. Observe that Bayesian estimations do not depend from the single variances σ_A^2 and σ_B^2, but they tend to $\bar{x}$ if $\frac{\sigma_A^2}{\sigma_B^2} \to 0$. Moreover t_i is the weighted average among the mean of distinct observations of i-th population, $\bar{x}_i$ and the general mean $\bar{x}$. These results are similar to Lindley (1971) when the groups of the population tend to be homogenous and the hypothesis of partial exchangeability tends to be close to the exchangeability assumption. Technically, Cifarelli (1979) obtained distinct observations $r_i \leq n_i$ instead of n_i and the mean of the clusters is for distinct observations $\bar{x}_i$. If $\alpha(\Re) \to +\infty$ then $r_i \to n_i$ and only this specific case Cifarelli's Bayesian estimators are equal to Lindley's results. In addition, if the cdf $(F_1, F_2, \ldots, F_k)$ are normal distributed, then nonparametric model corresponds to the parametric model defined by Lindley (1971).

1.5.1 Bayesian Nonparametric Linear Model

ANOVA and, more generally, regression models are examples that require one more level of dependence in the prior for Bayesian nonparametric inference. ANOVA is an example of symmetric dependence: it describes a relationship across the groups of a population. Linear regression is an example of asymmetric dependence: the outcomes depend on explanatory variables. We illustrate an application of the mixture

of products of Dirichlet processes prior to Bayesian nonparametric regression. Let $(X_1, X_2, \ldots, X_k, Y)$ be a $(k+1)$-dimensional random vector and $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$ n observations for $i = 1, 2, \ldots, n$.
Consider the multivariate linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.4}$$

where $\mathbf{Y}$ is the $(n \times 1)$ vector of outcomes, $\mathbf{X}$ is the $(n \times k)$ design matrix, $\boldsymbol{\beta}$ is the $(k \times 1)$ vector of parameters and ϵ the $(n \times 1)$ vector of errors.
In classical analysis, the ordinary least squares (OLS) estimator is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, which corresponds to the maximum likelihood estimator, if the errors vector is normally distributed. Stein (1962) proposed a ridge estimator which is a Bayesian estimator. Lindley (1962) provided an interpretation of this estimator in a Bayesian hierarchical model.

Cifarelli, Muliere and Scarsini (1981) described a Bayesian nonparametric estimator for a multivariate linear model, assuming partial exchangeability. Model (1.4) can be written as the following equations:

$$\begin{aligned}
\mathbf{y}_1 &= \mathbf{1}_{n_1}\mathbf{x}_1'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \\
\mathbf{y}_2 &= \mathbf{1}_{n_2}\mathbf{x}_2'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \\
\vdots \quad & \quad \vdots \quad \vdots \quad \vdots \\
\mathbf{y}_k &= \mathbf{1}_{n_k}\mathbf{x}_k'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k
\end{aligned}$$

The cumulative distribution functions, $F_1, F_2, \ldots F_k$, and the vectors $\mathbf{y_1}, \mathbf{y}_2, \ldots, \mathbf{y}_k$

are mutually independent. The joint cumulative distribution function is:

$$P(\mathbf{y}_1 \leq \boldsymbol{\xi}_1, \dots, \mathbf{y}_k \leq \boldsymbol{\xi}_k \mid X, \boldsymbol{\beta}, F_1, F_2, \dots, F_k) =$$

$$\prod_{j=1}^{n_1} F_1(\xi_{1,j}) \prod_{j=1}^{n_2} F_2(\xi_{2,j}) \dots \prod_{j=1}^{n_k} F_k(\xi_{k,j})$$

The cumulative distribution functions $F_1, F_2, \dots F_k$ are random and and they are given a mixture of products of Dirichlet processes prior. Namely, given $\mathbf{x}_i$ and $\boldsymbol{\beta}$, the cumulative distribution function F_i is a Dirichlet processes with parameter $\alpha_i(\mathbf{x}_i\boldsymbol{\beta}, \cdot)$ and $F_1, F_2, \dots, F_k \mid X, \boldsymbol{\beta} \sim \prod_{i=1}^{k} \mathcal{D}_i = DP(\alpha_i(\cdot; \mathbf{x}_i\boldsymbol{\beta})$, where

$$\alpha_i(\mathbf{x}_i, \boldsymbol{\beta}, \xi) = \alpha(\Re)\phi(\frac{\xi - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_i})$$

where ϕ is the cumulative distribution function of a standardized normal distribution. The cumulative distribution function for the vector $\boldsymbol{\beta}$, $\varphi(\boldsymbol{\beta})$, is such that

$$F_1, F_2, \dots, F_k \mid X \sim \int \prod_{i=1}^{k} \mathcal{D}_i d\varphi(\boldsymbol{\beta}). \tag{1.5}$$

In equation (1.5), the integral is with respect to a cumulative distribution function. Later in this chapter, we will consider Dirichlet process mixtures, where one has a distribution instead of a product of DPs inside the integral, and the integration will be with respect to a random probability measure.

The prior distribution for the parameter $\boldsymbol{\beta}$ is chosen as the conjugate prior for the Gaussian base measure of the DP:

$$\boldsymbol{\beta} \sim N(\mathbf{1}\beta_0, \sigma_\beta^2\mathbf{I}_p), \quad \sigma_\beta^2 > 0.$$

For the conjugacy property, one finds the following posterior density for $\boldsymbol{\beta}$:

$$\varphi^{(p)}(\boldsymbol{\beta} \mid \beta_0\sigma_\beta^2, \mathbf{y}_1, \dots \mathbf{y}_k, \mathbf{x}_1, \mathbf{x}_k) \propto$$

$$\prod_{i=1}^{k}\prod_{j=1}^{r_i} \phi'(\frac{y_{ij} - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_i})\varphi^{(p)}(\boldsymbol{\beta} \mid \beta_0, \sigma_\beta^2) \propto$$

$$e^{-\frac{1}{2}(\tilde{\mathbf{y}}-\tilde{\mathbf{x}}\boldsymbol{\beta})'\Sigma^{-1}(\tilde{\mathbf{y}}-\tilde{\mathbf{x}}\boldsymbol{\beta})+(\boldsymbol{\beta}-\mathbf{1}\beta_0)'(\boldsymbol{\beta}-\mathbf{1}\beta_0)\sigma_\beta^{-2}}$$

where r_i is the number of distinct observations in $(Y_{i1}, Y_{i2}, \dots, Y_{in_i})$, $\tilde{\mathbf{y}}$ is the vector of distinct elements of $\mathbf{y}$ and it is the same for the design matrix $\tilde{\mathbf{X}}$.
$\boldsymbol{\Sigma}$ is a blocked matrix such that $\Sigma_i = \sigma_i^2 Ir_i$ for $i = 1, 2, \dots, k$ and $r = \sum_{i=1}^{k} r_i$,
$\begin{pmatrix} \Sigma_1 & \\ & \ddots\ \Sigma_k \end{pmatrix}$.

Therefore, the posterior distribution for the vector of the parameters $\boldsymbol{\beta}$ is normal and the mean and the variance are respectively equal to $\mathbf{b}$ and $\mathbf{V}$, where

$$\mathbf{b} = E(\mathbf{b} \mid \beta_0, \sigma_\beta^2, \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_1, \dots, \mathbf{x}_k) = (\tilde{X}'\Sigma^{-1}\tilde{X} + \sigma_\beta^{-2}I_p)^{-1}(\tilde{X}'\Sigma^{-1}\tilde{X}\hat{\boldsymbol{\beta}} + \sigma_\beta^{-2}\mathbf{1}\beta_0)$$

where $\hat{\beta} = (\tilde{X}'\Sigma^{-1}\tilde{X})^{-1}\tilde{X}'\Sigma^{-1}\tilde{\mathbf{y}}$ and the covariance matrix is

$$V = Cov(\beta \mid \beta_0, \sigma_\beta^2, \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_1, \dots, \mathbf{x}_k) = (\tilde{X}'\Sigma^{-1}\tilde{X} + \sigma_\beta^{-2}I_p)^{-1}.$$

The Bayesian estimate of $\boldsymbol{\beta}$ is equal to the weighted average of the least square estimation $\hat{\mathbf{b}}$ and the expected value of the initial distribution, $\mathbf{1}\beta_0$.
If $\sigma_\beta^{-2} \to 0$ the initial distribution of $\boldsymbol{\beta}$ is the Jeffreys prior and in this case the posterior distribution is equal to the ordinary least squares (OLS) estimator: $\mathbf{b} = \hat{\boldsymbol{\beta}}$ and $V = (\tilde{X}'\Sigma^{-1}\tilde{X})^{-1}$.

If $\boldsymbol{\beta}_0$ is unknown, there are distinct observations, $r_1, r_2, \dots, r_k$, and not the sample size $n_1, n_2, \ldots, n_k$. The estimation of $\boldsymbol{\beta}_0$ does not depend apparently from $\alpha(\Re)$, which is the initial weight for different normal distributions. In addition, $\alpha(\Re)$ influences the number of distinct observations for each class: if $\alpha(\Re) \to +\infty$, than $r_i \to n_i$ almost surely and in this case these estimations are equal to Korwar and Hollander's results. However Cifarelli, Muliere and Scarsini (1981) choose $(F_1, F_2, \ldots F_k)$ such that the normal distributions and the nonparametric model are confused in Korwar and Hollander's parametric model.

Indeed, the predictive distribution function is similar to Cifarelli result.

Let $Y_1, Y_2, \ldots, Y_k$ be the results for k future observations in each of the k classes (or groups): one future observation for each class. The corresponding cumulative distribution function is:

$$\psi(y_1, y_2, \ldots, y_k) = Pr(Y_1 \leq y_1, Y_2 \leq y_2, \ldots, Y_k \leq y_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_k, \mathbf{x}_1, \ldots, \mathbf{x}_k).$$

So the linear equation is

$$\psi(y_1, y_2, \ldots, y_k) = \int_{\Re^p} \prod_{i=1}^{k} [\frac{\alpha(\Re)}{\alpha(\Re) + n_i} \phi(\frac{y_i - \mathbf{x}_i' \boldsymbol{\beta}}{\sigma_i}) + \frac{n_i}{\alpha(\Re) + n_i} \hat{F}_{i,n_i}(y_i)] \varphi^{(p)}(\boldsymbol{\beta} \mid \cdot)$$

where $\hat{F}_{i,n_i}(y_i)$ is the empirical cumulative distribution function computed on the observations of $i - th$ class $y_{i1}, y_{i2}, \ldots, y_{in_i}$ and $\varphi^{(p)}$ is one of the final distributions for $\boldsymbol{\beta}$.

Suppose that

$$\psi^{(p)}(\boldsymbol{\beta} \mid \cdot) = (2\pi)^{-p/2} \mid \mathbf{V} \mid^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})' V^{-1} (\boldsymbol{\beta} - \mathbf{b})}$$

where the general p-th conditional posterior distribution for $\boldsymbol{\beta}$

$$\mathbf{b} = (\tilde{X}'\Sigma^{-1}\tilde{X} + \sigma_{\beta}^{-2}I)^{-1}(\tilde{X}'\Sigma^{-1}\tilde{X}\tilde{\boldsymbol{\beta}} + \sigma_{\beta}^{-2}\mathbf{I}\beta_0)$$

$$\mathbf{V} = (\tilde{X}'\Sigma^{-1}\tilde{X} + \sigma_{\beta}^{-2}I)^{-1}$$

The predictive cumulative distribution function for the marginal Y_i is

$$\psi_i(y) = \frac{\alpha(\Re)}{\alpha(\Re) + n_i}\int_{\Re^p}\phi(\frac{y - \mathbf{x}_i\boldsymbol{\beta}}{\sigma_i})\varphi^{(p)}(\boldsymbol{\beta} \mid \cdot)d\boldsymbol{\beta} + \frac{n_i}{\alpha(\Re) + n_i}\hat{F}_{i,n_i}(y)$$

so

$$\psi_i(y) = \frac{\alpha(\Re)}{\alpha(\Re) + n_i}\phi(\frac{y - \mathbf{x}_i\boldsymbol{\beta}}{\sigma_i^*}) + \frac{n_i}{\alpha(\Re) + n_i}\hat{F}_{i,n_i}(y)$$

where

$$\sigma_i^{2*} = \sigma_i^2(1 - \mathbf{x}_i'(\mathbf{x}_i\mathbf{x}_i' + \sigma_i^2\mathbf{V}^{-1})^{-1}\mathbf{x}_i)^{-1} = \sigma_i^2 + \mathbf{x}_i'\mathbf{V}\mathbf{x}_i \tag{1.6}$$

Muliere and Scarisini (1983) considered a Bayesian linear parametric model that under some conditions correspond to this nonparametric construction that we reviewed in this section.

1.6 Dirichlet Process Mixtures

The discretness of the DP and of mixtures of DPs may be a drawback in inference for continuous data. We present here a further development given by Dirichlet Process mixtures, from the beginning untill recent papers. We will intend the central role of DP mixtures for the developments of the next chapters. The essential stucture of the DP mixture model is described as follows.

1. $X_i \mid G \overset{iid}{\sim} \int f(x_i \mid \theta) dG(\theta)$ for $i = 1, 2, \ldots$
2. $G \sim DP(\alpha)$ where $\alpha = MG^0$

Note the difference with Antoniak definition of mixtures of Dirichlet processes. Here, the conditional density of the data is a mixture of kernels $f(x \mid \theta)$, and the mixing distribution is a DP.

Ferguson (1983) studied DP mixtures of Gaussian distributions for Bayesian density estimation. We report briefly his results as first example of DP mixture models.

Let $f(x)$ be the mixture:

$$f(x) = \sum_{i=1}^{\infty} p_i h(x \mid \mu_i^*, \sigma_i^{2*}) \tag{1.7}$$

where $h(x \mid \mu_i^*, \sigma_i^{2*})$ denotes the density of normal distribution $N(\mu_i^*, \sigma_i^{2*})$ with mean μ_i^* and variance σ^{2*}. To obtain identifiability it is possible to rewrite equation (1.7) in terms of

$$f(x) = \int h(x \mid \mu, \sigma) dG(\mu, \sigma^2) \tag{1.8}$$

where G is the probability measure on $\{(\mu^*, \sigma^{2*}) : \sigma^{2*} > 0\}$ that gives point masses p_i at atoms (μ_i^*, σ_i^*). Ferguson (1983) chose the prior distribution for the hyperparameters $(p_1, p_2, \ldots, \mu_1, \mu_2, \ldots, \sigma_1^2, \sigma_2^2 \ldots)$ such that the distribution function, G, is a Dirichlet process with base measure $\alpha = MG^0$. Indeed, if $G \sim DP(\alpha, G^0)$, by the stick-breaking representation, presented in equation (1.1), G is almost surely equal to $G = \sum_{i=1}^{\infty} p_i \delta_{(\mu_i^*, \sigma_i^{2*})}$, where the atoms (μ_i^*, σ_i^{2*}) are i.i.d. according to G^0, and the weights (p_i) have a stick-breaking prior with parameter α, independently on the (μ_i, σ_i^2). Therefore, the mixture in equation (1.8) reduces to the countable mixture 1.7, with these priors on the weights and components parameters. Ferguson suggested to choose $G^0 = E(G)$, as the conjugate prior for (μ, σ^2), i.e. Normal-Gamma distribution. The model can be equivalently reformulated as

1. $X_i \overset{ind}{\sim} h(x \mid \theta_i)$, where $\theta_i = (\mu_i, \sigma_i^2)$ for $i = 1, 2, \dots n$

2. $\theta_i \mid G \overset{iid}{\sim} G$ and

3. $G \sim DP(\alpha G^0)$.

It can be shown that, given $(X_1, X_2, \dots, X_n)$, the posterior distribution of G is a mixture of Dirichlet processes in the sense of Antoniak, as we discussed in the third section of this chapter:

$$G \mid X_1, \dots, X_n \sim \int \dots \int DP(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}) dH(\theta_1, \theta_2, \dots, \theta_n \mid X_1, X_2, \dots, X_n)$$

Ferguson (1983) suggested to use the normal distribution as kernel function in the mixture model, and DP as mixing distribution. Lo (1984) described a general continous kernel function $K(\cdot, \cdot)$; the case of Gaussian kernel studied by Ferguson (1983) is thus a special case. However, Lo (1984) analyzed the choice of the kernel function, which defines the model. In his mathematical description, there are all the elements of the DP mixture model that we summarized briefly at the beginning of this section. Indeed, Lo (1984) defined the conditional distribution $f(x \mid G) = \int_{\Re} K(x, u) G(du), x \in \mathcal{X}, u \in \Re$, where $f(\cdot \mid G)$ is a density function by virtue Fubini‘s theorem. Observe that this property will be used in the third Chapter for two latent variables in our proposal. The other important property is the marginal density of x, $\int_{\Theta} f(x \mid G) \mathcal{P}_{\alpha}$, for each $x \in X$, is equal to $\int_{\Re} K(x, u)(\alpha(du)/\alpha(\Re))$, that we will use for inference in the next chapters. Lo (1984) and independently Kuo(1980) and Ghorai and Rubin (1982) investigated also the conditional expected value given the observations for $f(x \mid G)$, which is the core of Bayesian applied inference.

While theoretically interesting for Bayesian density estimation and many other nonparametric problems, inference for DP mixture is analytically complicated, and its application was initially limited to problems with fairly small sample size. The use of Markov Chain Monte Carlo (MCMC) simulation techniques in the 90‘s gave great impulse to the application of Bayesian nonparametric procedures. West (1993) tried to extend the DP mixture schemes for different kind of kernel functions in dynamic models. He also discussed the opportunity to evaluate the approximation of the posterior for the DP mixtures using Monte Carlo methods. Escobar (1994) and Escobar and West (1995) suggested a Gibbs sampling to simulate from the posterior distribution in DP mixtures of Gaussians. A more efficient Gibbs sampler algorithm was proposed by Müller and MacEachern (1998) who also extended DP mixtures to non-conjugate base measures.

1.7 Dependence across a Collection of Random Probability Measures

In the previous sections, we discussed dependence through random probability measures, and the different roles played by the Dirichlet process. Here we review two other possible constructions of dependent random measures. We illustrate the dependence for a collection of random probability measures. The following two sections represent the connections between the results that we described briefly in the previous sections and more recent research literature, which will be the basis for the structure of the next chapters. We will focus on the definition of the dependent Dirichlet process (DDP), which is the starting point for our research, and the ANOVA DDP, that we will extend for modeling time series.

1.7.1 Dependent Dirichlet Process: definition and main properties

In the previous review of nonparametric priors, we considered only one random probability measure, which is, in the basic case, a Dirichlet process. We introduce a collection of random probability measures, such that, marginally, each element is a Dirichlet process. MacEachern (1999, 2001) defined as *dependent Dirichlet process*, (DDP), underlying the relevance of the dependence between random probability measures. For the definition of DDPs, there are two basic results: the stick-breaking representation of the DP and the mixture of Dirichlet processes. The dependent Dirichlet process is based on the exchangeability assumption on the data. As we discussed in the previous section, Cifarelli and Regazzini (1978) proposed a DDP prior assuming partial exchangeability.

Mathematically, the dependent Dirichlet process is a prior probability law for a collection of random probability measures, $\{G_y\}$, such that each G_y is marginally a Dirichlet process. Let G be a random distribution function, $G_y \sim DP(M, G_y^0)$, where M is the total mass parameter and G_y^0 the base measure. By Sethuraman's representation it is possible to write down G as: $G_y = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h}(y)$ where $\sum_{h=1}^{\infty} \omega_h = 1$. The Dirichlet process places a prior on the space of distribution functions by creating a distribution on θ_h and ω_h. Such distribution is governed by the parameter of the DP, α, which is the measure absolutely continous with respect to Lebesgue measure. The total mass of the measure α is denoted by M and the shape of α is described by the probability measure or the corresponding distribution function $G^0(\cdot) = \alpha(\cdot)/M$.

The important assumptions for the DDP are:

1. θ_h and ω_h are mutually independent;

2. θ_h are independent and identically distributed as G^0.

There are three kind of different and alternative possible frameworks for dependent prior models. The dependence across the collection of random probability measures can be chacterized as follows. The simplest case of the DDP model is the *single-p DDP*, which has common weights, ω_h and dependent locations, $\theta_h(y)$. The dependence of the random probability measures is modeled only on the locations, it means that across the random probability measures can be inserted dependence. The stick-breaking representation of this process is $G_y = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h}(y)$. An other form of dependence is the $varying$ weights DDP, note that the locations are common and dependent weights, $\omega_h(y)$, it means that on the weights we add one more level of dependence. In this case the stick-breaking representation is $G_y = \sum_{h=1}^{\infty} \omega_h(y) \delta_{\theta_h}$, so the weights are varying with the variable y. The last construction introduces dependence on the weights, $\omega_h(y)$ and jointly on the locations, $\theta_h(y)$. This last framework for the DDP is less flexible to respect to the other proposals, for the high level of dependence. Indeed, the weights and the locations are varying with y, which are modeled as jointly functions of y. For many applications of the DDP are used the single-p DDP or the varying weights DDP for flexibility and parsimony of the number of parameters involved in the model and the capacity of describing dependence. We will use the varying locations and varying weights in Chapter 4 for an application to a real dataset. Gelfand *et al.* (2005) illustrated that the varying weights and varying locations DDP can be seen as a ‘limited‘ Gaussian process for spatial data. They proposed a nonstationary and neither Gaussian spatial Dirichlet process mixture model such that the varying weights and the varying locations DDP is a Gaussian process, but the stationary property has to be guaranteed.

Here we just mention one of the limits of the three DDP models is to establish the number of finite mixtures in the stick-breaking representation. Muliere and Tardella (1998) proposed a method for approximating the distributions also Ishwaran and Zarepour (2002) advanced a different solution. However, we will illustrate more details of this aspect in the next chapters.

Some of the main properties of the DDP, which can be used in the next chapters, has been introduced by MacEachern (1999).

1. The prior distribution on $G_{y_1}, G_{y_2}, \dots, G_{y_d}$ has full support, provided by the stochastic process θ_y is rich enough.

2. The DDP models are amenable to simulation based fits.

3. The marginal distribution, G_y follows a well-known distribution. In fact, $G_y \sim DP(M_y, G^0_y)$ for each $y \in \mathcal{Y}$.

4. The distributions G_y are continous in y. This feature is what produces distributions that evolve as the covariate changes. It can be obtained by working with stochastic processes which produce continous paths for θ_y and ω_y.
   G_{y_1} and G_{y_2} tend toward independent distributions as y_1 and y_2 become more distant. To accomplish this, we need θ_{y_1} to tend toward independence from θ_{y_2} and also V_{y_1} to tend toward independence from V_{y2}. This can be accomplished by writing stochastic processes which yield the decay toward independence.

5. In addition, a spectrum of inference can be captured, ranging from nearly parametric inference (take M_y nearly ∞ for all values of y) to inference that relies on a single nonparametric distribution (take the stochastic process θ_y for which $\theta_{iy} = \theta_h$ for all y) to inference that shows a strong dependence between distributions with nearby y (take slowly varying stochastic processes for θ_y and/or V_y) to inference that encourages quick changes in the distributions as y changes (take θ_y that change quickly in y).

The dependent Dirichlet process is useful for generalized linear models and can be also rewritten in terms of a Bayesian hierarchical model, which is an other inferential aspect for more levels of dependence. In particular, the single-p DDP has easy implementation in a Gibbs sampler algorithm.

The dependence of the DDP can be modeled as a linear regressive model. Let $Y_i = X_i\beta + \varepsilon_i$ for $i = 1, 2, \ldots, n$ be the linear regression and the errors are Gaussian distributed, $\varepsilon_i \sim N(0, \sigma^2)$. MacEachern (2001) suggested to replace the random sample of normal variables by a sample of independent variables, i.e. $\varepsilon_i \sim F_{y_i}$ for $i = 1, 2, \ldots, n$. The approximated continuity of the outcome can be adjusted by an additional hierarchical level and the distribution of the outcome is smoothed.

An other possible application of the DDP for linear models is the distribution of the errors as $x_i\beta$. The residuals correspond to a sample of independent variables, where $\varepsilon_i \sim F_{x_i\beta}$. In this second proposal for linear models, there is a reduction of the dimension of the space.

1.7.2 ANOVA and Dependent Dirichlet Process prior

One of the most popular variations of dependent prior model is the dependent Dirichlet process for analysis of variance, briefly, ANOVA DDP. De Iorio *et al.* (2004) proposed a model to describe dependence across random probability measures in an analysis of variance (ANOVA)-type fashion. They defined a probability model such that marginally the random probability measures follow a Dirichlet process and the dependent Dirichlet process describes the dependence across the related random probability measures.

Here, we introduce more details of this model, which is one of the starting point for the methodological aspects that we will present in the next chapters.

Let F_y be a random distribution for a p-dimensional vector $y = (y_1, y_2 \ldots, y_p)$ of categorical covariates. De Iorio *et al.* (2004) defined the model for F_y such that marginally to respect y the random distribution F_y is defined on the class $\mathcal{F} = \{F_y, y \in \mathcal{Y}\}$ and G_y is a Dirichlet process, $G_y \sim DP(M, G_y^0)$ with total mass parameter M and base measure G_y^0. The dependence across y is defined by the dependent Dirichlet process (MacEachern, 1999, 2001) mathematically this is $(G_y, y \in \mathcal{Y}) \sim DDP(M, G_y^0)$. The random measure G_y has the first property of a DP, being almost surely discrete with point masses generated marginally from the base measure G_y^0. The choice of a DDP prior means that there is dependence across y in the distributions of these point masses. The ANOVA model describes dependence through the trajectories for the point masses [1]. This model is based on DP prior

[1] Note that we indicate as ‘trajectories *for* the point masses‘ and not the ‘trajectories *of* the point masses‘, this small difference means that we study the dependence *across* the random probability

distribution and the DP, under these assumptions, is a probability model for random probability measures. The random measure $G_y \sim DP(M, G_y^0)$ is discrete a. s. and can be represented as a stick-breaking. In particular, the random probability measure, G_y, is $G_y = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h(y)}$, where ω_h are common weights for the point masses and $\theta_h(y)$ the locations, which are the dependent prior model defined as single-p DDP prior. The usual prior distribution for the weights is a rescaled Beta distribution $\omega_h / \prod_{i=1}^{h-1}(1-\omega_i) \sim Beta(1, M)$ and the locations θ_h are independent and identically distributed samples from the base measure G^0. However, in a lot of data analysis applications the discreteness of DP is inappropriate. A solution is the DP mixture models involving a continous kernel convolution. The DP mixture model is:

$$y_i \overset{iid}{\sim} H, \quad with\ H(y) = \int f(y \mid \mu) dG_y(\mu),$$

$$G_y \sim DP(M, G_y^0)$$

One of the most used continous kernel functions is the normal distribution for the conjugacy property:

$$f(y \mid \mu) = N(\mu, \sigma^2)$$

where μ and σ^2 are the average and the variance, respectively. Moreover, the function $H(y)$ leads a mixture of normals $H(y) = \sum_{h=1}^{\infty} \omega_h N(\mu_h, \sigma^2)$. Note the methodological difference of this proposal and Cifarelli's ANOVA DDP based on the mixture of the products of Dirichlet processes. In particular, the role of the base measures G: in the previous work the vector of the base measures were used as products of Dirichlet processes. Here we have G as mixing distribution. In this proposal there is the exchangeability assumption and it is an other important difference to respect

measures, in other words, the point masses are represented by the trajectories. More details we will show through the first plot in the next chapter.

Cifarelli (1979) which used the partial exchangeability assumption. This is just an application of the DP mixture. The statistical contribution is an explicit application of one more level of dependence in the DDP framework. The DDP construction allows different levels of dependence. MacEachern (2001) proposed a simple linear model as dependent prior model. The advance of the ANOVA DDP is the dependence of the dependent Dirichlet process under the exchangeability assumption. This is one of the most evident differences with Cifarelli (1979).
However, the ANOVA DDP can be seen also in terms of DDP for linear regression on a covariate y. MacEachern (1999) considered a family of random distributions on the class $\mathcal{F} = (F_y, y \in \mathcal{Y})$ indexed by a covariate of y. The probability model for $\mathcal{F}$ is marginally such that $G_y = \sum \omega_h \delta(\theta_{yh})$ follows a DP. De Iorio *et al.* (2004) added y on the point masses θ_{yh} to indicate dependent point masses in the random measure G_y. In the basic DDP model the weights ω_h are common for all the depending distributions F_ys held in the depending locations. The DDP model induces dependence across y by assuming that $\theta_h = (\theta_{hy}, y \in \mathcal{Y})$ are i.i.d. realizations of a stochastic process in y. In addition, independence across h and stick-breaking representation for the weights ω_h, garantees that G_y marginally follows a DP. Dependence in the sample path of the stochastic process θ_h introduces the desidered dependence across y. The DDP structure is the base for the ANOVA-like probability model over an array of random probability measures. They assume on class $\mathcal{F} = (F_y, y \in \mathcal{Y})$ an array of random distributions, indexed by categorical covariates y. In the next chapter we will discuss about the limit of categorical covariates and we will illustrate that it is possible to use continous variables for time series analysis.

1.8 Discussion

Recent statistical research is focused on complex problem sets such that known inferential methods are not completly able to produce satisfactory results. More aspects of the same problem can be described introducing dependence. The main idea of the use of the dependence is to keep jointly more aspects of the same phenomenon and share information. In Bayesian statistics simple priors can be assumed for simple problems, but for more complex problems depending priors are the methodological direction. In the last two sections, we underlined recent studies and applications which involved a collection of random probability measures that marginally have a DP prior. In addition, we illustrated the relevance of the DP mixture model and its flexible form for our further applications. However, we discussed the difference of theoretical developments of the beginning of 80's and the computational revolution of 90's. The aim of this chapter is a brief background of the principal roles of the DP and an historical excursus of the inferential evolution of the Bayesian nonparametric statistics. We reviewed here also recent statistical literature as the DDP and the ANOVA DDP which are the basical knowledges for better understanding the next chapters. Here we also clarified the substantial differences between the mixture of DPs and the DP mixtures, which are the two fundamental types of inferential methods proposed in 80's and 90's, respectively. The prediction is the core of the difference. The mixture of DPs is based on the mixture of distinct random probability measures, i.e. the ANOVA mixture of the product of DPs. Indeed, the DP mixture, i.e. ANOVA DDP, is a collection of random probability measures such that marginally is a DP and across the random probability measures there is dependence modeled as an ANOVA. The ANOVA is not the only possible way to describe dependence across depending prior models. In the next chapters we will illustrate autoregressive models such that these are able to depict dependence across all the three frameworks of DDP.

Bibliography

Antoniak, C. E. (1974). ”Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems”. *Annals of Statistics*, **2**, 1152-1174.

Blackwell, D. (1973). ”Discretness of Ferguson selections”. *Annals of Statistics*, **1**, 356-358.

Cifarelli, D. M. (1979). ”Impostazione Bayesiana di un Problema di Analisi della Varianza con Approccio Nonparametrico”. *Tech. Rep. Universitá di Torino*

Cifarelli, D. M. and Muliere, P. (1989). ”Statistica Bayesiana”. *Gianni Iuculano Editore*.

Cifarelli, D. M. Muliere, P. and Scarsini, M. (1981). ”Il Modello Lineare nellÁpproccio Bayesiano Non Parametrico.” *Istituto G. Castelnuovo.* Roma.

Cifarelli, D. M. and Regazzini, E. (1978). ”Problemi statistici non parametrici in condizioni di scambiabilit parziale: impiego di medie associative” *Quadermi Istituto di Matematica Finanziaria*. Serie III, nn. 12, Universitá Torino.

De Iorio, M., Müller, P., Rosner, G. R. and MacEachern, S. N., (2004). ”An ANOVA Model for Dependent Random Measures.” *Journal of the American Statistical Association*, **99**, 205-215.

Escobar, M. D. (1994). ”Estimating Normal Means with a Dirichlet Process Prior.” *Journal of American Statistical Association*, **425**, 268-277.

Escobar, M. D. and West, M. (1995). ”Bayesian Density Estimation and Inference using Mixtures”. *Journal of American Statistical Association*, **90**, 577-588.

Ferguson, T. S. (1973). ”A Bayesian Analysis of some Nonparametric Problems”. *The Annals of Statistics*, **1**, 209-230.

Ferguson, T. S. (1983). ”Bayesian Density Estimation by Mixtures of Normal Distributions”. *Recent Advices in Statistics*, eds. H. Rizvi and J.Rustagi, New York: Academic Press, 287-302.

Gelfand, A. E. Kottas, A. and MacEachern, S. N. (2005). ”Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing”. *Journal of American Statistical Association*. **475**, 1021-1035.

Ghorai, J. K. and Rubin, H. (1982). ”Bayes risk consistency of nonparametric Bayes density estimates”. *Australian Journal Statistics.*

Korwar, R. M. and Hollander, M. (1973). ”Contributions to the theory of Dirichlet processes”. *Annals of Probability*. **1**, 705-711.

Kuo, L. (1980). ”Computations of mixtures of Dirichlet processes”. Tech. rep., Department of Statistics. University of Mitchigan, A. Arbor, **96**.

Lindley, D. V. (1970). ”Introduction to probability and Statistics”. *Cambridge University Press*. **2**.

Lindley, D. V. (1971). ”The estimation of many parameters”. *Foundation of Statistical Inference*.

Lindley, D. V. and Smith, F. M. (1972). ”Bayes Estimates for the Linear Model.” *Journal of the Royal Statistical Society. Series B*. **34**, No. 1, 1-41.

Lo, A. Y. (1984). ”On a Class of Bayesian Nonparametric Estimates: I. Density Estimates”. *The Annals of Statistics*,**1**, 351-357.

MacEachern, S. N. (1999).”Dependent nonparametric processes”. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, V.A.. American Statistical Association.

MacEachern, S. N. (2000). ”Dependent Dirichlet processes”. Tech. rep., Department of Statistics. The Ohio State University.

McCloskey, J. W. (1965). ”A model for the distribution of individuals by species in an environment”. Ph.D. thesis, Michigan State University.

Muliere, P. and Scarsini, M. (1982). ”Il Modello Lineare: Inferenza e Previsione”. *Studi statistici n.1*. Universitá Bocconi.

Muliere, P. and Scarsini, M. (1983). ”Impostazione Bayesiana di un Problema di Analisi della Varianza a Due Criteri”. *Giornale degli Economisti e Annali di Economia.* **2**.

Müller, P., West, M. and MacEachern, N. S. (1997). ”Bayesian Models for Non-Linear Autoregressions”. *Journal of Time Series Analysis*. **18**, 593-614.

Müller, P. and Quintana, F. (2004). ”Non Parametric Bayesian Data Analysis”. *Statistical Science*, **19**, 95-110.

Pitman, J. (1996). ”Some developments of the Blackwell-MacQueen urn scheme”. *Statistics, Probability and Game Theory*. **30**, 245-267.

Rolin, J. M. (1993). ”On the Distribution of the Distribution of the jumps of the Dirichlet Process”. Technical Report. Catholique de Louvain. Institut de statistique.

Sethuraman, J. (1994). ”A Constructive Definition of the Dirichlet Process Prior”. *Statistica Sinica*. **2**, 639-650.

Sethuraman, J. and Tiwari, R.C. (1982). ”Convergence of Dirichlet Measures and the Interpretation of Their Parameters”. *Statistical Decision Theory and Related Topics III*. **2**, 305-315. eds. S. S. Gupta and J. O. Berger, New York: Academic Press.

Smith, C. A. B. (1970). ”Discussion of a paper by A. W. F. Edwards.”. *Journal of Royal Statistical Society, B.* **32**, 165-166.

Smith, A. F. M. (1973). ”Bayes Estimates un One-Way and Two-Way Models”. *Biometrika*, **60**, 2. 319-329.

Stein, C. (1966). ”An approach to the recovery of inter-block information in balanced incomplete block designs.” *Festschrift for J. Neyman: Research Papers in Statistics (F. N. David, ed.)*. 351-366. New York: Wiley.

Chapter 2

A Bayesian Nonparametric Autoregressive Model

Abstract.

We propose a Bayesian nonparametric autoregression for a sequence $(Y_t, t \geq 1)$*. We assume* $(Y_t \mid Y_{t-1} = y) \sim F_y$ *for a family of random probability measures* $\mathcal{F} = \{F_y;\ y \in Y\}$*. We define a prior probability model for* $\mathcal{F}$ *using a dependent Dirichlet process (DDP) prior. Specifically, we use common weights for* F_y *and define the point masses as a function of* y*. We refer to the model as DDP-AR(1). We illustrate the model and posterior computation using Old Faithful Geyser dataset.*

2.1 Introduction

We present a nonparametric extension of autoregressive models in Bayesian analysis. Let $(Y_t, t \geq 1)$ denote a time series of random variables. The standard autoregressive model with one lag, AR(1), for a stationary process assumes $Y_t \mid Y_{t-1} = y \sim N(\alpha y; \sigma^2)$. Instead, we introduce a flexible distribution F_y for $Y_t \mid Y_{t-1} = y$. We define a nonparametric prior probability measure on the class $\mathcal{F} = \{F_y, y \in \mathcal{Y}\}$ in such a way that, marginally, F_y is a mixture of Gaussian distributions: $F_y \mid G_y \sim$

$\int N(0,\sigma^2)dG_y(\theta)$. We assume a prior on $\{G_y, y \in Y\}$ such that the random measure G_y follows a Dirichlet process (DP) prior, $DP(M, G_y^0)$ where M is the total mass parameter and G_y^0 the base measure (Ferguson, 1973). We introduce dependence across y, i.e. dependence for $\mathcal{F} = \{F_y, y \in \mathcal{Y}\}$, using the dependent Dirichlet Process (DDP) as defined by MacEachern (1999, 2001). The random measures G_y are almost surely discrete with the point masses generated marginally from the base measure G_y^0. The DDP introduces dependence across G_y by imposing dependence in the distribution of these point masses. We use the DDP to define the desidered nonparametric autoregressive model by assuming AR models for these point masses. We use this DDP structure to develop an AR like probability model over an array of random distributions. The DDP model provides a convenient starting point for the discussion. De Iorio *et al.* (2004) define an ANOVA DDP (as we described in the previous chapter). They propose a model which describes dependence across random distributions in an ANOVA fashion type. The ANOVA DDP model requires discrete variables, but this limits the applicability of the model when we wish to include continous variables for time series analysis. An extention in these lines is given by De Iorio *et al.*(2009), who introduce a linear DDP for survival analysis. The proposed nonparametric AR model can be seen as a special case of the ANOVA when the linear model for each point mass is an autoregression on y_{t-1}. Our model is based on the Dirichlet process prior distribution (Ferguson, 1973; Antoniak, 1974). The DP is a probability model for random probability distributions. It plays a central role in Bayesian nonparametric inference and it has been successfully applied in many problems. Sethuraman (1994) gives a constructive representation of the DP, showing that, if $G_y \sim DP(M, G_y^0)$, then it can be a. s. represented as $G_y = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h(y)}$. Here ω_h for $h \geq 1$ are common weights of the point masses at locations $\theta_h(y)$, it means that the weights are costant to respect the point masses. The weights are generated from rescaled Beta distributions, $\omega_h = V_h \prod_{j=1}^{h-1}(1 - V_j) \sim Be(1, M)$, and the locations θ_h are i.i.d. samples from the base measure G_y^0. We will use a single-p DDP structure (MacEachern, 1999; 2001) for our proposal, while in the fourth Chapter we will use variable weights abandoning the stick breaking representation. Rodriguez and ter Horst (2008) use the single-p DDP for time series analysis. They suppose discrete time, and model the trajectories for the point masses, $\theta_h(y)$ as dynamic linear models. Caron *et al.* (2006) consider a time-dependent version of DDPs. Griffin and Steel (2006) introduce dependence across random permutations on the athoms. They propose variable weights and a stationary process for the point masses; in particular, they assume that the V_js depend over time. Specifically, they define a sequence of times $\tau_1, \tau_2, \ldots$ as a Poisson process. The size of $V(t)$ increases at time τ_j by introducing an extra variable $0 < V_j^* < 1$. This process defines distributions that change in continous time. Griffin and Steel (2011) extended these results obtaining the ANOVA DDP as special case. Other approaches that explicitly introduce covariate dependence in the weights include kernal-stick breaking of Dunson and Park (2008), and the probit-stick breaking of Chung and Dunson (2011). Additional references in Hjort *et al.* (2010). An early development of dependent Dirichlet process, as we discussed in the first Chapter, is in Cifarelli and Regazzini (1978), where the dependence on the covariates is introduced as a regression in the base measure of marginally Dirichlet process distributed random probability measures. Cruz-Marcelo *et al.* (2010) review and compare some covariate-dependent models. For an approach via parametric mixtures of autoregressive models with common unknown lag, see Wood *et al.* (2011).

This chapter is organized as follows. In Section 2.2, we describe our proposal in more details. In Subsection , we develop posterior distributions for a simple version of our model (Subsection 2.2.3, when the variance for the observations is unknown); then we include a prior for the variance of $Y_t \mid Y_{t-1}$, and we calculate the conditional posterior distributions of interest. In the last Subsection (2.2.4), we compute the predictive distribution for our basic model and evaluate the relevance of the introduction of a proper prior for the variance. In Section 2.3, we illustrate the model using the Old Faithful Geyser dataset (available in R software). We study possible extensions of the basic DDP-AR(1) model: the efficiency of the parameters for the trajectories for the point masses (Subsection 2.4.1). In Subsection 2.4.3, we consider a univariate mixture for the trajectories for the point masses, which is also an introduction for the next chapter. In Subsection 2.4.2, we extend the model by allowing quadratic trajectories. In Section 2.5, we extend our model to the multivariate case. We analyze different aspects linked to the covariance matrix. In Section 2.6, we discuss about possible further extentions of our model.

2.2 DDP-AR(1)

We define an autoregressive model with lag one, AR(1), such that $Y_t \mid Y_{t-1} = y, F_y \sim F_y(\cdot)$. The model rises up the dependent Dirichlet process model, by defining a dependent prior on the class $\mathcal{F} = \{F_y;\ y \in \mathcal{Y}\}$ using a DDP prior with common weights. Let $(G_y, y \in \mathcal{Y})$ be a family of random probability measures in the space $\mathcal{Y}$, where

$$G_y = \sum_{h=1}^{\infty} w_h \delta_{\theta_h(y)}$$

is a Dirichlet process. We use a DDP prior for $(G_y, y \in \mathcal{Y})$, such that the point masses $\theta_h(y)$ are trajectories indexed by y. As a starting point, we use the simplest possible linear trajectories

$$\theta_h(y) = \beta_h + \alpha_h y \ \text{ with } \ (\beta_h, \alpha_h) \overset{iid}{\sim} G^0 \tag{2.1}$$

with $\beta_h \sim N(m_\beta, \sigma^2_\beta)$ and $\alpha_h \sim N(m_\alpha, \sigma^2_\alpha)$. This implicitely defines $G^0(\theta_h(y),\ y \in \Re)$ with marginal $G^0(\theta_h(y)) = N(m_\beta + y m_\alpha, \sigma^2_\beta + y^2 \sigma^2_\alpha)$. In equation (2.1), we define trajectories across the point masses which describe functional dependence. Clearly, for the discrete nature, a Dirichlet process cannot be used directly as a prior for the unknown distribution of continous data, (Ferguson, 1983; Lo, 1984). We slightly extend the basic model by including an additional residual in the likelihood. The proposed model is for $t \geq 2$

$$\begin{aligned} Y_t \mid Y_{t-1} = y, G_y, \sigma^2 &\sim f_y(Y_t \mid Y_{t-1} = y, G_y, \sigma^2) = \int N(Y_t; \mid \mu_t, \sigma^2)\, dG_y(\mu_t) \\ G_y &\sim DP(M, G^0_y) \end{aligned} \quad (2.2)$$

In this first simple model, σ^2 is known; in the next subsection, we extend to the case of unknown variance, σ^2. In equation (2.2) there is the core of our methodologiacal contribution for this chapter and in some way this is the base for further developments of the next chapters. The contribution is the use of a single-p DDP prior in a DP mixture formalization for time series analysis. This framework has structural differences with i. e. the mixture of products of DP for Bayesian linear models proposed by Muliere and Scarsini (1982), (that we reviewed in the first Chapter). Firstly, we consider in equation (2.2) a mixture of parametric distributions instead of a product of Dirichlet processes based on the partial exchangeability assumption. Secondly, the mixing distribution is, marginally, a DP, instead of a mixture of DPs and we assume exchangeability on the data.

Model (2.2) can be equivalently formulated as a hierarchical model:

$$Y_t \mid Y_{t-1} = y, \mu_t \sim N(\mu_t(y), \sigma^2)$$

$$\mu_t(y) \mid G_y \sim G_y$$

$$(G_y, y \in Y) \sim DDP(M, G^0_y) \quad (2.3)$$

where the DDP prior is defined as before. We obtain a flexible model for the conditional density of $Y_t \mid Y_{t-1}$, given by the mixture of Gaussians (2.2), which resembles traditional kernel density estimation, as in Lo (1984), who introduced the convolution for continous kernel functions. In equation (2.3) we have replaced the mixture $\int N(y_t \mid \mu_t, \sigma^2)\ dF_y(\mu_t)$ by a hierarchical model with a (new) latent process $(\mu_t, t \geq 1)$. For the implementation of posterior simulations we find it is convenient to use an equivalent parametrization using the unique point masses θ_h and latent indicators $r_t = h$ if $\mu_t = \theta_h(y)$, following that

$$\begin{aligned} Y_t \mid Y_{t-1} = y, r_t = h, \sigma^2, \{(\beta_h, \alpha_h)\} &\sim N(\beta_h + \alpha_h y; \sigma^2) \\ p(r_t = h) = \omega_h \ \ (\alpha_h, \beta_h) &\overset{iid}{\sim} G^0 \text{ for } h = 1, 2, \ldots, H. \end{aligned} \tag{2.4}$$

For computational simplicity we introduce a further approximation using a finite mixture

$$G_y(Y_t) = \sum_{h=1}^{H} w_h \delta_{\theta_h(y)}. \tag{2.5}$$

We refer to equation (2.5) as the finite stick breaking approximation of G_y, denoted by DP$_H$. In particular, we have sticks, ω_h, such that $\omega_h = V_h \prod_{j=1}^{h-1}(1 - V_j)$, for $h = 1, \ldots, H$ with prior $V_h \sim$ Be$(1, M)$, $h = 1, \ldots, H - 1$ and $V_H = 1$, $\sum_{h=1}^{H} \omega_h = 1$ and marginally $G_y \sim$ DP$_H(M, G_y^0)$. According to equation (2.3), we denote the joint model as $(G_y, y \in Y) \sim$ DDP$_H(M, G_y^0)$. This is also the evolution of the hierírarchical structure of the equation (2.3). In summary, we have

$$\begin{aligned} Y_t \mid Y_{t-1} = y, r_t = h &\sim N(\beta_h + \alpha_h y, \sigma^2) \\ p(r_t = h) &= \omega_h \\ (\beta_h, \alpha_h) &\overset{i.i.d.}{\sim} G^0(\beta, \alpha) \end{aligned} \tag{2.6}$$

As said here, we suppose that the variance of $Y_t \mid Y_{t-1}$ is fixed. The i.i.d. prior on (β_h, α_h) and the stick-breaking prior for w_h together define the DP prior, in this case truncated as DP$_H$. See Figure 2.1. Note that we propose modeling a sequence of continous outcomes by means of a countable mixture of regressions on lagged terms. This approach differs from previous models for time series (such as Rodriguez and ter Horst, 2008, or Caron *et al.*, or Contreras-Cristan *et al.*). We fix the number of mixtures and it is well distinguished also from random functionals approximating distributions as in Muliere and Tardella (1998).

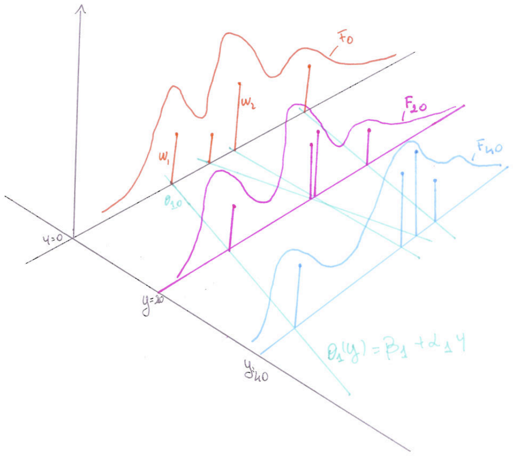

Figure 2.1: Stylized representation of the idea beyond the model. Red, pink and blue colors are for the weights of the point masses. On the top of the lines we have different point masses. We describe red, pink and blu curves with three different possible Dirichlet processes. The trajectories for the point masses are the links between the probability distributions and represent autoregressive modeling dependence across three Dirichlet processes.

2.2.1 Posterior Distributions

Having described our simplest model, we have all the elements to calculate the posterior distributions. We refer to the hyperparameter space and we illustrate the joint posterior distribution. References on MCMC algorithm for DP mixtures are Escobar and West (1995); MacEachern (1994), MacEachern and Bush (1996). MacEachern and Müller (1996) for the efficiency of a MCMC and MacEachern and Muller (1998) for the use of the Gibbs sampler algorithm in the case of non conjugate priors.

2.2.2 Bayesian Inference

Let ξ_1 be the vector of hyperparameters for the model (2.4), given by

$$\xi_1 = (r_1, r_2, .., r_T, V_1, V_2, ..., V_{H-1}, \alpha_1, \alpha_2, \ldots, \alpha_{H-1}, \beta_1, \beta_2, \ldots, \beta_{H-1}). \tag{2.7}$$

It is easy to compute the joint probability distribution

$$\begin{aligned} P(y, \xi_1) = \prod_{t=1}^{T} p(Y_t \mid r_t, \alpha_h, \beta_h) \prod_{t=1}^{T} p(r_t \mid \xi \setminus \{r_t\}) \\ \prod_{h=1}^{H} p(V_h \mid \xi \setminus \{V_h\}) p(\alpha_h \mid \xi \setminus \{\alpha_h\}) p(\beta_h \mid \xi \setminus \{\beta_h\}) = \\ = \prod_{t=2}^{T} N(Y_t \mid \beta_{r_t} + \alpha_{r_t} Y_{t-1}, \sigma^2) \prod_{t=1}^{T} \omega_{r_t} \\ \prod_{h=1}^{H} Be(V_h \mid 1, M) N(\alpha_h \mid m_\alpha, {\sigma_\alpha}^2) N(\beta_h \mid m_\beta, {\sigma_\beta}^2). \end{aligned} \tag{2.8}$$

We approximate posterior inference using Markov Chain Monte Carlo simulation. We provide the full conditional posterior distributions required for the Gibbs sampler, which are easily computed from (2.8). We define transition probabilities by generalizing from each of the complete conditional posterior distributions. In particular, let $Q_h = \{t : \ r_t = h\}$ denote the subset of observations for $h = 1, 2, \ldots, H$ with μ_t tied to $\theta_h(Y_{t-1})$. We find the following conditional posterior distributions:

$\alpha_h :$ for $h = 1, \ldots, H$. Let y_t^* be equal $(Y_t - \beta_h)$ for $t = 1, 2, \ldots, T$. Then

$$p(\alpha_h \mid \xi \setminus \{\alpha_h\}) \propto N(\alpha_h \mid m_\alpha, \sigma_\alpha^2) \prod_{t \in Q_h} N(y_t^* \mid \alpha_h Y_{t-1}, \sigma^2).$$

Therefore, $p(\alpha_h \mid \xi \setminus \{\alpha_h\}) \propto N(\alpha_h \mid m_1, V_1)$, with

$$m_1 = V_1(\sigma_\alpha^{-2} m_\alpha + \sigma^{-2} \sum_{t \in Q_h} Y_{t-1} y_t^*) \text{ and } V_1^{-1} = (\sigma_\alpha^{-2} + \sigma^{-2} \sum_{t \in Q_h} Y_{t-1}^2)$$

where $n_h =\mid Q_h \mid$ is the number of observations tied to $\theta_h(\cdot)$.

β_h**:** for $h = 1, \ldots, H$. Suppose $y_t^{**} = (Y_t - \alpha_h Y_{t-1})$ for $t = 1, 2, \ldots, T$, we have

$$p(\beta_h \mid \xi \setminus \{\beta_h\}) \propto N(\beta_h \mid m_\beta, \sigma_\beta^2) \prod_{t \in Q_h} N(y_t^{**} \mid \beta_h, \sigma^2);$$

thus, $p(\beta_h \mid r_t, V_h, \alpha_h) \propto N(\beta_h \mid m_2, V_2)$ with

$$m_2 = V_2(\sigma_\beta^{-2} m_\beta + n_h \sigma^{-2} \frac{1}{n_h} \sum_{Q_h} y_t^{**}) \text{ and } V_2^{-1} = (\sigma_\beta^{-2} + n_h \sigma^{-2})$$

V_h**:** for $h = 1, \ldots, H-1$. Let $S_h = \{t : r_t > h\}$ and, as before, $Q_h = \{t : r_t = h\}$ be the subsets for the stick breaking construction. From the joint posterior distribution, we see that

$$p(V_h \mid \xi \setminus \{V_h\}) \propto Be(1, M) \prod_{t \in S_h} (1 - V_h) \prod_{t \in Q_h} V_h = Be(1+ \mid Q_h \mid, M+ \mid S_h \mid)$$

.

r_t**:** for $t = 1, \ldots, T$. The latent indicators r_t for the mixture component are discrete random variables with

$$p(r_t \mid \xi \setminus \{r_t\}) \propto \omega_h^*$$

where $\omega_h^* = \omega_h N(Y_t \mid \alpha_h Y_{t-1} + \beta_h, \sigma^2)$ and $r_t \in \{1...H\}$.

The Gibbs sampler approximation is illustrated in the next section with an application to the Old Faithful Gayser dataset.

2.2.3 Model with unknown observations variance

Computations in the previous section assume a known value for the variance σ^2. We remove this restriction, and describe the resulting inference. We assume an Inverse Gamma prior distribution for σ^2, $\sigma^{-2} \sim Ga(a, b)$. The parameter space is

$$\xi_2 = (r_1, r_2, \ldots, r_T, V_1, V_2, \ldots, V_{H-1}, \alpha_1, \alpha_2, \ldots, \alpha_{H-1}, \beta_1, \beta_2, \ldots, \beta_{H-1}, \sigma^2), \quad (2.9)$$

and the previous equation (2.8) becames:

$$p(y, \xi_2) = \prod_{t=1}^{T} p(Y_t \mid r_t, \alpha, \beta, \sigma^2) \prod_{t=1}^{T} p(r_t \mid \xi \setminus \{r_t\}) \prod_{h=1}^{H} p(V_h) p(\alpha_h) p(\beta_h) p(\sigma^{-2}) =$$

$$= p(\sigma^{-2}) \prod_{t=1}^{T} N(Y_t \mid \beta_{r_t} + \alpha_{r_t} Y_{t-1}, \sigma^2) \prod_{t=1}^{T} \omega_{r_t} \prod_{h=1}^{H} Be(V_h \mid 1, M) N(\alpha_h \mid m_\alpha, {\sigma_\alpha}^2)$$

$$N(\beta_h \mid m_\beta, {\sigma_\beta}^2). \quad (2.10)$$

The full conditional for σ^2 is Inverse Gamma distributed, $IG(A, B)$, where the shape and rate are: $A = (a + \frac{1}{2}T)$ and $B = (b + \frac{1}{2}\sum_{t=1}^{T}(Y_t - \beta_{r_t} - \alpha_{r_t} Y_{t-1})^2)$, respectively.

2.2.4 Predictive Distribution

In the previous subsection, we obtained the full conditional distributions that are used in a Gibbs sampler from the posterior distribution of the parameters. Therefore, we can compute the posterior predictive distribution, using a Monte Carlo simulation. The total number of iterations is I and the estimator of the predictive distribution for our model equal to

$$p(Y_{T+1} \mid y) = \frac{1}{I} \sum_{i=1}^{I} [\sum_{h=1}^{H} w_h^{(i)} N(\cdot \mid \alpha_h^{(i)} y + \beta_h^{(i)}, \sigma^2)]$$

This probability represents also a random probability function and if we plug in a specific point it will be the average for the random posterior distributions and not the average of posterior distributions and we have

$$\frac{1}{I}\sum_{i=1}^{I}[\sum_{h=1}^{H} w_h^{(i)} N(\cdot \mid \alpha_h^{(i)} y + \beta_h^{(i)}, \sigma^2)] = E[f_y \mid y] = \bar{f} \tag{2.11}$$

In a Markov Chain Monte Carlo the chain gradually forgets the initial state and it eventually converges (we have to check the convergence) to a unique stationary distribution. Note that in equation (2.11) above we did not introduce yet the burn in. This is the number of the initial iterations will be discarded. So the previous equation will be rewritten as

$$\frac{1}{I-R}\sum_{i=R+1}^{I}[\sum_{h=1}^{H} w_h^{(i)} N(\cdot \mid \alpha_h^{(i)} y + \beta_h^{(i)}, \sigma^2)] = E[f_y \mid y] = \bar{f} \tag{2.12}$$

where R is the burn in. Equation (2.12) is the *ergodic average* for our model. One of the problems discussed in the literature is about the definition of the number of iterations that we have to discard. Geyer (1992) suggests to take out from the total number of iterations 1%, 2% of the initial iterations. We will explain more details about our choice in the next section, when we will illustrate our Bayesian predictive distributions applied to Old Faithful Gayser dataset. Note that $F_y(\cdot)$ is the posterior mean of the autoregressive model for y corresponding to a first-lag response. Let $\bar{F}_y = E(F_y \mid data)$ denote the posterior expectation and $\bar{f}_y(\cdot)$ be the corresponding probability density function. The posterior mean $\bar{F}_y$ is easiest evaluated as posterior predictive distribution as well as we illustrate in the application of the next sections.

2.3 Example: Old Faithful Geyser Dataset

We illustrate posterior inference for the Old Faithful Geyser data, a popular dataset for kernel density estimate and time series analysis. The dataset describes eruptions of the Old Faithful, which is a geyser in Yellowstone National Park, Wyoming (USA). The Old Faithful Geyser dataset includes three variables: the date of the gathering of statistical data, the duration of the current eruption expressed in minutes and the time between two succeding eruptions also expressed in minutes. During October 1980, the data were collected by volonteers and provided by R. Hutchinson *et al.*. The park services looked for a prediction of the next eruption of this geyser, which is an attraction for tourists.

Weisberg (1980) employed this dataset for density estimation in a linear regression model. Silverman (1986) selected 107 observations of the eruptions to illustrate density estimation. Azzalini and Bowman (1990) conducted several analysis on this dataset using 299 observations and adding one more variable for different times of the day (morning, noon and evening). They illustrated several examples about kernel density estimation and time series analysis, focusing on the asymmetric relation of the duration of the eruptions and the time between eruptions.

We consider 272 observations and two variables: the duration of the eruption and the waiting time among two succeding eruptions. This is the version of the dataset given by Härdle (1991), available in R software. In addition, for the first following applications we consider only the waiting time among succeding eruptions, y_t.

The waiting time variable, y_{t-1},for $t = 2, \dots, 272$ is the lagged response, which is the covariate. The number of observations for y_{t-1} is 271 and we add as first observation the mean of y_t . In Figure 2.2, we show a scatter plot labelling x-axes as y_{t-1} and y-axes y_t, respectively. We distinguish three clusters, which are three possible subsets to respect x-axes. These three intervals for y_t are $[45, 55]$, $[60, 70]$, $[75, 85]$. Looking at this plot to respect y-axes, we note a different form of clouds of points.

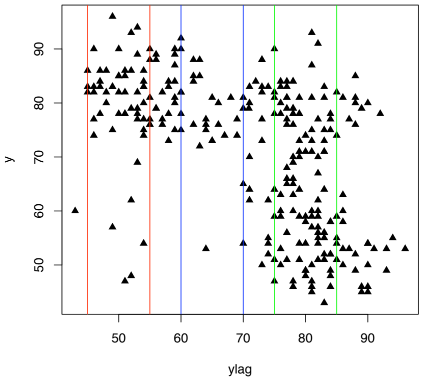

Figure 2.2: Old Faithful gayser data. We study Bayesian density estimations for three distinguished subsets. Red lines describe $y_{t-1} \in [45; 55]$; blue lines are for $y_{t-1} \in [60; 70]$ and green lines define $y_{t-1} \in [75; 85]$.

In particular, for the third subset in $[75, 85]$, we have a bimodal cloud of points. We have enough analitical elements to apply our models to this dataset. Specifically we compute a blocked Gibbs sampler algorithm for the variable waiting time, Y_t for whole observations. We graphically compare posterior density distributions computed on the three subsets for the first model that we illustrated in the Subsection 2.2.2 and the posterior distributions for the second model in the Subsection 2.2.3. We recall the two models using the vector of hyperparameters. For the first model we consider the parameters of the equation (2.7) and for the second model we refer to equation (2.9). We focus on the relevance of the variance σ^2. We evaluate if the variance is useful for better fitting of the predictive distribution. We use a diagnostic proposed by Geweke (1992) to decide termination of the simulates Markov chain. Geyser (1992) proposed a burn in equal to 1 or 2% of runs to guarantee that the chain reaches the stationary. We fixed 200,000 iterations, burn in 100,000 and thinning 20, in accordance with Geweke, 1992. Therefore, the inspection of the trace plots revealed slowly mixing, which is the dependence over the total number of iterations. We augmented the total number of iterations, 450,000, the burn in is equal to 200,000 and thinning 50 obtaining a stationary chain and reaching the convergence. For Bayesian predictive posterior distribution, we refer to the previous equation (2.12) (which is in the previous Section 2.2.4). Specifically, for the total number of iterations $I = 450,000$, burn in equal to $R = 200,000$ and thinning equal to 20, we used 5,000 samples of the MCMC output. However, we fixed the number of sticks, $H = 20$ and the evaluation is on the intermediate point of each intervals $[45, 55]$ $[60, 70]$ $[75, 85]$ and these are $y = 50, 65, 80$, respectively

$$\sum_{i=200{,}001}^{450{,}000} [\sum_{h=1}^{20} w_h^{(i)} N(. \mid \alpha_h^{(i)} y + \beta_h^{(i)}, \sigma^2)]$$

fixed variance σ^2 is equal to 25 for the first hyperparameter space and random variance is Gamma distributed with location parameter equal to 2 and rate parameter equal to 2, $\sigma^2 \sim Inv - Ga(2, 2)$, i.e. $E[\sigma^2] = 2$, $E[1/\sigma^2] = 1$ and $Var[1/\sigma^2] = 0.5$, which implies higher and more dispersed precision than in the case of fixed variance $1/25 = 0.04$. In Figure 2.3, we illustrate a comparison between model 2.8 and model 2.10. We assume that the variance σ^2 is equal to 25 and we plot it using kernel density estimate (otherwise this is a number and not a distribution) for the unknown variance we suppose the Inverse-Gamma(2,10)-distributed prior distribution. We notice that there is not significant difference. Both of the distributions are centered on the value 25.

One more detail is for the fixed thinning, we selected the values of α_h, β_h, and ω_h every 20 iterations, such that we have in the new step approximatively the same predictive posterior distribution. In Figure 2.4, we compare posterior distributions for two possible initial values for $\sigma^2 \sim IG(2, 2)$ and $\sigma^2 \sim IG(2, 10)$. We observe that there is not a substantial difference and we can conclude that we can choose the first model which is in equation (2.7).

In Figure 2.5, we illustrate Bayesian estimated posterior predictive distributions for the values of $y = 50, 65, 80$ on a grid in $[40, 100]$ similar to the real range,

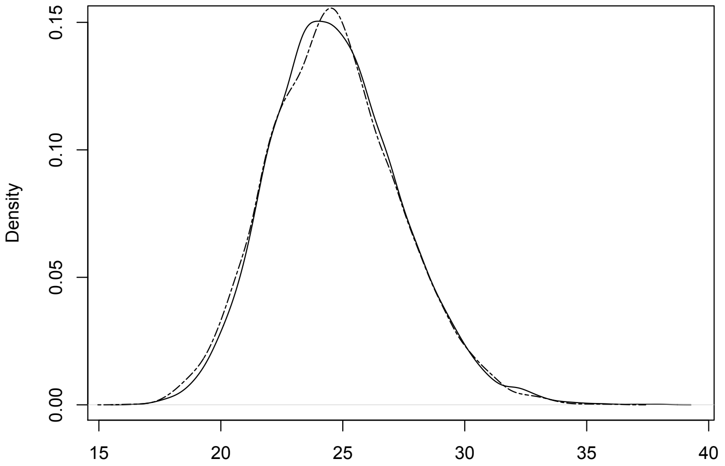

Figure 2.3: Comparison between fixed variance $\sigma^2 = 25$ using a kernel density estimation (dashed line) and unknown variance assuming prior distribution: $\sigma^2 \sim IG(2, 10)$ (solid line).

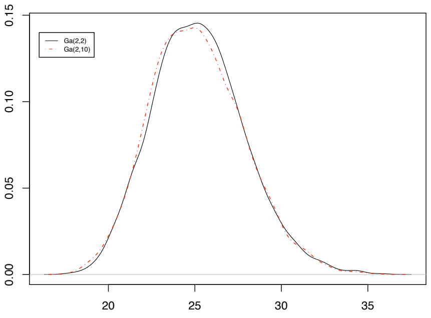

Figure 2.4: Posterior distribution of the variance σ^2 of the normal components in the nonparametric mixture, under an Inverse-Gamma(2,2), (black continous) and an Inverse-Gamma(2,10) distributions (red dot-dashed).

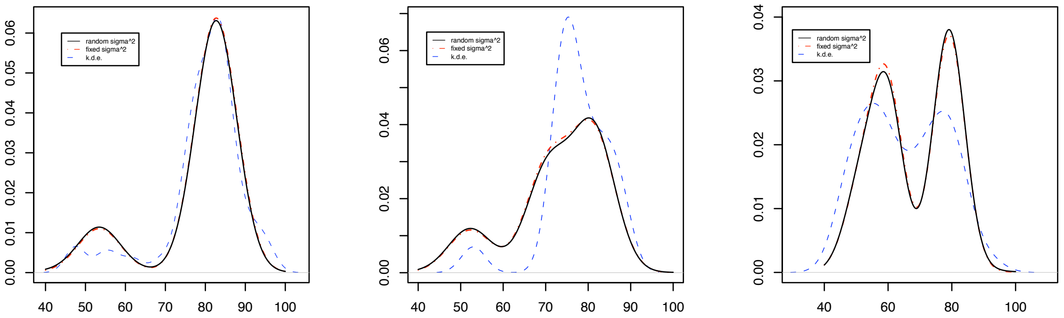

Figure 2.5: Bayesian density estimation, $E(F_y(\cdot) \mid y)$, for few selected values of $y = 50, 65, 80$.

$[43, 96]$. We checked convergences and choosed the initial values using diagnostics in CODA package for R software. For screaning and final decision of the initial values we use Geweke's diagnostics. Note that all these hyperparameters values were fixed starting from sample values when available (the sample mean and variance of the data are 70.9 and 184.8, respectively), but a fairly massive robustness analysis and MCMC performance were also conducted. For instance, we found that when increasing the fixed Gaussian component variance to 50 or 100, the converge diagnostics were worse. On the other hand, we assumed a higher prior mean for σ^2, i.e. Inverse-Gamma(2,10)-distributed and we found moderate differences on the inferences; for instance, see Figure 2.4, where the corresponding distributions of σ^2 are depicted.

In Figure 2.5, we illustrate a comparison of Bayesian predictive density estimations. We observe that there is not a substantial difference if the variance of the likelihood is fixed or random, in particular, when density estimation is $y = 80$: bimodal curves are still present, as in the kernel density estimation.

Graphical comparison of predictive distributions is not statistically enough. We calculate also the mean square error and the minimum mean square error for numerical comparison. We evaluate which is the best of the two proposed models for the three subsets. We propose the following table with the mean square errors, MSE, and the minimum mean square errors, MMSE, for the first model that we defined in equation (2.8) and for the second model in (2.10). We compare the predictive distribution function illustrated in equation (2.11), imputed for the two hyperparameter spaces. We indicate one step ahead distribution for the first model as $\bar{f}_y(y)$ and $\bar{f}^0_y(y)$ for the second proposed model. We apply the following formula for the mean square error: $E(\int[\bar{f}_y(y) - \bar{f}^0_y(y)]^2 dy$, where the subindex $_y$ assumes values 50, 65 or 80, respectively. For the minimum mean square error estimator, we compute the variance over the integral: $Var(\int[\bar{f}_y(y) - \bar{f}^0_y(y)]^2 dy$.

Model	**y**	**MSE**	**MMSE**
2.8	50	0.01747147	0.0003798467
2.10	50	0.01740426	0.0003625283
2.8	65	0.01899085	0.0001826035
2.10	65	0.01899063	0.0001709452
2.8	80	0.01909110	8.773023e-05
2.10	80	0.01924572	9.8905e-05

Table 2.1: Comparison between two models - the first model has fixed variance, $\sigma^2 = 25$ and the second model σ^2 is Inverse Gamma(2,10) distibuted.

In Table 2.1, the two models are quite similar. The minimum mean square error suggests the second model, so the model with random variance σ^2 for $Y_t \mid Y_{t-1}$. We compare the two models for $y = 50, 65, 80$ on the two subsets and we introduce one more prior on the first model which includes more information. In this specific case, the choice of the fixed or random variance plays a marginal role due to the framework of the simply initial model, which is able to fit the data, and also for the differences showed in Figure 2.3 and Figure 2.4. Indeed, for a huge dataset the role of the variance is crucial as well as the DNA-sequencing dataset that we discuss in the next chapter.

In the previous methodological sections, we assumed a finite number of dependent Dirichlet process priors which is related to the total mass parameter. In Bayesian

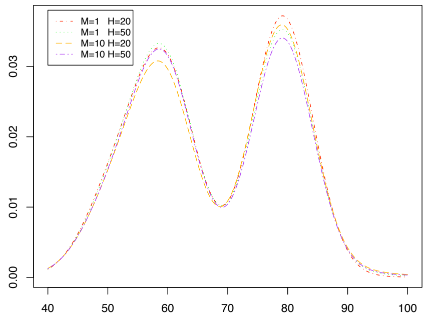

Figure 2.6: Bayesian posterior density $E(F_y(\cdot) \mid y = 80)$ for 4 possible combinations of the initial values for the total mass parameter, M and the number of sticks, H.

density estimations, we obeserved that the predictive density for y equal to 80 is bimodal. This result has been useful for the final decision of the initial number of sticks and for the trade off between the finite number of sticks (of the stick-breaking representation) and the total mass parameter of DDP-prior model. For the choice of the finite number of sticks, we analyze different possible combinations of initial values of the parameters. In Figure 2.6, we show Bayesian density estimate, $E(F_y(\cdot) \mid y)$ fixing y equal to 80 and our best solutions considering the number of sticks H equal to 1 or 10 and the total mass parameter M equal to 20 or 50, respectively. We observe that all these combinations are able to describe the bimodality of the Bayesian density and these differ from each others for a small distance. For this reason we plug in specific labels.

In Figure 2.7, we illustrate the posterior mean of $f_{y_{t-1}}(\cdot)$ for $y_{t-1} = 80$, choosed all the initial values for the parameters under the $1/\sigma^2 \sim InverseGa(2, 2)$ prior, $M = 1$ and $H = 50$, together with 95% pointwise posterior credible bands.

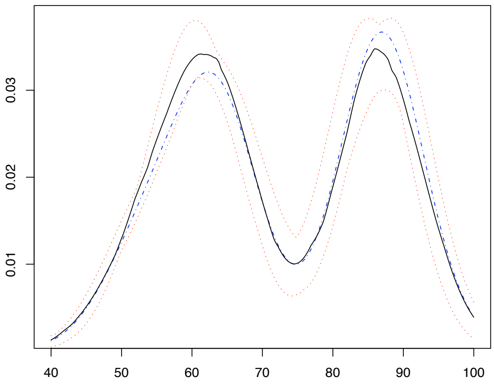

Figure 2.7: Bayesian estimates of $\bar{f}_{y_{t-1}}(\cdot)$ for $y_{t-1} = 80$ (blue semidashed line), together with point wise 95% credible bands (red dotted lines) and median (solid black line).

2.3.1 Co-Clustering

In this section we underline some aspects of the estimation of the posterior distributions using the Gibbs sampler algorithm. We focus on the probability of the latent variable r_t when this is equal to one of the sticks, h, that technically, we defined $P(r_t = h) = \omega_h$. This is a simplified formalization of the $P(r_t = h) = E(\omega_h)$ and $P(r_t = h \mid \omega_h) = \omega_h$. The estimation of the posterior distribution of r_t can be obtained by a functional, $g_1(r_t)$

$$g_1(r_t) = \frac{1}{I}\sum_{i=R+1}^{I} g(r_t^{(i)})$$

where R is the burn in and I is the total number of iterations [1]. This is a no-Rao Blackwellised estimator (Gelfand and Smith, 1990; Bush and MacEachern, 1994). We count the number of times that $r_t = r_{t'}$ where $t \neq t'$ for $t = 1, 2, \dots T$ and the

[1] in the literature, the number of iterations is usually indicated by M, but it could be confused with the total mass parameter of the DP

associated probability is $P(r_t = r_{t'})$. Rao-Blackwellised estimator is

$$g_2(r_t) = \frac{1}{I} \sum_{i=R+1}^{R+I} E(g(r_t^{(i)} \mid \xi^{(i)} \setminus \{r_t\}))$$

In this specific case, these two different estimators are equivalent, we have:

$$E(g(r_t^{(i)} \mid \xi^{(i)} \setminus \{r_t^{(i)}\})) = E(\omega_h).$$

We use the first estimator, $g_1(r_t)$ to tabulate $P(r_t = r_{t'})$. This implies that we also analyze the relevance of the posterior distribution of the latent variable r_t given all the other parameters on the subset Q_h.
MacEachern *et al.* (1999) discussed about efficiency of Gibbs sampler algorithm, which is very important if the model is to be useful in pratice. Dahl (2006) proposed a least-squares model-based clustering for gene expression data using DP mixtures model. The advantage of this method is the selection of a sum clustering which consider the pairwise probability matrix. We applied this method to our posterior distribution for the latent variable r_t and computed our natural number of clusters obtained via DP mixtures model. We chose the number of mixtures, cheking the convergences of the parameters (that we discussed in the previous subsection), here we study the similarities of the estimated mixtures. In Figure 2.8 each square is a random partition obtained by the closest points. Each point is a probability. The yellow squares indicate the set of points such that $p(r_t = r_{t'}) \leq 0.30$. The green-yellow squares exihibit $0.30 < p(r_t = r_{t'}) < 0.60$; green squares display $p(r_t = r_{t'}) \geq 0.60$. For the conjugacy of the DP mixture model, the pairwise probability matrix of the latent variables is the posterior predictive distribution for a new vector $r_{t'}$ evaluated at r_t.

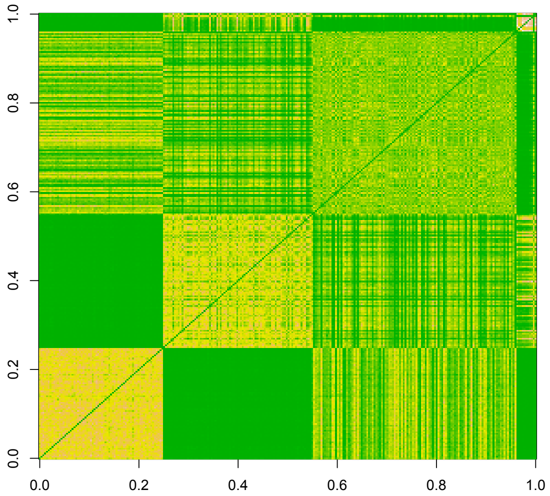

Figure 2.8: Co-clustering plot on the posterior distribution of the latent variable r_t. Each point in the squares represent $p(r_t = r_{t'} \mid \xi)$ and a point estimate of the true clustering is based on squared distances for the pairwise probability matrix.

2.4 Extensions of the base DDP-AR(1) model

In the previous Section 2.2, we proposed a DDP prior for a collection of random probability measures, indexed by lagged covariates. In this section we discuss different possible variations of the base DDP-AR(1) model. We focus on different aspects of the trajectories: in Subsection 2.4.1 we consider the efficiency of the implemented MCMC algorithm, in Subsection 2.4.2 we extend linear trajectories to quadratic trajectories. In the last Subsection 2.4.3 we study univariate mixtures of trajectories, these represent methodological base for the construction of the model in the next Chapter 3.

2.4.1 Efficient Trajectories

The trajectories for the point masses in equation (2.4) are two indipendent prior parameters normal distributed. However, we discuss here a more efficient construction. For parsimony and efficiency of the model and better mixing in the Gibbs sampler algorithm, we define a bivariate parameter $\boldsymbol{\theta}_h$ such that it will describe jointly the two parameters of the trajectories, α_h and β_h, respectively. This method was introduced by MacEachern and M$\ddot{u}$ller (2000) for an efficient MCMC schemes. Let α_h and β_h be two different parameters for the trajectories for the point masses as we described in equation (2.8). We study jointly these two parameters as: $\boldsymbol{\theta}_h = (\alpha_h, \beta_h)$. The model is:

$\boldsymbol{\theta}_h \sim N_2(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ where the vector of the means is $\boldsymbol{\mu}_{\boldsymbol{\theta}} = (m_\alpha, m_\beta)$ and the covariance matrix is $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = diag(\sigma^2_\alpha, \sigma^2_\beta)$. The hyperparameter space is $\xi_2 = (\boldsymbol{\theta}_h, V_h, r_t)$ and the full conditionals change only on the new parameter, $\boldsymbol{\theta}_h$, as

$$p(\boldsymbol{\theta}_h \mid r_t, V_h) \propto N_2(\boldsymbol{\theta}_h \mid \mathbf{m}_3, \mathbf{V}_3)$$

with posterior location $\mathbf{m}_3 = \mathbf{V}_3(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \mathbf{X}'_t y^*_t/\sigma^2)$ and posterior variance equal to $\mathbf{V}^{-1}_3 = (\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}} + \mathbf{X}'_t\mathbf{X}_t/\sigma^2)$.

In addition, we indicate design matrix, $\mathbf{X}_t = (y_{t-1}, 1)$. As above, the dependent variable is y^*_t on the set $\{t \in Q_h\}$. Note that the outcome, y^*_t is univariate and its fixed variance, σ^2. We also introduce the common variance random. In this case we estimate the posterior distributions when there is one parameter for the trajectory and the common variance is random. The hyperparameter space is $\xi_4 = (\boldsymbol{\theta}_h, \sigma^2, V_h, r_t)$. Mean and variance of the posterior distributions are formally the same, we have just substituted in σ^2 the full conditional distribution because it is random. So we compare graphically the random probability distributions considering α_h and β_h separated and for $\boldsymbol{\theta}_h = (\alpha_h, \beta_h)$ jointly in both cases we assume that also the common variance is random. The inference on the random posterior distributions gives uni-

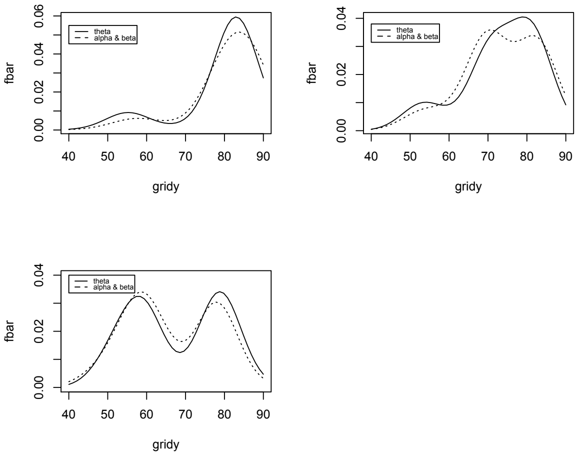

Figure 2.9: Comparison of Bayesian density estimation for α_h and β_h computed separately and $\boldsymbol{\theta}_h = (\alpha_h, \beta_h)$. We consider the model when σ^2 is fixed parameter.

variate distributions and the tails are heavier than the prior densities. In Figure 2.9, we illustrate how the parameters on the locations differ if they are studied jointly or separetly, $\boldsymbol{\theta}_h$ or α_h and β_h, respectively.

2.4.2 Linear and quadratic trajectories

Linear trajectories for point masses are the simplest possible case as we showed in Section 2.2.2 in equation (2.1). Extensions are clearly possible. Here, we introduce a quadratic form for the trajectories. The model becomes

$$\begin{array}{rcl} Y_t \mid Y_{t-1} = y, r_t = h & \sim & N(m_t, \sigma^2) \\ m_t & = & \beta_h + \alpha_h y + \gamma_h Y_{t-1}^2 \\ p(r_t = h) & = & w_h \\ (\beta_h, \alpha_h, \gamma_h) & \overset{iid}{\sim} & G^0(\beta, \alpha, \gamma) \end{array} \tag{2.13}$$

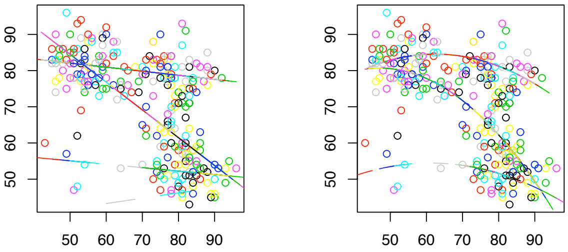

Figure 2.10: Last (5000-th) MCMC iteration of all the atoms θ_h, $h = 1, \dots, H$ in the linear (left) and the quadratic (right) case. The same color is used for points associated to the corresponding clusters. Linear and quadratic estimated trajectories for fixed variance $\sigma^2 = 25$. In each plot there are 20 different colors for 20 different values assumed by the latent variable r_t.

where $\beta_h \sim N(m_\beta, \sigma^2_\beta)$, $\alpha_h \sim N(m_\alpha, \sigma^2_\alpha)$, and $\gamma_h \sim (m_\gamma, \sigma^2_\gamma)$. We assume known mean and variances of $\beta_h, \alpha_h, \gamma_h$, respectively. Note that in this hypothesis G^0 is defined on $\Re^3$.

We apply the estimation of the trajectories for whole observations for the variable waiting time. We illustrate how our trajectories represent dependence among the athoms.

In Figure 2.10, we plot both the linear and quadratic trajectories. We count four curves and five lines. The number of lines and curves is the number of activated clusters. We suppose twenty clusters in the model, but the real number of estimated clusters is less. The number of clusters is naturally different for the different number of parameters involved in the models: more parameters as in the quadratic trajectories conducts less clusters; viceversa, less parameters as well as in the linear model, products more clusters and also more lines in the plot.

2.4.3 Mixture of trajectories

In this section we express the trajectories for the point masses as a mixture of different functions. There are two kind of motivations. We focus on more flexibility for the trajectories and secondly, this evolution of our basic model is the starting point for the next chapter. In particular, we add one more latent variable for the mixture of the trajectories $k_t = j$ for j = 1,2,..J. In this application we introduce mixtures of four components, the motivation will become more clear in the next chapter, when we will apply this model to the DNA-sequencing, where the component will represent the four labels of the nucleotides. The model is the following

$$\begin{aligned} Y_t \mid Y_{t-1} = y, r_t = h, k_t = j &\sim N(\theta_h^j, \sigma^2) \\ \theta_h^j &= \beta_h^j + \alpha_h^j y \\ p(r_t = h) &= w_h \quad p(k_t = j) = p_j \\ (\beta_h^j, \alpha_h^j) &\sim G_j^0(\beta_h^j, \alpha_h^j) \\ \text{or } p(Y_t \mid Y_{t-1} = y, r_t = h) = \sum_{j=1}^{J} p_j N(\theta_h^j, \sigma^2) \\ \text{or } p(Y_t \mid Y_{t-1} = y, k_t = j) = \sum_{h=1}^{H} w_h N(\theta_h^j, \sigma^2) \end{aligned}$$

The hyperparameter vector is

$$\xi_5 = (k_1, \ldots, k_T, r_1, \ldots, r_T, V_1, ...V_H, \beta_1^1, ..., \beta_H^1, \alpha_1^1, ..., \alpha_H^1, \beta_1^2, ..., \beta_H^2, \alpha_1^2, ..., \alpha_H^2,$$
$$\beta_1^3, ..., \beta_H^3, \alpha_1^3, ..., \alpha_H^3, \beta_1^4, ..., \beta_H^4, \alpha_1^4, ..., \alpha_H^4, p_1, p_2, p_3, p_4)$$

The finite approximation for the stick breaking is the same as in the previous sections. Note that we now have an additional prior for the common weights of the mixture of the trajectories. We assume that
$(p_1 \ldots p_J) \sim Dir(\frac{a}{J}, \frac{a}{J}, \frac{a}{J}, \frac{a}{J})$; in particular, we will take $a = 1$ and $J = 4$.

The joint distribution is:

$$p(y,\xi) = \prod_{h=1}^{H} Be(V_h \mid 1, M) \prod_{h=1}^{H}\prod_{j=1}^{J} N(\alpha_h^j \mid m_\alpha, \sigma_\alpha^2) \prod_{h=1}^{H}\prod_{j=1}^{J} N(\beta_h^j \mid m_\beta, \sigma_\beta^2) p(p_1, p_2, p_3, p_4)$$
$$\prod_{t=1}^{T} N(y_t \mid \alpha_{r_t}^{k_t} Y_{t-1} + \beta_{r_t}^{k_t}, \sigma^2) p(r_t \mid V_h) p(k_t \mid p_1, p_2, p_3 p_4)$$

Let $Q_h^j = \{t : r_t = h \; and \; k_t = j\}$ define the subset of observations with $\theta_h^j(Y_{t-1})$ and $n_{h_j} =\mid Q_{h_j} \mid$ is the number of observations tied to $\theta_h^j(.)$.

So we can impute now the complete conditional posterior distributions for each parameter involved in this model.

α_h^j: for $h = 1...H$ and $j = 1...J$. Let $y_t^* = (Y_t - \beta_h^j)$ for $t = 1, 2, \ldots, T$
$p(\alpha_h^j \mid r_t, k_t, V_h, \beta_h^j) \propto N(\alpha_h^j \mid m_\alpha, \sigma_\alpha^2) \prod_{t \in Q_{h_j}} N(y_t^* \mid \alpha_h^j Y_{t-1}, \sigma^2) = N(\alpha_h^j \mid m_1^j, V_1^j)$
where $m_1^j = V_1^j(\sigma_\alpha^{-2} m_\alpha + \sigma^{-2} \sum_{t \in Q_{h_j}} Y_{t-1} y_t^*)$ and $V_1^{(-1)j} = (\sigma_\alpha^{-2} + \sigma^{-2} \sum_{t \in Q_{h_j}} Y_{t-1}^2)$

β_h^j: for $h = 1...H$ and $j = 1...J$ and let $y_t^{**} = (Y_t - \alpha_h^j Y_{t-1})$

$p(\beta_h^j \mid r_t, k_t, V_h, \alpha_h^j) \propto N(\beta_h^j \mid m_\beta, \sigma_\beta^2) \prod_{t \in Q_{h_j}} N(y_t^{**} \mid \beta_h^j, \sigma^2) = N(\beta_h^j \mid m_2^j, V_2^j)$

where

$m_2^j = V_2^j(\sigma_\beta^{-2} m_\beta + n_{h_j} \sigma^{-2} \frac{1}{n_{h_j}} \sum_{t \in Q_{h_j}} y_t^{**})$
$V_2^{(-1)j} = (\sigma_\beta^{-2} + n_{h_j} \sigma^{-2})$.

The conditional posterior distribution for V_h given all the other parameters and the data does not change. The posterior distribution for r_t jointly with k_t is such that,

if $(r_t = h)$ and $(k_t = j)$ then

$$p(r_t = h, k_t = j \mid V_h, \alpha_h^j Y_{t-1} + \beta_h^j, p_1, p_2, p_3 p_4) \propto w_h p_j N(Y_t \mid \alpha_h^j Y_{t-1} + \beta_h^j, \sigma^2).$$

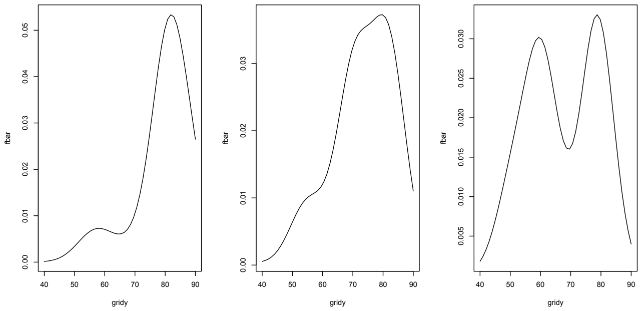

Figure 2.11: Bayesian density estimates for $E(F_{50} \mid y)$, $E(F_{65} \mid y)$, $E(F_{80} \mid y)$.

For Fubini's theorem, it is possible to integrate with respect to h and j. In the next chapter we will illustrate that the order of the conditioning is relevant for the application to the DNA-sequencing. We have

$$p(r_t = h \mid V_h, \alpha_h^j Y_{t-1} + \beta_h^j, k_t = j) \propto w_h \sum_{j=1}^{J} p_j N(Y_t \mid \alpha_h^j Y_{t-1} + \beta_h^j, \sigma^2)$$
$$p(k_t = j \mid r_t = h, V_h, \alpha_h^j Y_{t-1} + \beta_h^j) \propto p_j N(Y_t \mid \alpha_h^j Y_{t-1} + \beta_h^j, \sigma^2)$$

In Figure 2.11, we show Bayesian density estimate for the mixture of trajectories on the three subsets for a complete inferential point of view. We used the established initial values for the parameters that we discussed in the previous Subsection 2.3, and we check the form of the estimated densities for each subset, which is similar to the simple model discussed in equation 2.10.

2.5 Multivariate DDP-AR(1)

In this section we extend univariate model for the trajectories for the point masses to the multivariate case, considering separated equations. More details about this chice will be clarified in the next chapter, where we will consider single linear equations for each nucleotide.

2.5.1 Seemingly Unrelated Linear Equations

Here we illustrate multivariate linear trajectories extending univariate linear trajectories. The equations are

$$Y_{1,t} = \beta_1 + \alpha_{1,1} Y_{1,t-1} + \varepsilon_{1,t}$$
$$Y_{2,t} = \beta_2 + \alpha_{2,2} Y_{2,t-1} + \varepsilon_{2,t}$$

We do not assume interactions between the variables $Y_{1,t-1}$ and $Y_{2,t-1}$; we assume the *seemingly unrelated regressions* (S.U.R.). The errors are independent and unrelated and Gaussian distributed, so each of the equations is like the univariate case, showed in the previous sections. We analyze this special case, even if we should consider also the dependence between the two variables, waiting time, $y_{1,t}$ and the eruptions, $y_{2,t}$. This is a simplified version of the vector of Autoregressive model (V.A.R).

For this application, we consider the complete dataset. We have two variables the waiting time between eruptions, $y_{1,t}$, and the duration of the eruptions for the Old Faithful Gayser dataset, we call $y_{2,t}$. We know that the iteraction between the waiting time and the duration of the eruption is not zero, because more authours used the initial dataset Old Faithful Gayser for asimmetric dependence. In Figure 2.12, we show the duration of the eruptions (in minutes) at time $t-1$ for the x-axis and the duration of the eruption at time t for the y-axis.

In Figure 2.12, we detect three groups of points as for the waiting time. However, for this variable we should apply our univariate model. For the variable eruption, we should think about two possible subsets $y_{2,t-1} \in [1.5, 2.5]$ and $y_{2,t-1} \in [3.5, 5.0]$. For this second subgroup, looking through the y-axis we note two separate subgroups as a possible mixture of two distributions. For the waiting time $y_{1,t-1} \in [75, 85]$ there were long the y-axis one bimodal distribution. This is the reason that we point out the example for the univariate DDP-AR(1) using the waiting time and not the

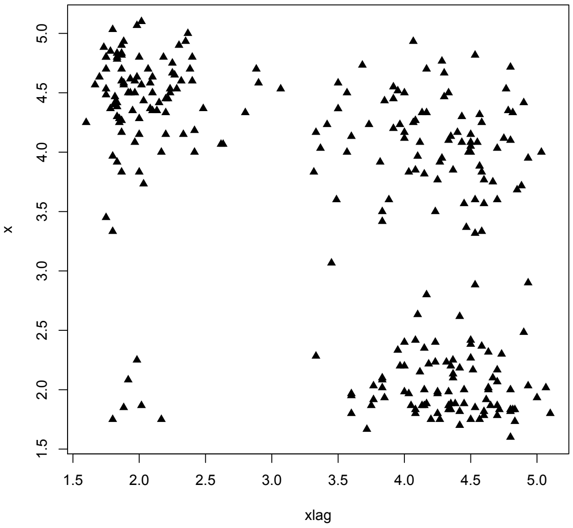

Figure 2.12: Scatterplot of the duration of the eruptions in Old Faithful Gayser dataset.

duration of the eruptions.

We estimate the following multivariate model

$$\mathbf{Y}_t = \mathbf{X}_t\mathbf{\Theta} + \boldsymbol{\varepsilon}_t$$

where:

$\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})$, $\mathbf{Y}_t = (Y_{1,t-1}, Y_{2,t-1}, 1, 1)$, $\mathbf{\Theta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \begin{pmatrix} \alpha_{1,1} & 0 \\ 0 & \alpha_{2,2} \end{pmatrix}$

and $\boldsymbol{\varepsilon}_t \sim N_2(\mathbf{0}, \mathbf{\Sigma})$.

For our hypothesis the covariance matrix is: $\mathbf{\Sigma} = diag(\sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \sigma^2_{\beta_1}, \sigma^2_{\beta_2})$.

2.5.2 Quadratic Form for the Trajectories

Here we use the properties of the matrices to extend the univariate quadratic trajectories to the multivariate case, assuming unrelated equations. This extension is useful for checking the number of activated clusters in the MCMC algorithm. When we will discuss about this point looking at the inferential results. We have the following structure:

$$Y_{1,t} = \alpha_1 Y_{1,t-1} + \beta_1 + \gamma_1 Y_{1,t-1}^2$$
$$Y_{2,t} = \alpha_2 Y_{2,t-1} + \beta_2 + \gamma_2 Y_{2,t-1}^2$$

to respect $Y_{1,t} \mid Y_{1,t-1}$ we assume the same priors:

$\alpha_1 \sim N(m_{\alpha_1}, \sigma^2_{\alpha_1}), \beta_1 \sim N(m_{\beta_1}, \sigma^2_{\beta_1})$ and $\gamma_1 \sim N(m_{\gamma_1}, \sigma^2_{\gamma_1})$.

For $Y_{2,t} \mid Y_{2,t-1}$ we assume that:

$\alpha_2 \sim N(m_{\alpha_2}, \sigma^2_{\alpha_2}), \beta_2 \sim N(m_{\beta_2}, \sigma^2_{\beta_2})$ and $\gamma_2 \sim N(m_{\gamma_2}, \sigma^2_{\gamma_2})$.

We found more convinient to reparametrize, as follows:

$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \boldsymbol{\varepsilon}_t$ for $t = 1, 2, \ldots, T$
where

$\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})$ in particular, $Y_{1,t}$ is the variable waiting time and $Y_{2,t}$ is the duration of the eruptions,

$\mathbf{Y}_{t-1} = (Y_{1,t-1}, Y_{2,t-1}, Y^2_{1,t-1}, Y^2_{2,t-1})$

$\mathbf{A} = (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2)$

$\boldsymbol{\varepsilon}_t \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma)$

Moreover we do not consider the interactions between the variables, α_{12} and α_{21}, that we suppose equal to zero. In general, we have the following equations for the model:

$$Y_{1t} = \alpha_{11} Y_{1t-1} + \beta_{11} + \alpha_{12} Y_{1t-1} + \gamma_{11}(Y_{1t-1} - \bar{Y_{1t}})^2$$
$$Y_{2t} = \alpha_{21} Y_{2t-1} + \beta_{21} + \alpha_{22} Y_{2t-1} + \gamma_{21}(Y_{2t-1} - \bar{Y}_{2t})^2$$

where $\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \beta_{11} & \gamma_{11} \\ \alpha_{21} & \alpha_{22} & \beta_{21} & \gamma_{21} \end{pmatrix}$

Note that in these two subsections we indicate as $Y_{1,t}$ for the generic form of the trajectories and also for the waiting time, the reason is only for easily notation. In these equations we consider $(Y_{1,t-1}-\bar{Y_{1,t}})^2$ and $(Y_{2,t-1}-\bar{Y_{2,t}})^2$; if the priors $\alpha_{.t}$, $\beta_{.t}$ and $\gamma_{.t}$[1] have mean equal to zero, i.e. $m_{\alpha_{1t}} = 0$, then the entire equation can simplified as $Y_{1t} = \alpha_{11}Y_{1t-1} + \beta_{11} + \gamma_{11}Y_{1t-1}^2$. For our application, we checked convergences diagnostics on the posterior distributions and we conclude that the quadratic form for the variables are sufficient.

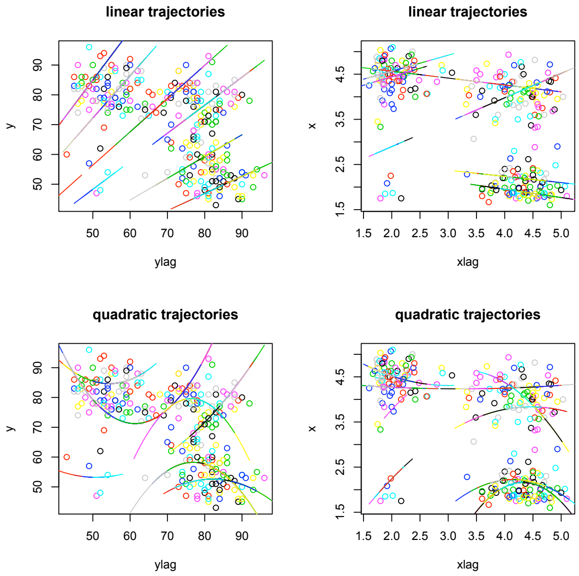

Figure 2.13: These plot are the linear and quadratic estimated trajectories for fixed variance of the likelihoods $\sigma_1^2 = 25$ and $\sigma_2^2 = 0.5$, respectively.

In Figure 2.13, we use the separation of the equations for a double comparison. Looking at the plot as rows of a matrix (2×2), we estimate linear trajectories for the two variables, $Y_{1,t} \mid Y_{1,t-1}$ and $Y_{2,t} \mid Y_{2,t-1}$. If we consider colomn direction, we can study how the quadratic trajectories are able to fit the real dataset for both the variables, waiting time between the eruptions and duration of the eruptions.

[1] the points indicates the case 1 and 2, i.e., α_{1t}, α_{2t}

2.5.3 Unseparable Linear Equations

We discuss about an other possible extension. Here we show the limits and the starting points for the assumptions of the next chapter. In the previous two subsections, we always considered the absence of correlation between the errors. The reason is that we propose a variation of the ANOVA-DDP model for time series. In the construction, we have two strong assumptions on the errors: uncorrelation and the normal distribution. We did not consider the linear relation between the waiting time and the eruptions, $y_{1t} = \beta + \alpha y_{2t} + \varepsilon_t$. Considering this variation, general linear equation is $\mathbf{Y}_t = \mathbf{Y}_{t-1}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$ for the DDP-AR(1) it means that will change the number of columns and rows of the regression, so we have

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & Y_{1,t-1} & Y_{t,2} & 0 & 0 \\ 0 & 0 & 0 & 1 & Y_{2,t-1} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \alpha_{11} \\ \alpha_{12} \\ \beta_2 \\ \alpha_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}$$

the corresponding equations are

$$\begin{aligned} Y_{t,1} &= \beta_1 + \alpha_{11} y_{1t-1} + \alpha_{12} Y_{t,2} + \varepsilon_{t,1} \\ Y_{t,2} &= \beta_2 + \alpha_{22} y_{2,t-1} + \varepsilon_{t,2} \end{aligned}$$

This kind of framework could hold in classical statistics. However, we cannot apply this construction to our DDP-AR(1) model, because we supposed that the parameters of the trajectories are i.i.d. in the base measure it means that the errors have to be unrelated. Nevertheless, we could insert an interaction on the independent variables as $Y_{t,2}Y_{1,t-1}$ for the parameter α_{12}, but this is not our case and the problem it is also the interpretation of a possible result, when it is reached. We assume that the errors are normal distributed, $\varepsilon_t \sim N(\mathbf{0}, \mathbf{\Sigma})$ and in this case the covariance matrix has to be:

$$\mathbf{\Sigma} = \begin{pmatrix} Var(Y_{t,1}) & 0 \\ 0 & Var(Y_{t,2}) \end{pmatrix}$$

If we suppose in the covariance matrix of the errors a correlation between the dependent variable of the waiting time $Y_{t,1}$ and the eruption, $Y_{t,2}$ which is the outcome in the second equation, we do not respect the assumptions of the DDP prior model in the ANOVA fashion-type. In the next chapter, we propose a multivariate DDP-AR(1), assuming that the errors are not correlated, but the existing correlations between the outcomes will be described in the prior of the covariance matrix introducing a fitting Wishart distribution.

2.6 Discussion

In this chapter we considered a single-p DDP prior for an autoregressive lag one model. We focused on the description of different kind of trajectories for the point masses. Mathematically, the single-p DDP prior for time series can be seen as a special case of the ANOVA-DDP proposed by De Iorio *et al.* (2004). The innovation is the asimmetric dependence as well as regressive linear model over time. Yet, we studied more possible forms of trajectories for point masses and we analyzed possible multivariate extentions. In particular, we analyzed the problem of density estimates in a Bayesian point of view, applying methodological results to Old Faithful Geyser dataset. The choice of this dataset is not random. In the literature, Azzalini and Bowman (1990) used the same dataset as an example for kernel density estimates for classical inference.

We proposed a mixture of the trajectories for the point masses, which has two relevant roles. Firstly, we decided to illustrate this possible extention in this chapter because here we based on the trajectories, evenif in terms of inference the mixture does not look really relevant. Secondly, this simple example of univariate mixture is the starting point for the next chapter where we will study the DNA-sequencing dataset. In particular, we will develop a multivariate mixture of the trajectories and we will illustrate the relevance of the latent variable, k_t, which will represent the label for the four nucleotides of the DNA. The other aspect of our contribution is that for the first time we introduced costant weights and the trajectories for the point masses will describe the data, Nieto-Barajas, Müller *et al.* (2011) introduced variable weights for the description of the data over time. We will discuss more details in the fourth Chapter.

Bibliography

Antoniak, C. E. (1974). ”Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems”. *Annals of Statistics*, **2**, 1152-1174.

Azzalini A. and Bowman A. W., (1990). ”A look at some data on the Old Faithful Geyser”. *Journal of the Royal Statistical Society*. SERIES C-APPLIED STATISTICS. **39**, 357-365.

Azzalini A. and Bowman A. W., (1997). ”Applied Smoothing Techniques for Data Analysis”. *Oxford Science Pubblications*.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2008). ”Bayesian inference for linear dynamic models with Dirichlet process mixtures”. *IEEE Transactions on Signal Processing*, **56**, 71-84.

Cifarelli, D. M. Regazzini, E. (1978). ”Problemi statistici non parametrici in condizioni di scambiabilit parziale: impiego di medie associative” *Quadermi Istituto di Matematica Finanziaria*. Serie III, nn. 12, Universitá Torino.

Contreras-Cristan, A., Mena, R. H. and Walker, S. G. (2009). ”On the construction of stationary AR(1) models via random distributions”. *Statistics*. **43**, 227-240.

Cruz-Marcelo, A., Rosner, G. R., Müller, P. and Stewart, C. (2010). ”Modeling Covariates with Nonparametric Bayesian Methods”. Available at SSRN: http://ssrn.com/abstract=1576665.

Dahl, D. B. (2006). ”Model-Based Clustering for Expression Data via Dirichlet Process Mixture Model”. In Vannucci, M. Do, K. A. and Müller, P. eds. *Bayesian Inference for Gene Expression and Proteomics.* Cambridge University Press.

De Iorio, M., M$\ddot{u}$ller, P., Rosner, G. R. and MacEachern, S. N., (2004).”An ANOVA Model for Dependent Random Measures”. *Journal of the American Statistical Association*, **99**, 205-215.

Doss, H. (1994). ”Bayesian Nonparametric Estimation for Incomplete Data via Successive Substitution Sampling”. *Annals of Statistics*. **22**, 1763-1786.

Escobar, M. (1988). ”Estimating the means of several normal populations by estimating the distribution of the means. Ph.D. dissertation, Dept. Statistics, Yale Univ.

Escobar, M. D. and West, M. (1995). ”Bayesian density estimation and inference using mixtures”. *Journal of American Statistical Association*, **90**, 577-588.

Ferguson, T. S., (1973). ”A Bayesian analysis of some nonparametric problems”. *Annals of Statistics*, **1**, 209-230.

Gasparini, M., (1996). ”Bayesian density estimation via mixture of Dirichlet processes”. *Journal Non Parametric Statistics*. **6**, 355-366.

Geyer, C. J. (1995). ”Conditioning in Markov Chain Monte Carlo”. *Journal of Computational and Graphical Statistics*. **4**, 2, 148-154.

Geweke, J. ”Evalueting the accuracy of sampling-based approaches to calculating posterior moments”. *In Bayesian Statistics 4* eds. JM Bernado, JO Berger, AP Dawid and AFM Smith. Clarendon Press, Oxford, UK.

Griffin, J. E., Steel, M., (2006). ”Order-Based Dependent Dirichlet Processes”. *Journal of the American Statistical Association* Theory and Methods, **101**, 179-194.

Griffin, J. E., Steel, M., (2011). ”Order-Based Dependent Dirichlet Processes”. *Journal of Econometrics*, **162**, 2, 383-396.

Härdle, W. (1991). *Smoothing techniques: with implementation in S*. Springer, New York.

Hjort, N., Holmes, C., Müller, P. and Walker, S. G. (2010). ”Bayesian Nonparametrics” *Cambridge: Cambridge University Press*.

Hutchinson, R. A., Westphal, J. A. and Kieffer, S. W. (1997). ”In situ observations of old faithful geyser.” *Geology*, **25**, 875-878.

Ishwaran, H. and James, L. F. (2001). ”Gibbs sampling methods for stick-breaking priors”. *Journal of American Statistical Association*, **96**, 161-173.

Ishwaran, H. and James, L. F. (2002). ”Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information.” *Journal of Computational Graphical Statistics*. **11**, 508-532.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23**, 727-741.

MacEachern, S. N., M$\ddot{u}$ller, P. (2000). ”Estimating Mixture of Dirichlet Process Models”. *Journal of Computational and Grafical Statistics*. **7** 223-239.

MacEachern, S. N. and M$\ddot{u}$ller, P. (2000). ”Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichelt Process Mixture Models”. *Robust Bayesian Analysis. Lecture Notes in Statistics*. **152**, 295-316.

Mena, R. H. and Walker, S. G. (2005). ”Stationary Autoregressive Models via A Bayesian Nonparametric Approach”. *Journal of Time Series Analysis*. **26**, 6.

Muliere, P. and Scarsini, M. (1982). ”Il Modello Lineare: Inferenza e Previsione.” *Studi statistici n.1*. Universitá Bocconi.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors”. *Canadian Journal of Statistics*. **26**, 283-298.

Müller, P., West, M. and MacEachern, N. S., (1997). ”Bayesian Models for Non-Linear Autoregressions”. *Journal of Time Series Analysis*. **18**, 593-614.

Rodriguez, A. and ter Horst, E. (2008). ”Bayesian Dynamic Density Estimation”. *Bayesian Analysis*. **4**, 793-816.

Sethuraman, J. (1994). ”A constructive definition of the Dirichlet process prior”, *Statistica Sinica*. **2**, 639-650.

Silverman, B. W. (1986). ”Density Estimation for Statistics and Data Analysis”. *Chapman and Hall, London*.

Weisberg, S. (1994). ”Applied Linear Regression”, third Edition. *Wiley-Interscience, New York*. **18**.

Wood, S., Rosem, O. and Kohn, R. (2011). ”Bayesian mixtures of autoregressive models”. *Journal of Computational and Graphical Statistics*. **20**, 174-195.

Chapter 3

A Nonparametric Autoregressive Model for DNA-sequencing

Abstract.
We consider the problem of base calling for data from high throughput sequencing (HTS) experiments. We propose a Bayesian nonparametric approach. The proposed model generalizes earlier approches based on mixtures of normals to mixtures of random probability measures. Complication arises from the inherently autoregressive nature of real data (phasing, fading and cross talk between channels). We use a variation of dependent Dirichlet process models (DDP) that define a nonparametric vector autoregressive model for the four-dimensional output from the four channels of the sequencing experiment.

3.1 Introduction

The term DNA-sequencing refers to sequencing methods which aim to determine the primary structure of an unbranched biopolymer. Solexa sequencing is the digital version of the classical microarray technology, it measures the exact number of gene copies rather than the relative aboundance. This is the new version for the new ultra highthrouput sequencing. Solexa sequencing technology breaks a long genome in fregments of short DNA tags. Each DNA tag gives rises to a set of quadriple vectors. Each vector contains four fluorescent intensities of nucleotides. The sequencing method determines the order of the nucleotide bases - Adenine, Guanine, Cytosine and Thymine - in a molecule of DNA. For simplicity we call them A, G, C and T respectively. Knowledge of DNA is fundamental for biological research and also for other branches which utilize DNA-sequencing. The determination of the DNA-sequencing from nucleotides fluorescent intensities is called Base Calling. Ji *et al.* (2011) used a DNA-sequencing dataset for a Bayesian parametric model. We implement a Bayesian nonparametric inference for the same dataset. The dataset contains 1000 colonies. Each colony corresponds to a DNA segment with 36 bases. For each base there are four nucleotide intensity measurements. Substantially, the dataset has 36,000 intensity measurements (observations) and four nucleotides (variables). The focus is to estimate this sequence of each short tag. The experiment involves three major sources of noises: *fading*, *phasing* and *cross-talk of channels*. These three sources of noises represent a motivation for a different probability model for Ji *et al.* (2011) and also for our proposal. The accuracy of Base calling is affected also from other kind of noises that we are able to resolve standardizing the initial dataset, for more details see Rougemont *et al.*, (2008). For example, Ji *et al.* (2011) standardized the initial DNA dataset to respect the minimum value and dividing by the stardard deviation of the data. The reason was a methodological and empirical comparison to respect the different methods. We will operate a traditional stardardization: to respect the vector of the mean of the data and divide by the stardard deviation of the initial data.

For this dataset is really relevant the accuracy of an initial explorative analysis.

Fading noise refers to the exponential decay in the intensity as a function of cycle number. Within each colony, as the cycle number increases, the intensity measurement decreases. This is usually caused by material loss during the sequencing process.

Phasing noise is an error which involves more than one nucleotide in one cycle, thus increasing the noise in the signal output for down-stream cycles. The precision of base calling drops as cycle number increases.

Cross-talk between channels induces high correlations between A and C intensities, and between G and T intensities, respectively.

Ji *et al.* (2011) resolved these three kind of problems modeling a four dimensional mixture of truncated Gaussian distributions. However, we study the same problem using the same dataset in a Bayesian nonparametric context, because we use the efficiency and flexibility properties of the nonparametric framework. We assume a dependent Dirichlet process (DDP) prior for a vector autoregressive lag one, briefly, DDP-VAR(1). Finally, we compare the two models. Figure 3.1 shows the data for the first 1,000 colonies, i.e., $\mathbf{y}_t$ for the first $T = 36 \cdot 1000 = 36,000$ bases, for $t = 1, \ldots, T$. The DNA dataset has 36,000 intensity measurements and four variables. Figure 3.1 illustrates on the left hand side y_{t2} versus y_{t1} and on the right hand side y_{t4} versus y_{t3}. The first two dimensions correspond to the channels recording A and C. The last two dimensions correspond to G and T. The plot clearly shows the correlation between the channels. Phasing and fading noises can not directly be seen in this plot. Figure 3.2 illustrates two scatterplots of the intensity measurements and we plug in the labels of the true sequence of these intensity measurements. Ideally, the true sequence is the dataset with the labels for the four nucleotides with highest fluorescence intensities. The true sequence is possible for this specific dataset on enterobactaria phage. The kind of bacteria gives the true sequence. For human genome is not biological possible, see for more details Ji *et al.* (2011). We call true sequence a sequence of tags aligned to phage genome. There is a sequencing error when the observed nucleotide is coming from the base calling method and it does not match the nucleotide in the true sequence. In principal, the sequence is true for two reasons: there is not polymorphism in the phage genome and the small genome size makes a mistaken sequence match over 36 nucleotides highly unlikely. Statistically, we do not have a label switching problem because we have the true sequence. Follow that we define a variable, k_t such that

- $k_t = 1$ if the highest intensity measurement is Adenine
- $k_t = 2$ if the highest intensity measurement is Cytosine
- $k_t = 3$ if the highest intensity measurement is Guanine
- $k_t = 4$ if the highest intensity measurement is Thymine

  for $t = 1, 2, \ldots, T$

For example, consider in Figure 3.2 the plot on the left hand side. The x-axis is the intensity measurements of the nucleotide Adenine and the y-axis is the Cytosine intensity measurement. Each point in the scatterplot represent a couple of the coordinates of these nucleotides. The color for each point is the label of the true sequence of the intensity measurements. Indeed, the intensity corresponding to the correct base is highest. The red color of the points comes from the attributed labels and corresponds to Cytosine intensity measurement. Assume that a generic point of Figure 3.2 has coordinates (0,2), this point has bigger Cytosine intensity measurement than Adenine. Follow that the variable k_t assumes value 2 and the point in the scatter plot is red. A counter example is a generic point which has coordinates (2,0), in this case the color of the point is black because the Adenine intensity measurement is bigger than Cytosine and it corresponds to the label $k_t = 1$. Similarly, for the scatterplot on the right hand side. Figure 3.2 has also an other important role. We observe that the form of the clouds is not a mixture of normal distributions. In particular, the black cloud in the plot on the left hand side which represents the strong presence of high intensity measurements for Adenine it does not look like a bivariate normal distribution, but it is closer to a mixture of two bivariate normal distributions. This is one more reason that motivated us to develop a Bayesian nonparametric model. Ji *et al.* (2011) defined a true sequence in a separate dataset which is the entire phase genome of 5386 positions to search the best matches. They used Solexa Software Phage Align. To clarify this point, suppose to have a tag of DNA which will be aligned to the phage genome, when the tag is aligned to the phage genome the matched sequence on the phage genome is considered to be the true sequence and any mismatched nucleotide is considered a sequence error. Ji *et al.* (2011) compare the performance of the Base Calling method with the true sequence and specifically they considered mismatching when there is a difference between the measurement of the true sequence and the measurement associated to the observed Base Calling dataset. The Base Calling method produces the fluorescent intensities measuremtents. Ewing and Green (1998) do not have the true sequence, so they focus on the error probabilities in the sequence. Lawrence and Solovyev (1994) used the discriminant analysis to separate the true sequence by the Base Calling fluorescent intensity measurements. For the Base Calling of Solexa sequencing data, Rougemont *et al.* (2008) identified the presence of phasing noise and cross-talk between channels and they used a mixture of Gaussian distributions. Each element of the mixture was keeping one of the four variables but they did not consider the phasing noise. The consequence of this mistake was reveled in them inferential results: they showed that the limit of them methodology was in the errors.

We illustrate more explorative aspects for the noises. Phasing noise is present when the intensity scores at cycle t depend (at least) on the ones at cycle $(t-1)$.

Figure 3.3 shows strong presence of the dependence between the cycles and this represents why we introduce an autoregressive model. Specifically, in each plot of Figure 3.3 we have the intensity scores at time t depending on more than one cycle. which is one of the caracteristics of a real DNA dataset. We will use a multivariate autoregressive lag one linear model for each nucleotide, because it is

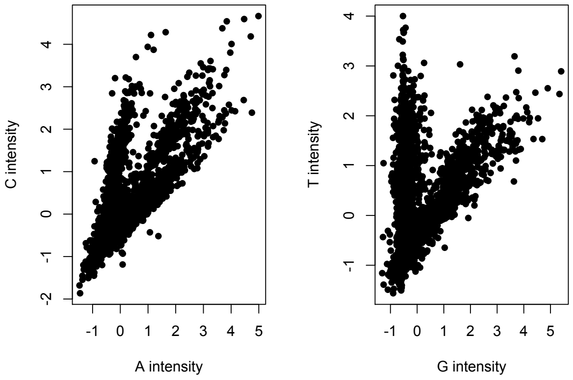

Figure 3.1: Standardized DNA-sequencing dataset with (35000×4) observations.

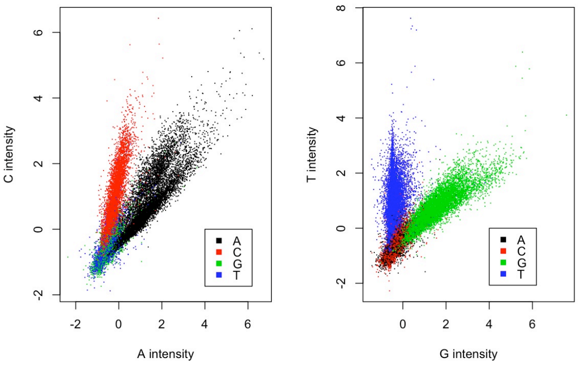

Figure 3.2: Dataset with (35000×4) observations. The different colors come from the true sequences.

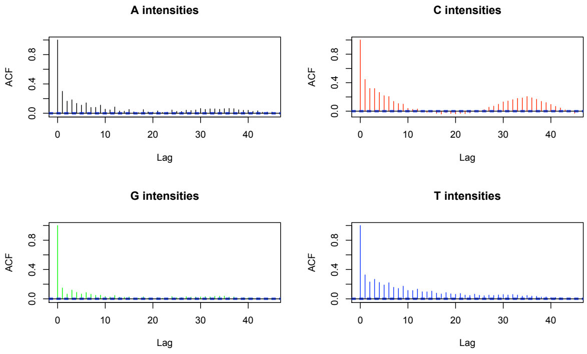

Figure 3.3: The colors of the bars for the autocorrelation function correspond to the four intensities: black for A, red for C, green for G and blue for T.

the highest level of phasing noise for each intensity scores.

Fading noise is not easy to single out on the data, because the intensities decay over cycles.

Figure 3.4 and Figure 3.5 show two of the four boxplots for the intensity measurements A and G, respectively on 35 cycles. We choose the boxplot graphical representation for the presence of outliers in the standardized dataset. On each plot we indicate dashed line for the average of the standardized data and dotdashed line for the exponential function. Focusing on Figure 3.5, for example, we observe positive asymmetry of each boxplot at each cycle, this is the effect of the presence of the fading noise.

The plan of this chapter is as follows. In Section 3.2, we describe more details of our methodological choices. In Section 3.3, we illustrate our assumptions and in Section 3.4 we show posterior distributions for our model. In Section 3.5 we illustrate our inferential results. A final discussion in Section 3.6 will conclude this chapter.

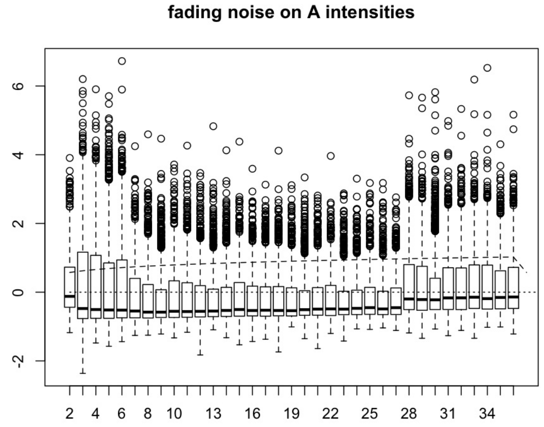

Figure 3.4: Fading noise for Adenine

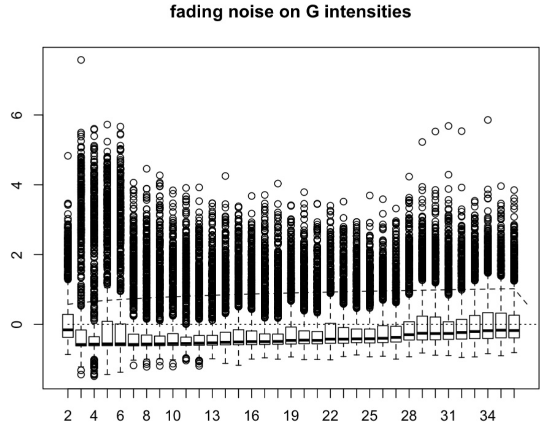

Figure 3.5: Fading noise for Guanine

3.2 Multivariate Mixture for DNA-Sequencing

We briefly trace an overview of previous modeling for this dataset, underling the critical points and the reasons of our decisions of proceeding in a nonparametric modeling.

3.2.1 Parametric Vector Autoregressive Model

Let $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{36}\}$ denote the 36 quadruples of nucleotide intensity measurements for one colony. The four dimensions of $\mathbf{y}_t$ are the four channels of the sequencing machine corresponding to the nucleotides A, C, G, and T. We write, $\mathbf{y}_t = (y_{t1}, y_{t2}, y_{t3}, y_{t4})$ for each cycle $t = 1, 2, ..., 36$. The (4×1) vector, $\mathbf{y}_t$ represents respectively the intensities of four nucleotides at location t of the DNA tag.
We assume that y_{tj} have been standardized to zero mean and unit standard deviation (across the entire data). Let $k_t \in \{1, 2, 3, 4\}$ denote the (unknown) true basis, i.e. $i(t) = (t \; mod \; 36) + 1$. Conditional on k_t we build a sampling model for $\mathbf{y}_t$ that represents fading, phasing and cross-talk noises. Fading noise is simple. Let $i(t) \in \{1, \ldots, 36\}$ denote the cycle for the t-th base, when t is obvious from the context we simply write i. We include a term $\boldsymbol{\beta}_j \cdot e^{-\lambda i^{\gamma}}$ with parameters $\boldsymbol{\beta}_j \in R^4$, and $\lambda, \gamma \in [0, 1]$. The term represents an exponential fading signal. We use $\boldsymbol{\beta}_j$ specific to the (latent) base $k_t = j$, and common λ, γ across bases. Cross-talk is represented by allowing for non-zero correlation of the 4-dimensional response $\mathbf{y}_t$ using a (4×4) non-diagonal covariance matrix $\boldsymbol{\Sigma}_j$. Phasing is more difficult. It requires modeling of dependence across cycles. We use the simplest possible structure by assuming a first-order autoregressive process, regressing $\mathbf{y}_t$ on $\mathbf{y}_{t-1}$. For two vectors $\mathbf{x}, \mathbf{y} \in R^4$, let $\mathbf{x} * \mathbf{y}$ denote the elementwise product $(x_1 y_1, \ldots, x_4 y_4)'$. We include a term $\boldsymbol{\alpha}_j * \mathbf{y}_{t-1}$, allowing for different $\boldsymbol{\alpha}_j$ for each base j. In summary, we consider

a vector autoregression with lag one, VAR(1)

$$\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{Y}, k_t = j \sim N_4(\boldsymbol{\theta}_j(\mathbf{y}, i), \Sigma_j) \text{ with } \boldsymbol{\theta}_j(\mathbf{y}, i) = \boldsymbol{\beta}_j e^{-\lambda i^\gamma} + \boldsymbol{\alpha}_j * \mathbf{y} \qquad (3.1)$$

for $i = 2, \ldots, 36$.

Phasing only occurs within the cycles of a colony, i.e., the VAR model only applies to cycle 2 through 36. One could include a separate model for cycle 1 data. For simplicity we drop cycle 1 responses, focusing only on data for cycles 2 through 36.

3.2.2 Nonparametric VAR(1)

Comparing with Figure 3.2 we notice that the assumption of a single multivariate normal distribution, as above, is not sufficient to model this dataset. The normal assumption for each component of the mixture is too restrictive. This leads us to consider a nonparametric extension, replacing the multivariate normal model $N_4(\cdot, \cdot)$ with a random probability measure G. In other words, we treat G as an unknown quantity and complete the model with a prior for the unknown G.

Probability models, $p(G)$, for random distributions are known as Bayesian nonparametric priors. See, for example, Hjort, Holmes, Müller and Walker (2010) for a recent review. One of the most commonly used prior models for an unknown distribution is the Dirichlet process (DP) prior (Ferguson, 1973). We write $G \sim$ DP(G^0, M). One of the important features of the DP prior is the discrete nature of the random probability measure G. We can therefore write G as a sum of point masses as a constructive representation of the Dirichlet process

$$G = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h}.$$

The DP model is indexed by two parameters, the base measure G^0 and the total mass parameter M. One of the definiting properties of the DP is that the point masses are $\theta_h \sim G^0$, independent and identically distributed and the weights are defined as $w_h = V_h \prod_{\ell<h}(1 - V_\ell)$ with fractions $V_h \sim \text{Be}(1, M)$. This construction is known as stick-breaking representation (Sethuraman, 1994). For many applications, the discrete nature of G is inappropriate. An easy and commonly used is a convolution with an additional (continuos) kernel, for example a Gaussian kernel:

$$G = \sum_{h=1}^{\infty} \omega_h N(\theta_h, \Sigma) \tag{3.2}$$

This construction is known as DP mixture model and widely used in the literature as we discussed in the previous first two chapters.

The challenge in this application is that the multivariate normal model in (3.1) includes a regression on $\mathbf{y}_{t-1}$, i.e., an autoregression. In other words, rather than *one* unknown probability measure F, we need a probability model for an entire *family* of unknown probability measures on the class $\mathcal{F}_j = \{F^j_{\mathbf{y},i}(\cdot);\ \ \mathbf{y} \in R^4, i = 2, \ldots, 36\}$ to replace (3.1) by

$$\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{y}, k_t = j \sim F^j_{\mathbf{y},i}(\mathbf{y}_t).$$

We use an extension of the DP prior to define a probability model for $\mathcal{F}_j$. The dependent Dirichlet process (DDP) (MacEachern, 1999; 2001) defines a prior for a family of distributions $\mathcal{F} = \{F_y,\ y \in \mathcal{Y}\}$ indexed by a covariate y as a natural extension of the stick-breaking construction:

$$F_y = \sum_{h=1}^{\infty} \omega_h N(\theta_h(y), \Sigma). \tag{3.3}$$

The point mass θ_h of the stick-breaking representation for the DP is replaced by a stochastic process $\theta_h(y)$. In the simplest case $\theta_h(y)$ could be a parametric model, for example $\theta_h(y) = a_h + b_h y$. Many variations are possible, including alternative constructions with varying weights that replace w_h by $w_h(y)$. We will discuss about varying weights in the next chapter. Here we restrict the model to the simple case of costant weights w_h and a parametric trajectory for the point masses. We construct a prior for $\mathcal{F}_j$ by using a special instance of the DDP

$$\theta^j_h(\mathbf{y}, i) = \boldsymbol{\beta}^j_h e^{-\lambda i^\gamma} + \boldsymbol{\alpha}^j_h * \mathbf{y}.$$

This defines a nonparametric extension of the normal VAR (3.1) by defining

$$\mathbf{Y}_t \mid \mathbf{Y}_{t-1}, \mathbf{y}, k_t = j \sim F^j_{\mathbf{y},i} \text{ with } F^j_{\mathbf{y},i} = \sum_{h=1}^{\infty} w_h N_4(\theta^j_h(\mathbf{y}, i), \boldsymbol{\Sigma}_j). \tag{3.4}$$

This model resolves jointly these three kind of noises. The autoregression in the DDP captures the dependence between the consecutive cycles, resolving the phasing noise. We use an unknown covariance matrix, $\boldsymbol{\Sigma}_j$, for $\mathbf{y}_t$, which is able to capture high correlation between channels and finally for the fading noise we introduced the exponential function in the trajectories. Note that the trajectories include two indices: h, j and i. The h indexes the elements of the stick-breaking (we will clarify this point in the next equation 3.5) and j is the (unknown) base. In summary, the model for the DNA-sequencing is

$$\begin{aligned} \mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{y}, r_t = h, k_t = j \quad &\sim \quad N(\theta^j_h(\mathbf{y}, i), \boldsymbol{\Sigma}_j) \\ \theta^j_h(\mathbf{y}, i) = \boldsymbol{\beta}^j_h e^{-\lambda i^\gamma} + \boldsymbol{\alpha}^j_h \times \mathbf{y} & \end{aligned} \tag{3.5}$$

for $h = 1, 2..., H,\ j = 1, 2, \ldots, J$ and $t = 2, \ldots, 36$

In equation 3.5, on the left hand side, we only highlight the conditioning on the latent variables r_t and k_t, dropping all other variables in the conditioning set for notational simplicity. For the latent indicator r_t the prior is

$$p(r_t = h) \quad = \quad w_h.$$

The model is completed with a prior for the latent unknown true base, indexed by k_t

$$p(k_t = j) \quad = \quad p_j.$$

The DDP model includes a prior for the parameters of the trajectories for the point masses:

$$(\boldsymbol{\beta}_h^j, \boldsymbol{\alpha}_h^j) \quad \sim \quad G_j^0(\boldsymbol{\beta}_h^j, \boldsymbol{\alpha}_h^j). \tag{3.6}$$

We make one more important semplification. We use a finite DP by replacing (3.3) with

$$G_y = \sum_{h=1}^{H} w_h \delta_{\theta_h}(y) \tag{3.7}$$

with a finite number of $H < \infty$ terms instead of the infinite discrete measure in (3.3). The stick-breaking prior for the weights is simply truncated with $V_H = 1.0$, as we defined in the previous chapter.

Follow that our model (3.5) can be marginalized to respect k_t

$$\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{y}, r_t = h \sim \sum_{j=1}^{J} p_j N_4(\boldsymbol{\alpha}_h^j \times \mathbf{y}_{t-1} + e^{-\lambda i^\gamma} \boldsymbol{\beta}_h^j, \boldsymbol{\Sigma}_j)$$

or equivalently to respect r_t

$$\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{Y}, k_t = j \sim \sum_{h=1}^{H} w_h N_4(\boldsymbol{\alpha}_h^j \times \mathbf{Y}_{t-1} + e^{-\lambda i^\gamma} \boldsymbol{\beta}_h^j, \boldsymbol{\Sigma}_j).$$

3.3 Prior Probability Model

For reference we indicate the full vector of all parameters and latent variables in the model (3.1) as:

$$\xi = (r_1, \ldots, r_T, k_1, \ldots, k_T, \ldots, V_1, \ldots, V_{H-1}, \boldsymbol{\beta}_1^1, \ldots, \boldsymbol{\beta}_H^1, \boldsymbol{\alpha}_1^1, \ldots, \boldsymbol{\alpha}_H^1,$$

$$\boldsymbol{\beta}_1^2, \ldots, \boldsymbol{\beta}_H^2, \boldsymbol{\alpha}_1^2, \ldots, \boldsymbol{\alpha}_H^2, \boldsymbol{\beta}_1^3, \ldots, \boldsymbol{\beta}_H^3, \boldsymbol{\alpha}_1^3, \ldots, \boldsymbol{\alpha}_H^3, \boldsymbol{\beta}_1^4, \ldots, \boldsymbol{\beta}_H^4, \boldsymbol{\alpha}_1^4, \ldots, \boldsymbol{\alpha}_H^4,$$

$$\gamma, \lambda, p_1, p_2, p_3, p_4, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \boldsymbol{\Sigma}_4)$$

Recall that the fractions V_h define the weights $\omega_h = V_h \prod_{l<h}(1-V_l)$. The DP prior for G_y implies $V_h \sim Be(1, M)$ for $h = 1...H-1$ and $V_H = 1$. The condition $V_H = 1$ arises from the approximation of the finite Dirichlet process truncated at H, DP_H, as we defined in the previous chapter.

The base measure G_j^0 of the DDP is defined by

$$\boldsymbol{\alpha}_h^j \sim N_4(\mathbf{m}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \text{ for } h = 1, \ldots, H, \ j = 1, \ldots, J$$

and

$$\boldsymbol{\beta}_h^j \sim N_4(\mathbf{m}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \text{ for } h = 1, \ldots, H, \ j = 1, \ldots, J$$

Here $\mathbf{m}_{\boldsymbol{\alpha}} = (m_{\alpha_1}, m_{\alpha_2}, m_{\alpha_3}, m_{\alpha_4})^t$ and $\mathbf{m}_{\boldsymbol{\beta}} = (m_{\beta_1}, m_{\beta_2}, m_{\beta_3}, m_{\beta_4})^t$ are vectors (4×1) mean for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Also $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = diag(\sigma^2_{\alpha_1}, \sigma^2_{\alpha_2}, \sigma^2_{\alpha_3}, \sigma^2_{\alpha_4})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = diag(\sigma^2_{\beta_1}, \sigma^2_{\beta_2}, \sigma^2_{\beta_3}, \sigma^2_{\beta_4})$ are fixed (4×4) covariance matrices. The model for $(\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{y}, r_t = h, k_t = j)$ is completed with priors on the remaining parameters. In particular, we introduce unknown weights for the elements of the mixture and for the prior probabilities p_j, for $j = 1, \ldots, 4$ we assume a conditional conjugate Dirichlet prior distribution $p_j \sim Dir(\frac{a}{J}, \frac{a}{J}, \frac{a}{J}, \frac{a}{J})$ and we suppose $a = 1$ and $J = 4$.

For the two parameters of the base number, λ and γ, we assume (for each of them) a prior discrete Uniform distribution, as in Ji *at al.* (2011): $\lambda \sim Be(1,1)$ and $\gamma \sim Be(1,1)$, respectively. They describe the exponential function for fading noise. Finally we use a Wishart random variable for the covariance matrix of the likelihood: $\mathbf{\Sigma}_j^{-1} \sim W(\mathbf{\Psi}^{-1}, m)$ for $j = 1, 2, 3, 4$

where $\mathbf{\Psi} = \begin{pmatrix} \hat{S}_A^2 \hat{S}_{AC} \hat{S}_{AG} \hat{S}_{AT} \\ \hat{S}_{CA} \hat{S}_C^2 \hat{S}_{CG} \hat{S}_{CT} \\ \hat{S}_{GA} \hat{S}_{GC} \hat{S}_G^2 \hat{S}_{GT} \\ \hat{S}_{TA} \hat{S}_{TC} \hat{S}_{TG} \hat{S}_T^2 \end{pmatrix}$ is the empirical covariance matrix, the degree of freedom are $m = 6$ and i.e. $E[\mathbf{\Sigma}_j^{-1}] = \mathbf{\Psi}^{-1} m$.

3.4 Posterior Distributions

We start by studying the joint distribution for the parameters described in the hyper-parameter space.

$$p(y, \xi) = \prod_{h=1}^{H-1} Be(V_h \mid 1, M) \left[\prod_{h=1}^{H} \prod_{j=1}^{J} N_4(\boldsymbol{\alpha}_h^j \mid \mathbf{m}_{\boldsymbol{\alpha}}, \mathbf{\Sigma}_{\boldsymbol{\alpha}}) N_4(\boldsymbol{\beta}_h^j \mid \mathbf{m}_{\boldsymbol{\beta}}, \mathbf{\Sigma}_{\boldsymbol{\beta}}) \right] p(\lambda) p(\gamma)$$

$$p(p_1, p_2, p_3, p_4) \prod_{j=1}^{J} p(\mathbf{\Sigma}_j) \prod_{t=2}^{T} p(r_t \mid V_h) N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \mathbf{\Sigma}_{k_t}) p(k_t \mid p_1 p_2 p_3 p_4) =$$

$$= p(\lambda) p(\gamma) p(p_1, p_2, p_3, p_4) \prod_{j=1}^{J} p(\mathbf{\Sigma}_j) \prod_{h=1}^{H-1} Be(V_h \mid 1, M) \prod_{h=1}^{H} \prod_{j=1}^{J} N_4(\boldsymbol{\alpha}_h^j \mid \mathbf{m}_{\boldsymbol{\alpha}}, \mathbf{\Sigma}_{\boldsymbol{\alpha}})$$

$$\prod_{h=1}^{H} \prod_{j=1}^{J} N_4(\boldsymbol{\beta}_h^j \mid \mathbf{m}_{\boldsymbol{\beta}}, \mathbf{\Sigma}_{\boldsymbol{\beta}}) \prod_{t=2}^{T} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \mathbf{\Sigma}_{k_t}) \prod_{t=1}^{T} \omega_{r_t} \prod_{t=1}^{T} p_{k_t}$$

From the joint distribution, we realize that it is not possible to have more analitical results. To implement an MCMC algorithm, (we use a blocked Gibbs sampler algorithm,) we calculate the full conditional distributions.

$$\begin{aligned} p(\lambda \mid \xi \setminus \lambda \mid \mathbf{y}) \propto p(\lambda) \prod_{t=1}^{T} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_{k_t}) &= \\ 1 * \prod_{t=1}^{T} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_{k_t}) & \end{aligned}$$

$$\begin{aligned} p(\gamma \mid \xi \setminus \gamma \mid \mathbf{y}) \propto p(\gamma) \prod_{t=1}^{T} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_{k_t}) &= \\ 1 * \prod_{t=1}^{T} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_{k_t}) & \end{aligned}$$

To implement these two complete conditional distributions, we have to use two Metropolis-Hasting steps in the blocked Gibbs sampler algorithm because they are not two multivariate normal distributions to respect the two parameters of interest, λ and γ, respectively. Indeed, we define the two conditional posterior distributions considering the domain on the interval $[0, 1]$. Ji *et al.* (2011) introduced this smaller domain for the presence of the truncated Gaussian distribution in the model. This transformation for each parameter will give closer results to the real posterior distribution. We adopt the logit transformation to reduce the domain on the same interval, because we believe that the presence of the exponential function, in the locations of the mixtures, is stronger in a smaller domain than in a bigger domain. We define: $L_\lambda = logit(\lambda)$ so $L_\lambda = log\lambda - log(1-\lambda)$. The focus is to apply the formula $P_\lambda(\lambda(L)) \mid \frac{\partial L_\lambda}{\partial \lambda} \mid$ and the first derivative to respect the parameter of interest is $\frac{\partial L_\lambda}{\partial \lambda} = \frac{1}{\lambda} + \frac{1}{1-\lambda} = \frac{1}{\lambda(1-\lambda)}$ obtaining $P_L(L) = P_\lambda(\lambda = \frac{e^L}{1+e^L}) \mid \lambda(1-\lambda) \mid$. Note that $\frac{e^{L_\lambda}}{1+e^{L_\lambda}}$

is the candidate value for the Metropolis-Hasting step. The full conditional posterior distribution for λ is

$$p(L_\lambda \mid \xi \setminus \{\lambda \mid \mathbf{y}\}) \propto log(\lambda) + log(1-\lambda) + \sum_{t=2}^{T} logN_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{rt}^{k_t}\mathbf{Y}_{t-1} + \boldsymbol{\beta}_{rt}^{k_t} e^{-\lambda i^\gamma}, \boldsymbol{\Sigma}_{k_t})$$

the equation above is the density for the ratio in the Metropolis-Hastings step. The presence of the Gaussian distribution semplifies the ratio, because it is a symmetric density and it is sufficient to substitute the candidate value and the initial value in this density and check when the ratio is accepted.
Similarly, we compute the conditional posterior distribution function for γ given all the other parameters. The density for the ratio in the Metropolis-Hastings algorithm is on the domain $[0,1]$, obtaining:

$$p(L_\gamma \mid \xi \setminus \{\gamma \mid \mathbf{y}\}) \propto log(\gamma) + log(1-\gamma) + \sum_{t=1}^{T} logN_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{rt}^{k_t}\mathbf{Y}_{t-1} + \boldsymbol{\beta}_{rt}^{k_t} e^{-\lambda i^\gamma}, \boldsymbol{\Sigma}_{k_t}).$$

The next conditional posterior distribution is the covariance matrix, $\boldsymbol{\Sigma}_j$, which plays a relevant role in the model, this captures the *cross-talk between the channels*,

$$p(\boldsymbol{\Sigma}_j \mid \xi \setminus \{\boldsymbol{\Sigma}_j \mid \mathbf{y}\}) \propto p(\boldsymbol{\Sigma}_j) \prod_{t \in B_j} \mid \boldsymbol{\Sigma}_j \mid^{-\frac{n_{B_j}}{2}} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{j}\mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{j} e^{-\lambda i^\gamma}, \boldsymbol{\Sigma}_j)$$

We use $\boldsymbol{\theta}_{r_t}^{j} := \left[\boldsymbol{\alpha}_{r_t}^{j}\mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{j} e^{-\lambda i^\gamma}\right]$ (hiding the dependence on $\mathbf{Y}_{t-1}$ and i for notational simplicity). So the complete conditional posterior distribution is

$$p(\boldsymbol{\Sigma}_j \mid \xi \backslash \{\boldsymbol{\Sigma}_j\}) = \frac{\mid \boldsymbol{\Psi} \mid^3 \mid \boldsymbol{\Sigma}_j \mid^{-\left(\frac{n_{B_j}}{2}+\frac{1}{2}\right)}}{2^{\frac{1}{2}\left(n_{B_j}+24\right)} \pi^{\frac{n_{B_j}}{2}} \Gamma_4\left(3\right)} e^{-\frac{1}{2}\left[tr\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_j^{-1}\right) + \sum_{t \in B_j} (\mathbf{Y}_t - \boldsymbol{\theta}_{r_t}^{j})^t \boldsymbol{\Sigma}_j^{-1} (\mathbf{Y}_t - \boldsymbol{\theta}_{r_t}^{j})\right]} \quad (3.8)$$

We also define $\mathbf{Z}_t := (\mathbf{y}_t - \boldsymbol{\theta}_{r_t}^{j})$ and $S = \sum \mathbf{Z}_t^t \mathbf{Z}_t$ then

$$\textstyle\sum_{t\in B_j}(\mathbf{Y}_t-\boldsymbol{\theta}^j_{r_t})^t\boldsymbol{\Sigma}_j^{-1}(\mathbf{Y}_t-\boldsymbol{\theta}^j_{r_t})$$

so we have:

$$tr(\textstyle\sum_{i\in B_j}\mathbf{Z}_t^t\boldsymbol{\Sigma}_j^{-1}\mathbf{Z}_t).$$ [1]

We rewrite the posterior distribution in (3.8) as

$$\begin{aligned} p(\boldsymbol{\Sigma}_j \mid \xi \setminus \{\boldsymbol{\Sigma}_j \mid \mathbf{y}\}) \propto \mid \boldsymbol{\Sigma}_j \mid^{-\left(\frac{n_{B_j}}{2}+\frac{1}{2}\right)} e^{-\frac{1}{2}\left[tr\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_j^{-1}\right)+tr(\sum_{t\in B_j}\mathbf{Z}_t\mathbf{Z}_t^t\boldsymbol{\Sigma}_j^{-1})\right]} \\ = \mid \boldsymbol{\Sigma}_j \mid^{-\left(\frac{n_{B_j}}{2}+\frac{11}{2}\right)} e^{-\frac{1}{2}\left[tr(\boldsymbol{\Psi}+\mathbf{S})\boldsymbol{\Sigma}_j^{-1}\right]} \end{aligned}$$

We recognize the kernel of a Wishart distribution for $\boldsymbol{\Sigma}_j^{-1}$:

$$\boldsymbol{\Sigma}_j^{-1} \sim W\left[n_{B_j}+m,(\boldsymbol{\Psi}+S)^{-1}\right]$$

where we defined $\boldsymbol{\Psi}$ as the empirical covariance matrix, the subset $B_j := \{t: \ k_t = j\}$ and $n_{B_j} :=\mid B_j \mid$. Note that we introduced the statement $p(\boldsymbol{\Sigma}_j \mid \xi \setminus \{\boldsymbol{\Sigma}_j\})$ it means that we considered all the variables in the hyperparameter space without the variable of interest.

For the weights of the point masses we supposed the Beta distribution as prior, due to the stick-breaking representation (see more details in the introduction) with common weights. So the conditional posterior distribution is

$$P(V_h \mid \xi \setminus \{V_h\}) \propto Be(1+\mid Q_h \mid, M+\mid S_h \mid)$$

where $Q_h := \{t : r_t = h\}$ and $S_h := \{t : r_t > h\}$.

[1] We apply the property of matrices: $tr(ABC) = tr(BCA)$

The relevance of the two latent variables, r_t and k_t in the model, induced us to calculate them jointly, as follows:

$$P(r_t = h, k_t = j \mid \xi \setminus \{r_t, k_t\},\ \mathbf{y}) \propto w_h p_j N_4(\mathbf{y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t}\mathbf{y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda t^{\gamma}}, \boldsymbol{\Sigma}_j).$$

This distribution is the substantial difference with the parametric model: for each costant weight of the stick-breaking and for each selected label of the nucleotides *jointly* we have for the errors of the vector of the autoregressive lag one, a multivariate normal distribution centred on an exponential function and unknown covariance matrix.

Hence we marginalize the posterior distribution for each latent variable. In principle, for Fubini's theorem is completely irrelevant to marginalize first to respect r_t and then k_t, because analitically there is symmetry. Moreover, the motivation is different. The model has hierarchical order. We use the latent variable r_t across costant weights, w_h and the trajectories for the point masses $\boldsymbol{\theta}_h^j(\mathbf{y})$ and then we study the latent variable k_t for the mixture of the trajectories.

1. $p(r_t = h \mid \xi \setminus \{r_t, \mathbf{y}\}) \propto \omega_h^*$

   where $\omega_h^* = \omega_h \sum_{j=1}^{4} p_j N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_h^j \mathbf{Y}_{t-1} + \boldsymbol{\beta}_h^j e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_j)$

   or alternatively

   $p(r_t = h \mid \xi \setminus \{r_t,\ k_t,\ \mathbf{y}\}) \propto \omega_h N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{k_t} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{k_t} e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_{k_t})$

2. $p(k_t = j \mid \xi \setminus \{k_t,\ r_t,\ \mathbf{y}\}) \propto p_j N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_{r_t}^{j} \mathbf{Y}_{t-1} + \boldsymbol{\beta}_{r_t}^{j} e^{-\lambda i^{\gamma}}, \boldsymbol{\Sigma}_j)$

We will use step $1.$ and $2.$ to define a transition probability in the MCMC. In addition, we used these two steps to compare the true bases and the simulated bases. More details are explained in the next section.

To respect the hyperparameter space we have to calculate two more conditional posterior distributions: these are for the parameters of the trajectories for the point masses $\boldsymbol{\alpha}_h^j$ and $\boldsymbol{\beta}_h^j$.

$$p(\boldsymbol{\alpha}_h^j \mid \xi \setminus \{\boldsymbol{\alpha}_h^j \mid \mathbf{y}\}) \propto N_4(\boldsymbol{\alpha}_h^j \mid \mathbf{m}_{\boldsymbol{\alpha}}, \boldsymbol{\sigma}_{\boldsymbol{\alpha}}) \prod_{t \in R_{h_j}} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_h^j \mathbf{Y}_{t-1} + \boldsymbol{\beta}_h^j e^{-\lambda i^\gamma}, \boldsymbol{\Sigma}_j)$$

where $R_{h_j} = \{i : r_t = h, k_t = j\}$ and $n_{h_j} =\mid R_{h_j} \mid$ and we also define

$\mathbf{y}_t^* = [\mathbf{Y}_t - \boldsymbol{\beta}_h^j e^{-\lambda i^\gamma}]$

so we can rewrite the posterior distribution as

$$p(\boldsymbol{\alpha}_h^j \mid \xi \setminus \{\boldsymbol{\alpha}_h^j \mid \mathbf{y}\}) \propto N_4(\boldsymbol{\alpha}_h^j \mid \mathbf{m}_1, \mathbf{V}_1)$$

where $\mathbf{m}_1 = \mathbf{V}_1(\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}\mathbf{m}_{\boldsymbol{\alpha}} + \sum_{t \in R_{h_j}} \mathbf{T}_t^{-1}\tilde{\mathbf{Y}}_t)$ and $\mathbf{V}_1^{-1} = (\mathbf{S}_{\boldsymbol{\alpha}}^{-1} + \sum_{t \in R_{h_j}} \mathbf{T}_t^{-1})$.

The dimension of the vector of the means $\mathbf{m}_1$ is (4×1) and the covariance matrix $\mathbf{V}_1$ is (4×4). We introduced the matrix, $\mathbf{T}_t$, which is defined by $\mathbf{T_t} = \mathbf{L}_t \boldsymbol{\Sigma}_j \mathbf{L}_t$ where $\mathbf{L}_t$ is $diag\left(\frac{1}{y_{t-1,1}}, \frac{1}{y_{t-1,2}}, \frac{1}{y_{t-1,3}}, \frac{1}{y_{t-1,4}}\right)$ and $\tilde{\mathbf{y}}$ is $\mathbf{L}_t \mathbf{y}_t^*$.

The conditional posterior distribution for $\boldsymbol{\beta}_h^j$ has easier calculation than $\boldsymbol{\alpha}_h^j$ here there is the simply product between scalar values, $e^{-\lambda i^\gamma}$, and the vector of parameters.

$$p(\boldsymbol{\beta}_h^j \mid \xi \setminus \{\boldsymbol{\beta}_h^j \mid \mathbf{y}\}) \propto p(\boldsymbol{\beta}_h^j \mid \mathbf{m}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \prod_{t \in R_{h_j}} N_4(\mathbf{Y}_t \mid \boldsymbol{\alpha}_h^j \mathbf{Y}_{t-1} + \boldsymbol{\beta}_h^j e^{-\lambda i^\gamma}, \boldsymbol{\Sigma}_j).$$

Define $\mathbf{y}_t^{**} = [\mathbf{Y}_t - \boldsymbol{\alpha}_h^j \times \mathbf{Y}_{t-1}] e^{\lambda i^\gamma}$ to wit

$$p(\boldsymbol{\beta}_h^j \mid \xi \setminus \{\boldsymbol{\beta}_h^j \mid \mathbf{y}\}) \propto N_4(\boldsymbol{\beta}_h^j \mid \mathbf{m}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \prod_{t \in R_{h_j}} N_4(\mathbf{y}_t^{**} \mid \boldsymbol{\beta}_h^j, \mathbf{T_t}) = N_4(\boldsymbol{\beta}_h \mid \mathbf{m}_2, \mathbf{V}_2)$$

where $\mathbf{m}_2 = \mathbf{V}_2(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_j^{-1} \sum_{t \in Q_{h_j}} \frac{\mathbf{y}_t^{**}}{l_t^2})$ and $\mathbf{V}_2^{-1} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{\Sigma}_j^{-1} \sum_{t \in Q_{h_j}} \frac{1}{l_t^2})$ and $l_t = e^{\lambda i^\gamma}$. The dimension for the vector of the means, $\mathbf{m}_2$, is (4×1) and the covariance matrix $\mathbf{V}_2$ is (4×4).

3.5 Inference on the Posterior Distributions

The posterior distributions for each parameter are complicate to calculate analitically. We use the posterior distributions obtained in the last section to implement a blocked Gibbs sampler algorithm with two steps of random walk of Metropolis-Hastings algorithm for the parameters λ and γ, respectively. There are a lot of strategies to do inference.

In this step we do not show many plots about the choice of the initial values choosed for running the model, but we conduct as in the second Chapter diagnostic analysis, using CODA package in R software. The number of sticks, H, are 10, the number of labels k_t is fixed as 4. The base number i is equal to 10 for the concentration of the exponential function and the parameters λ and γ are respectively equal to 0.2 and 0.5, i. e. $e^{-\lambda i^{\gamma}} = 0.45$. We found also $\mathbf{m}_{\boldsymbol{\alpha}} = (0.30, 0.40, 0.30, 0.40)$, $\mathbf{m}_{\boldsymbol{\beta}} = (0.10, 2, 2.5, 5)$, $\mathbf{\Sigma}_{\boldsymbol{\alpha}} = diag(0.10, 1, 0.10, 1)$ and $\mathbf{\Sigma}_{\boldsymbol{\beta}} = diag(2.5, 5, 2.5, 5)$, respectively. The number of iterations is 100,000. It could be bigger for better results. The time of running each iteration is around 15 minutes.

One of our first purposes is to check if our model is really able to resolve the problems for this dataset. So we compute the elipses and we plug-in our posterior distributions. Then we put the elipses on the real dataset. In Figure 3.6, we show that the elipses are on the dataset and also they are separated which means that our nonparametric mixture is more efficient then the former mixture of multivariate normal distributions proposed by Ji *et al.* (2011). Furthermore, the triangles are the centroids of the posterior distributions plugged in the elipses.

An other interesting inferential aspect is the relevance of the latent variable k_t. It plays a central role to determine the elements of the mixture in the locations of point masses for the DDP prior model. A priori we assume exchangeability between the weights of the intensities of the mixtures for the trajectories for the point masses, it means that the probability of each label, k_t, is equal for $j = 1, 2, 3, 4$. We split on the

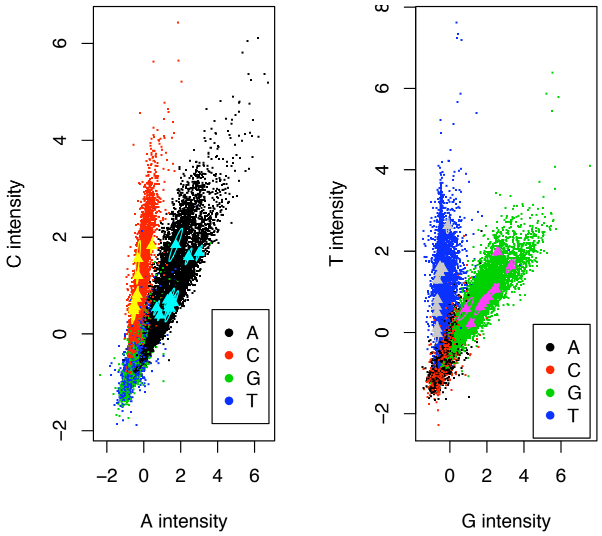

Figure 3.6: Yellow elipses represent the estimated DDP-AR(1) for A, light blue elipses are for C, gray's for G and pink's for T.

function for the posterior distribution of k_t in two parts. We use for the first half part only the simply probability p_j and for the second half part the posterior distribution for k_t which we obtained in the last section at point 2.

In Figure 3.7 we illustrate that a posteriori the elements of the mixtures on the trajectories for the point masses are not exchangeable. Further Figures 3.7 are simulated dataset, where the color of the observations depends from the simulated elements of the label k_t. Note that the position of the simulated observations is the same of the real observations and the color of the observations comes from the simulated posterior distribution of k_t. We conclude that the latent variable k_t is able to split the observations in the same number of groups as in the real dataset and also is able to assign the same bases as the true bases.

Finally, we calculate how much our simulated dataset is similar to the real dataset. We consider the last half part of the true bases which comes from Ji *et al.* (2011) paper and we compare it with our second half part of the matrix of the simulated bases. We have the last 17500 labels of our simulated dataset and 17500 labels of

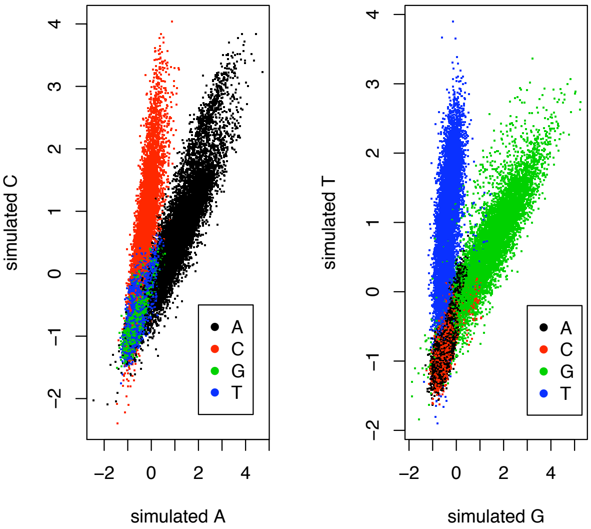

Figure 3.7: On the left hand side, the simulated dataset shows the same position of the clusters of the points as in the initial dataset. x-axes is the simulated intensity measurements for Adenine and y-axes is for Cythosine, respectively. On the right hand side we illustate the simulated dataset for Guanine and Thymine intensity measurements.

true bases. We count how many simulated labels correspond to the true bases. Our simulated labels are the posterior mode.The next table shows numerical results.

	k_t=1	k_t=2	k_t=3	k_t=4	**total**
right	26.78%	19.88%	23.42%	21.31%	**88.22%**
wrong	1.87%	1.37%	2.10%	6.43%	**11.78%**
total	**26.78%**	**19.88%**	**23.42%**	**29.92%**	**100%**

Table 3.1: **Comparison between simulated labels and true labels**

So we consider for a comparison the posterior mode for k_t. In the table above, we have 17500 labels; each single value represents how many right/wrong labels are assigned to the simulated dataset. In the next table, we count for each assigned label, how many times the assignment is right, i. e. $(k_t = 1 \mid k_t \setminus \{k_t = 1\})$. In Figure 3.8 we report the following results in the table as a barplot for a graphical comparison.

The two proposed tables differ for the denominator for a practical interpretation

	k_t=1	k_t=2	k_t=3	k_t=4
right	93.00%	93.10%	91.02%	78.51%
wrong	7%	6.89%	8.08%	21.49%
total	**100%**	**100%**	**100%**	**100%**

Table 3.2: **Comparison between simulated labels and true labels for each kind of label, given the others.**

of our obtained results. We had 88.22% of simulated labels out of the total labels. In the second table, we have that, given the right labels assignments for a nucleotide, we count how many times our simulated labels are rightly calculated. For example, in the first table, 88.22% is the percentage of the right number of times that we assigned the labels. In the second table, 27% of the total number of right assigned labels is the simulated label for Adenine nucleotide. In other words, given the total number of times that we assigned the right label to the nucleotide Adenine (which is about 27%), we want to know for the nucleotide Adenine how many times the simulated labels are the right labels. So, we have that for $k_t = 1$, (Adenine) 93.00% of right simulated labels, it means that in 93 cases of 100 we assigned right simulated labels to right simulated intensity measurements; for $k_t = 2$ (Cytosine) we obtained 93.10%, for $k_t = 3$, (Thymine) 91.02% and for $k_t = 4$, (Guanine) 78.51%.

To appreciate our results we insert Figure 3.8, which is a barplot, where the heights of the bars are dominated by the true bases so high percentage of similarities between the true bases and the simulated labels k_t. In Figure 3.8, we see only that for $k_t = 4$ the similarity between true bases and the simulated labels is less good then in all the other three cases.

Follow also an other straightforward interpretation. Suppose that we want to verify if there is a misleading when the probability of the true bases is equal to the probability of the simulated true bases versus these two probabilities are different. We reject the hypothesis of misleading and we conclude that we rightly assign our simulated labels to the real labels, see also Figure 3.9.

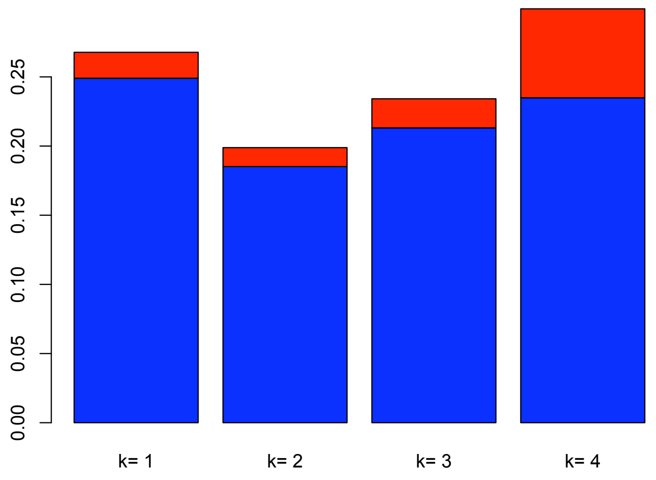

Figure 3.8: Blue color for the labels selected from the true bases dataset and red color for the labels selected from our simulated labels k_t.

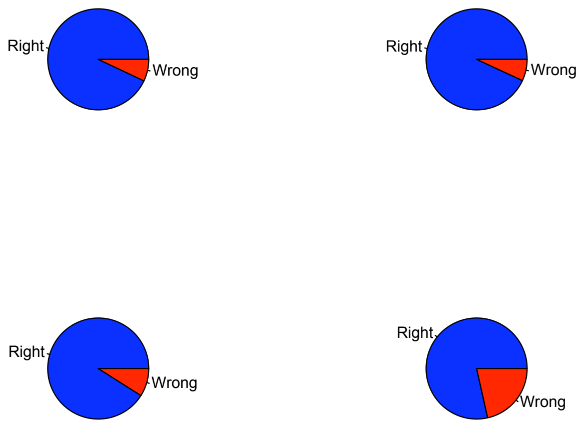

Figure 3.9: Pie representation of table 3.2. Right vs Wrong labels for each k_t. Blue color for the labels selected from the true bases dataset for each label given the others and red color for the labels selected from our simulated labels k_t.

3.6 Final Discussion

Our purpose is a steightforword proposal in a nonparametric framework to resolve the same problem of the DNA-sequencing dataset that Ji *et al.* (2011) studied in a Bayesian parametric context. However, one of our points of research was to improve the model adequacy for this dataset. The mixture of truncated Normal distributions was not able to represent the real dataset as well as we showed in Figure 3.2 in particular, for high Adenine intensity measurements. The advantage of our model is the flexibility of Bayesian nonparametric models and the capacity of fitting the dataset using the mixtures of random probabilities measures. The huge dataset had three different kind of problems that we modeled introducing a variation of the dependent Dirichlet processes. The dependent Dirichlet process is based on the random probability measures. We characterized the random probability measures such that the connected trajectories for the point masses resolved the fading and the phasing noises. The unknown covariance matrix of the errors associated to the autoregressive likelihood modeled the cross-talk between the channels. The latent variable k_t defined the number of random probability measures and r_t the number of finite dependent Dirichlet process we needed to fit the dataset. In the second Chapter we proposed a single-p DDP for autoregressive models and we concentrated on methodological aspects, in this chapter we applied the single-p DDP autoregressive model to a huge and real dataset. For the DNA-sequencing we considered a mixture of four multivariate autoregressive DDP model. The latent variable k_t played a central role: from the exploring through the inferential aspects, to the final results. The label of the true sequence at the beginning was explaining the high concentration of a specific nucleotide to respect the intensity measurement. In the modeling k_t was a random latent variable such that was characterizing the components of the mixture. For the inference was also relevant because we had the true bases and we did not have label switching problems tipical after an MCMC

algorithm and it was able to show the absence of exchangeability a posteriori.
In this chapter we reveled also a different role of the covariance matrix of the model, Σ_j. In the last chapter, the assumption of a random or fixed variance was not relevant. for the density estimation. In that case, we preferred models with fixed variance for the parsimony of the construction. Here we showed that the covariance matrix modeled the high correlations between the channels. A limit of our obtained results is the number of iterations runned in the MCMC algorithm. We gave a look to the convergences and to the diagnostics, but more accuracy for the details could be done. However, we used for this dataset the structure of the multivariate DDP-AR(1) similar to the construction of the previous chapter. In addition, we inserted an exponential function in the trajectories for the point masses.

Bibliography

Cokus, S. J., Feng, S., Zang, X., Zugen, C., Merriman, B. Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. and Jacobsen, S. E. (2008). *Nature*. **452**.

De Iorio, M., M$\ddot{u}$ller, P., Rosner, G., MacEachern, S. N., (2004).”An ANOVA Model for Dependent Random Measures”. *Journal of the American Statistical Association*, **99**, 205-215.

Dohm, J. C., Lottaz, C., Borodina, T., Himmelbauer, H. (2008). ”Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.” *Nucleic Acid Research.* **36**, 16.

Doss, H. (1994). ”Bayesian Nonparametric Estimation for Incomplete Data via Successive Substitution Sampling”. *Annals of Statistics*. **22** vol.4. 1763-1786.

Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. and Hannon, G. J. (2008). ”Alta-Cyclic: a self-optimizing bae caller for next-generation sequencing”. *Nature Methods.* **5**, 8.

Escobar, M. D., West, M. (1995). ”Bayesian density estimation and inference using mixtures”. *Journal of American Statistical Association*, **90**, 577-588.

Erwing, B. and Green, P. (1998). ”Base-Calling of Automated Sequencer Traces Using Phred. II Error Probabilities”. *Cold Spring Harbor Laboratory Press*.

Ferguson, T. S., (1973). ” A Bayesian analysis of some nonparametric problems”. *Annals of Statistics*, **1**, 209-230.

Hjort, N. L., Holmes, C., Müller, P., Walker, S. G. (2010). ”Bayesian Nonparametrics”. *Cambridge Series in Statistcal and Probabilistic Mathematics*.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). ”Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling”. *Statistical Science*. **20**, 1, 50-67.

Ji, Y., Mitra, R., Quintana, F., Müller, P., Jara, A. Liu, P., Lu, Y., Liang, S. (2011). ”BM-BC: A Bayesian method of base calling for Solexa sequence data”. (submitted paper).

Kahvejian, A. Quackenbush, J. and Thompson, J. F. (2008). ”What would you like do if you could sequence everything?” *Nature Biotechnology*. **26**, 10.

Kao, W., Stevens, K. and Song, Y. S. (2009). ”BayesCall: A model-based basecalling algorithm for high-throughput short-read sequencing”. *Cold SpringHarbor Laboratory Press.* **14**.

MacEachern, S.N., M$\ddot{u}$ller, P., (2000). ”Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichelt Process Mixture Models”. *Robust Bayesian Analysis. Lecture Notes in Statistics*. **152** 295-316.

Rougemont, J. Amzallag, A. Iseli, C. Farinelli, L., Xenarios, I., Naef, F. (2008). ”Probabilistic base calling of Solexa sequencing data”. *Biomedical Central Bioinformatics*. **9**.

Sethuraman J. (1994), ”A constructive definition of the Dirichlet process prior”, *Statistica Sinica*. **2**, 639-650.

Shendure, J. and Ji, H. (2008). ”Next-generation DNA-sequencing”. *Nature Biotechnology*. **26**, 10.

Chapter 4

A Bayesian Nonparametric Extention for Autoregressive Models

Abstract *We propose a Bayesian nonparametric autoregression for a sequence* $(Y_t, t \geq 1)$*. Autoregressive models of order* p *can be characterized as defining a family of random probability measures on a class* $\mathcal{F} = \{F_y,\ y \in \mathcal{Y}\}$ *of sampling models with* $Y_t \mid (Y_{t-1}, Y_{t-2}) = y \sim F_y$ *for any* $t \geq 3$*. The traditional linear AR(p) model restricts* $\mathcal{F}$ *to a family of normal linear models. We relax the parametric assumption while maintaining the characteristic AR(p) assumption. We restrict the model to lag 2, keeping in mind that the discussion remains valid almost without change also for* $p > 2$*. We define a prior probability model for* $\mathcal{F}$ *using a dependent Dirichlet process (DDP) model. Specifically, we use variable weights for* F_y *and define the point masses as a function of* $y = (Y_{t-1}, Y_{t-2})$*. Referring to the model as DDP-AR(2), we illustrate the model and posterior computation using the Annual Canadian Lynx trappings dataset.*

4.1 Introduction

In this chapter, we discuss an extension of nonparametric autoregressive models developed in the second and third Chapter in two important directions. We will extend the autoregressive mean function to higher order lagged terms to allow AR(p) structure, and we will move from single-p DDP models to variable weight DDP models.

As a motivating application we focus on the representation of harmonic models as an AR(2) model. We consider a linear autoregression, such that it depends jointly from the first two lags, $Y_t \mid (Y_{t-1}, Y_{t-2})$, therefore, conditional on Y_{t-1} and Y_{t-2}, we have $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$ for $t \geq 3$, where $\{\varepsilon_t\}$ is the usual sequence of Gaussian random variables. We assume a white noise process $\varepsilon_t \mid \sigma^2 \overset{iid}{\sim} N(0, \sigma^2)$, follows that $Y_t \mid \sigma^2 \sim N(\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}; \sigma^2)$. For further application on the annual Canadian lynx, we consider a simplified version of the AR(2) model (as Müller et al. 1997): $Y_t \mid (Y_{t-1}, Y_{t-2}) \sim N(Y_t \mid \beta Y_{t-1} - Y_{t-2}; \sigma^2)$. We indicate the labels y_1 for Y_{t-1} and y_2 for Y_{t-2} respectively and with y the two lags jointly, $(Y_{t-1} = y_1, Y_{t-2} = y_2)$. We present a Bayesian nonparametric autoregressive model such that $Y_t \mid y \sim F_y$ for any $t \geq 3$. Note that $Y_t \mid y$ does not change with t. This represent the first relevant difference with Griffin and Steel (2006; 2011), here implicitly we assume homogeneity over time. We define a family of random probability measures on the class $\mathcal{F} = \{F_y : y \in \mathcal{Y}\}$. On the class $\mathcal{F}$, we assume that the family of random probability measures is, marginally, a Dirichlet process (Ferguson, 1973). We indicate $G \sim DP(\alpha, G^0)$, where M is the total mass parameter and G^0 is the baseline distribution. A constructive representation of the DP is given by Sethuraman (1994), as we described in the equation (1.1). MacEachern (1999; 2000) used the stick-breaking representation reviewed in the equation (1.1) to propose a family of dependent Dirichlet processes (DDPs), i.e., a collection of random probability measures of the form $G_y = \sum_{h \geq 1} \omega_h(y) \delta_{\theta_h}(y)$ for $y \in \mathcal{Y}$ such that each random probability measure G_y is in some way marginally distributed as a DP, and with the property that G_y varies smootly with y. In the litterature there are a lot of variations of the DDP, for example Rodriguez and ter Horst (2008) proposed a DDP applied to the dynamic linear models. De Iorio *et al.*(2004) considered a single-p DDP for the ANOVA categorical covariates. We organized the contents of this chapter as follows.

In Section 4.2, we consider the single-p DDP prior and the autoregressive model with two lags jointly. In Section 4.3, we describe a possible different prior in a variation of the DDP with costant weights. In Section 4.4, we illustrate an example of the application for the annual Canadian lynx dataset for the two proposed models and we evaluate the model adequacy of these two models in Section 4.5. We conclude this chapter, in Section 4.6, with a discussion about these two proposals and further possible developments.

4.2 DDP-AR(2)

Autoregressive models of order p can be characterized with a generic nonparametric distribution F_y. We define a family $\mathcal{F} = \{F_y,\ y \in \mathcal{Y}\}$ of sampling models with a nonparametric distribution F_y and we indicate the label $y = (Y_{t-1} = y_1, Y_{t-2} = y_2)$ for $Y_t \mid y_1,\ y_2$. The traditional linear AR(p) model restricts $\mathcal{F}$ to a family of normal linear models. We relax the parametric assumption while maintaining the characteristic AR(p) assumption.

In this subsection we consider the dependent Dirichlet process with costant weights, it means that the collection of random probability measures can be represented as $G_y = \sum_{h\geq 1} \omega_h \delta_{\theta_h}(y)$ where the point masses $\theta_h(y)$ represent the atoms corresponding to the random distribution of Y_t. They are modeled as trajectories indexed by $y = (Y_{t-1}, Y_{t-2})$. Note that G^0 defines an harmonic autoregression $g(\cdot \mid Y_{t-1} = y_1, Y_{t-2} = y_2)$. Assume that the conditional distribution $Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_1$

depends only on Y_{t-1} and Y_{t-2} for any $t \geq 3$. Similarly, to the AR(1) that we proposed in the second Chapter, here we have for $t \geq 3$

$$Y_t \mid Y_{t-1} = y_1, Y_{t-2} = y_2 \sim \int_{\Re} N(Y_t \mid \beta y_1 - y_2, \sigma^2) dG(\beta),\ G \sim DP(M, G^0)$$

This is equivalent to assume

$$Y_t \mid Y_{t-1} = y_1, Y_{t-2} = y_2,\ m_t, \sigma^2 \sim N(Y_t \mid m_t, \sigma^2),\ m_t \sim F_y$$

where $y = (y_1, y_2)$ and $F_y = \sum_{h \geq 1} \omega_h \delta_{\theta_h(y)}$ with

$$\theta_h(y) = \beta y_1 - y_2,\ \beta_h \overset{iid}{\sim} G^0 = N(m_\beta, \sigma^2_\beta)$$

we assume a Gaussian distribution as prior for β_h such that mean, m_β and variance, σ^2_β, are known. The marginal prior is $G^0(\theta_h(y)) \sim N(y_{t-1} m_\beta - y_{t-2}; y^2_{t-1}\sigma^2_\beta)$. To compute the posterior distribution and implement a bloked Gibbs sampler we assume the finite Dirichlet process (Iswaran and Zarepur, 2002):

$$G_y(Y_t) = \sum_{h=1}^{H} \omega_h \delta_{\theta_h(y)}. \tag{4.1}$$

This is a simplified version of this model, with a finite mixture of H components as mixing measure $F_y = \sum_{h=1}^{H} \omega_h \delta_{\theta_h(y)}$, where the weights are costant and defined as in equation 1.1. For the implementation of the posterior inferences we find it convinient to use the equivalent hierarchical modeling

$$\begin{aligned} Y_t \mid Y_{t-1} = y_1, Y_{t-2} = y_2,\ r_t = h\ \sigma^2 \sim N(Y_t \mid \beta_h y_1 - y_2, \sigma^2) \\ p(r_t = h) = \omega_h \\ \beta_h \overset{iid}{\sim} N(m_\beta, \sigma^2_\beta) \times \sigma^{-2} \sim IG(a, b) \end{aligned}$$

Note that we introduced the Inverse-Gamma distribution for the parameter $1/\sigma^2$ with known shape and rate parameters, a and b, respectively.

4.2.1 Posterior Inference

The latent variable r_t is the link between costant weights of the stick-breaking representation (see in equation 1.1) and the trajectories for the point masses (4.2). The vector of hyperparameters for the single-p DDP-AR(2) is defined as follows:

$$\xi_1 = (V_1, V_2, ..., V_{H-1}, \beta_1, \beta_2, ..., \beta_{H-1}, r_1, r_2, ..., r_T, \sigma^2) \tag{4.2}$$

The joint posterior distribution is

$$p(Y, \xi_1) = \prod_{t=3}^{T} P(Y_t \mid \xi_1) \prod_{t=3}^{T} p(r_t \mid \xi_1 \setminus \{r_t\}) \prod_{h=1}^{H} p(V_h) p(\beta_h) p(\sigma^{-2}) =$$

$$p(\sigma^{-2}) \prod_{t=3}^{T} N(Y_t \mid \beta_{r_t} Y_{t-1} - Y_{t-2}, \sigma^2) \prod_{t=3}^{T} \omega_{r_t} \prod_{h=1}^{H} Be(V_h \mid 1, M) N(\beta_h \mid m_\beta, \sigma_\beta^2)$$

and we compute full posterior distributions for each parameter.
We assume exchangeability, therefore, the order of the posterior distributions in the Gibbs sampling is not relevant. We illustrate one of the possible orders for full posterior distributions.
This model has a similar structure of the DDP-AR(1), illustrated in second Chapter, with a prior on the variance for Y_t. The substantial difference is in the autoregressive representation of the trajectories. The parameter of the trajectories, β_h, which is random for the first lag and jointly to the second lag describes the harmonic trajectories.
Therefore, the posterior distribution for the parameter V_h has the same result, as in

the second Chapter, that here briefly, remind, and it is given by

$$p(V_h \mid \xi_1 \setminus \{V_h\}) \propto Be(1, M) \prod_{t \in S_h} (1 - V_h) \prod_{t \in Q_h} V_h =$$

$$Be(1+ \mid Q_h \mid, M+ \mid S_h \mid)$$

where $S_h = \{t : r_t > h\}$ and $Q_h = \{t : r_t = h\}$.

For the latent variable, r_t, we have

$$p(r_t = h \mid \xi_1 \setminus \{r_t\}) \propto \omega_h N(Y_t \mid \beta_h Y_{t-1} - Y_{t-2}, \sigma^2).$$

Keeping in mind that the autoregressive component has two lags, the posterior distribution is calculated on the parameter for the trajectories for the point masses, β_h, for $h = 1, 2, \ldots, H$. In principle, we have to consider

$$Y_t = \beta_h Y_{t-1} - Y_{t-2} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2).$$

and we define $Y_t^* = \frac{Y_t + Y_{t-2}}{Y_{t-1}}$, obtaining

$$Y_t^* = \beta_h + \epsilon_t \qquad \epsilon_t \sim N(0, W_t)$$

it means that

$$Y_t^* \mid \beta_h \sim N(\beta_h, W_t)$$

where $W_t = \frac{\sigma^2}{y_{t-1}^2}$. We also define the subset $Q_h = \{t : r_t = h\}$ for the link between the latent variable r_t and the trajectories for the point masses, β_h. We indicate the number of elements in this subset as $n_h =\mid Q_h \mid$.

$$p(\beta_h \mid \xi_1 \setminus \{\beta_h\}) \propto N(\beta_h \mid m_\beta, \sigma^2_\beta) \prod_{t \in Q_h} N(Y^*_t \mid \beta_h, W_t)$$

$$p(\beta_h \mid \xi_1 \setminus \{\beta_h\}) \propto N(\beta_h \mid m_1, V_1)$$

$$m_1 = V_1(\frac{m_\beta}{\sigma^2_\beta} + \sum_{t \in Q_h} \frac{y^*_t}{W_t}) \;\; V_1^{-1} = (\sigma_\beta^{-2} + \sum_{t \in Q_h} \frac{\sigma^{-2}}{W_t})$$

The last posterior distribution is for the variance of the model Y_t, σ^2, which is very similar to the result obtained in the second Chapter for the DDP-AR(1)

$$p(\sigma^2 \mid \xi_1 \setminus \{\sigma^2\}) \propto p(\sigma^{-2}) \prod_{t=3}^{T} N(Y_t \mid \beta_{r_t} Y_{t-1} - Y_{t-2}, \sigma^2) =$$

$$\frac{b^a}{\Gamma(a)} \sigma^{-2(a-1)} e^{-b\sigma^{-2}} (\frac{1}{\sqrt(2\pi)})^T (\sigma^{-2})^{\frac{T}{2}} e^{\{-(\frac{1}{2})\sigma^{-2} \sum_{t=1}^{T} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2\}}$$

$$\frac{b^a}{\Gamma(a)} \sigma^{-2[(a+\frac{1}{2}T)-1]} (\frac{1}{\sqrt(2\pi)})^T e^{\{-\sigma^{-2}(b+\frac{1}{2}) \sum_{t=1}^{T} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2\}}$$

Therefore, $\sigma^2 \sim IG(A, B)$ with $A = (a + \frac{T}{2})$ and $B = [b + \frac{1}{2} \sum_{t=1}^{T} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2] = \{b + \frac{1}{2} \sum_{h=1}^{H} [\sum_{t \in Q_h} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2]\}$.

4.3 Varying Weights DDP-AR(2)

Introduce dependence across trajectories for the point masses is not the only one possible proposal. An other possible point of research is the dependence across variable weights and fixed trajectories, but this case is not fitting for the application to the annual Canadian lynx trappings. However, Nieto-Barajas *et al.* (2011) suggested to use varying weights for autoregressive models, because this structure is more flexible on the data.

Single-p DDP prior is not able to describe extreme values present in lynx dataset. More details will be shown in the application. Here we study *varying* weights dependent Dirichlet process for autoregressive model with two lags. For further comparisons between the model proposed in equation (4.2) and the following model, we use an AR(2), but we should consider an AR(p) for $p > 2$.

Let AR(2) be the form of the trajectories for the point masses as in the previous proposed model in equation (4.2)

$Y_t = \beta Y_{t-1} - Y_{t-2} + \epsilon_t$ for $t = 3, 4, \ldots, T$ where $\epsilon_t \sim N(0, \sigma^2)$

$$Y_t \mid Y_{t-1}, Y_{t-2} \sim N(\beta Y_{t-1} - Y_{t-2}, \sigma^2)$$

Let Y_t be a continous random variable. We define a DP prior for the mixing distribution on $\Re$,

$$Y_t \mid Y_{t-1} = y_1, Y_{t-2} = y_2, G_{y=y_1,y_2} \sim f_{y_1,y_2}(Y_t) \tag{4.3}$$

and $f_y(Y_t)$ is the density distribution of the trajectories for the point masses and we write it in terms of DP mixtures

$$f_{y_1,y_2}(Y_t) = \int N(Y_t \mid \beta y_1 - y_2, \sigma^2) dG_y(\beta), G_y \sim DP(M, G_y^0) \tag{4.4}$$

In the DDP-AR(2) with common weights, we had $\omega_h = V_h \prod_{i=1}^{h-1}(1 - V_i)$ with $V_h \sim Be(1, M)$. In particular, the weights ω_h are invariable across $y = (y_1, y_2)$. Here we replace the costant weights ω_h by *varying weights*. We assume variable weights $\omega_h(y_1, y_2)$. We define the variable weights in a variation of the basic single-p DDP model. We introduce two changes. First, we approximate the infinite sum of point masses by a finite discrete model (Ishwaran and Zarepour, 2002; Hjort, 2000). Secondly, we abandon the stick-breaking prior. This is a variation of the DDP prior model that MacEachern (1999) defined, but it is not often applied. The reason is that in a lot of applications it is easier for computations to introduce costant weights. However, for the application on the annual lynx trappings, the single-p DDP prior with autoregressive lag 2 is not sufficient to describe this dataset as we will show in the next Section. Indeed, we allow to the weights ω_h to vary with the first two lagged terms (y_1, y_2) jointly, in other words we have a process (MacEachern, 1999) for the weights and a process for the athoms and both of them depending simultaneously from the lagged terms. The core of this chapter is the application of a logistic function, which describes varying weights of the DDP and in the same time we use an autoregressive model with two lags instead of one lag. The variable weights DDP model yields more elicitations. Yet, the functional form of the variable weights is coming from the utility function $U_h = \alpha_{0h} + \alpha_{1h} Y_{t-1} + \alpha_{2h} Y_{t-2} + \gamma_h$ such that $p(U_h > U_{\bar{h}})$, for $h \neq \bar{h}$. Our model will be more close to a parametric solution: we define more prior distributions reducing the flexibility of the initial model that we illustrated in equation 4.2. In addition, note that for the following model we consider the variance of the likelihood, σ^2, fixed, we will show in the Section 4.5.1 that warying weights are sufficient for Y_t's dispersion. The model is

$$Y_t \mid Y_{t-1} = y_1, Y_{t-2} = y_2, r_t = h \sim N(Y_t \mid \beta_h y_1 - y_2, \sigma^2)$$

$$\omega_h(y) = p(r_t = h \mid y_1, y_2) = \frac{e^{\alpha_{0h} + \alpha_{1h} y_{t-1} + \alpha_{2h} y_{t-2}}}{\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}} + \alpha_{1\bar{h}} y_{t-1} + \alpha_{2\bar{h}} y_{t-2}}}$$

$$\beta_h \overset{iid}{\sim} N(m_\beta, \sigma^2_\beta) \times \alpha_{0h} \overset{iid}{\sim} N(m_{\alpha_0}, \sigma^2_{\alpha_0}) \times \alpha_{1h} \overset{iid}{\sim} N(m_{\alpha_1}, \sigma^2_{\alpha_1}) \times \alpha_{2h} \overset{iid}{\sim} N(m_{\alpha_2}, \sigma^2_{\alpha_2})$$

for $h = 1, 2, \ldots, H$ and $t = 3, 4, \ldots, T.$

The prior of the base measure is still the same, $\beta_h \overset{iid}{\sim} N(m_\beta, \sigma^2_\beta)$ which implies a marginal prior for the trajectories $\theta_h(y_1, y_2) \sim N(Y_t \mid m_\beta y_1 - y_2; y_1^2 \sigma^2_\beta)$. Prior mean and variance for α_{0h}, α_{1h} and α_{2h} are known. For this proposal (in equation 4.5) we illustrate how the hyperparameter space is changed.

$$\xi_2 = (r_3, r_4, ..., r_T, \beta_1, \beta_2, \ldots, \beta_H, \alpha_{10}, \alpha_{20}, \ldots, \alpha_{H0}, \alpha_{11}, \alpha_{21}, \ldots, \alpha_{H1}, \alpha_{12}, \alpha_{22}, \ldots, \alpha_{H2}) \tag{4.6}$$

Note that the vector of parameters in equation (4.6) has the same parameters of the DDP-AR(2) with costant weights (4.2), but in this case we consider, σ^2 fixed (this is Y_t's variance). For the hyperparameter in equation (4.6), the joint posterior distribution is

$$p(y, \xi_2) = \prod_{h=1}^{H} p(\beta_h \mid m_\beta, \sigma^2_\beta) p(\alpha_{h0} \mid m_{\alpha_0}, \sigma^2_{\alpha 0}) p(\alpha_{h1} \mid m_{\alpha_1}, \sigma^2_{\alpha_1}) p(\alpha_{h2} \mid m_{\alpha_2}, \sigma^2_{\alpha_2})$$
$$\prod_{t=3}^{T} p(Y_t \mid \beta_{r_t} Y_{t-1} - Y_{t-2}, \sigma^2) p(r_t \mid \xi \setminus \{r_t\}) =$$
$$\prod_{h=1}^{H} N(\beta_h \mid m_\beta, \sigma^2_\beta) N(\alpha_{h0} \mid m_{\alpha_0}, \sigma^2_{\alpha 0}) N(\alpha_{h1} \mid m_{\alpha_1}, \sigma^2_{\alpha_1}) N(\alpha_{h2} \mid m_{\alpha_2}, \sigma^2_{\alpha_2})$$
$$\prod_{t=3}^{T} N(Y_t \mid \beta_{r_t} y_1 - y_2, \sigma^2) \omega_{r_t}(y_1, y_2)$$

Observe that even if there are many normal distributions, for the joint posterior distribution it is not possible to apply conjugacy property and also this is not a close form, so we can conclude that it is not possible to proceed analitically. Specifically, it is also not easy the calculation for the product of normals, but the real difficult part is in the last row: the product of the likelihoods of the parameters on the subset of r_t and the effect of the covariates on the latent variables. For this model we also assume excheangebility and consequently we marginalize the joint posterior distribution to respect each of the involved parameters and implement a bloched Gibbs sampler algorithm such that the empirical results will be close to the theoretical conditional posterior distributions.

4.3.1 Posterior Distributions for Varying Weights DDP-AR(2) Model

We use the new hyperparameter space of the equation (4.6), where the relevant changes are for the role of the latent variable, r_t. The conditional posterior distribution for the parameter of trajectories, β_h for $h = 1, 2, \ldots, H$ has the same conditional posterior distribution of the trajectories for the costant weights (4.3).

The full posterior distribution for the next three parameters represents the realization of the novel model. Let α_{h0} be for $h = 1, 2, \ldots, H$ the first parameter of interest for warying weights.
We consider $e^{\alpha_{0h}+\alpha_{1h}y_{t-1}+\alpha_{2h}y_{t-2}}$ which is the logistic model and the posterior distribution for the intercept is

$$p(\alpha_{h0} \mid \xi \setminus \{\alpha_{h0}\}) \propto N(\alpha_{h0} \mid m_{\alpha_0}, \sigma^2_{\alpha 0}) \prod_{t=3}^{T} \frac{e^{\alpha_{0r_t}+\alpha_{1r_t}y_{t-1}+\alpha_{2r_t}y_{t-2}}}{\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}y_{t-1}+\alpha_{2\bar{h}}y_{t-2}}}$$

Unfortunally, the parameter, α_{0h}, does not have a conjugate posterior distribution, even if the prior is a normal distribution and we supposed to use the conjugacy property. We need to develop for it a Metropolis-Hastings step, which will be in the Gibbs sampler algorithm.

$$p(\alpha_{h0} \mid \xi \setminus \{\alpha_{h0}\}) \propto N(\alpha_{h0} \mid m_{\alpha_0}, \sigma^2_{\alpha 0}) \prod_{t \in Q_h} \frac{e^{\alpha_{0h}+\alpha_{1h}Y_{t-1}+\alpha_{2h}Y_{t-2}}}{\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}}$$
$$\prod_{t \in Q_h^c} \frac{e^{\alpha_{0r_t}+\alpha_{1r_t}Y_{t-1}+\alpha_{2r_t}Y_{t-2}}}{\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}}$$

where we define $Q_h = \{r_t = h\}$ and $Q_h^c = \{r_t \neq h\}$. To simplify some computations we introduce the logarithm which is a monoton function. The posterior distribution, for the Metropolis-Hastings step, is

$$p(\alpha_{h0} \mid \xi \setminus \{\alpha_{h0}\}) \propto logN(\alpha_{h0} \mid m_{\alpha_0}, \sigma^2_{\alpha 0}) + \sum_{t \in Q_h} (\alpha_{0h} + \alpha_{1h}Y_{t-1} + \alpha_{2h}Y_{t-2}) +$$
$$-log(\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}) +$$
$$+ \sum_{t \in Q_h^c} (\alpha_{0r_t} + \alpha_{1r_t}Y_{t-1} + \alpha_{2r_t}Y_{t-2}) - log(\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}})$$

Let α_{h1} be the slope of the logistic function for $h = 1, 2, \dots, H$ we have a similar conditional posterior distribution, where the relevant change is on the prior,

$$p(\alpha_{h1} \mid \xi \setminus \{\alpha_{h1}\}) \propto N(\alpha_{h1} \mid m_{\alpha_1}, \sigma^2_{\alpha_1}) \prod_{t=3}^{T} \frac{e^{\alpha_{0r_t}+\alpha_{1r_t}Y_{t-1}+\alpha_{2r_t}Y_{t-2}}}{\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}}$$

Like in equation (4.7) we use the logarithmic transformation

$$\begin{aligned} p(\alpha_{h1} \mid \xi \setminus \{\alpha_{h1}\}) \propto log N(\alpha_{h1} \mid m_{\alpha_1}, \sigma^2_{\alpha_1}) + \sum_{t \in Q_h} (\alpha_{0h} + \alpha_{1h} Y_{t-1} + \alpha_{2h} Y_{t-2}) + \\ - log(\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}) + \\ + \sum_{t \in Q_h^c} (\alpha_{0r_t} + \alpha_{1r_t} Y_{t-1} + \alpha_{2r_t} Y_{t-2}) - log(\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}) \end{aligned}$$

For α_{h2} for $h = 1, 2, \dots, H$ we have

$$p(\alpha_{h2} \mid \xi \setminus \{\alpha_{h2}\}) \propto N(\alpha_{h2} \mid m_{\alpha_2}, \sigma^2_{\alpha_2}) \prod_{t=3}^{T} \frac{e^{\alpha_{0r_t}+\alpha_{1r_t}Y_{t-1}+\alpha_{2r_t}Y_{t-2}}}{\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}}$$

$$\begin{aligned} p(\alpha_{h2} \mid \xi \setminus \{\alpha_{h2}\}) \propto log N(\alpha_{h2} \mid m_{\alpha_2}, \sigma^2_{\alpha_2}) + \sum_{t \in Q_h} (\alpha_{0h} + \alpha_{1h} Y_{t-1} + \alpha_{2h} Y_{t-2}) + \\ - log(\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}) + \\ + \sum_{t \in Q_h^c} (\alpha_{0r_t} + \alpha_{1r_t} Y_{t-1} + \alpha_{2r_t} Y_{t-2}) - log(\sum_{\bar{h}=1}^{H} e^{\alpha_{0\bar{h}}+\alpha_{1\bar{h}}Y_{t-1}+\alpha_{2\bar{h}}Y_{t-2}}) \end{aligned}$$

The conditional posterior distribution for the latent variable r_t has the essential variation on the weights, which comes from the product of the logistic function and the trajectories for the point masses.

$p(r_t = h \mid \xi \setminus \{r_t\}) \propto \omega^*_h(y)$ for t = 3, 4, . . . ,T and for h = 1, 2, . . . ,H

where

$$\omega_h^*(y_1, y_2) = \omega_h(y) N(Y_t \mid \beta_h y_1 - y_2, \sigma^2)$$

4.3.1.1 A variation for varying weights DDP- AR(2) model

We discussed in the second Chapter the relevance of the random variance of the likelihood, σ^2. We illustrated that it is not always relevant, in particular, when there are small samples. Indeed, in the third Chapter, we showed the use of a prior on the variance for better modeling the correlation (cross-talk between channels). between the nucleotides. Here, we refer just as completeness the posterior distribution for random variance, σ^2. We add for the model in equation (4.5) the prior distribution for σ^2 as an Inverse Gamma, $IG(a, b)$, where the slope and rate are known a and b respectively. The first consequence of this hypothesis is on the hyperparameter space that we studied above in equation (4.6)

$$\xi_3 = (r_1, r_2, \ldots, r_T, \beta_1, \beta_2, \ldots, \beta_H, \alpha_{10}, \alpha_{20}, \ldots, \alpha_{H0},$$
$$\alpha_{11}, \alpha_{21}, \ldots, \alpha_{H1}, \alpha_{12}, \alpha_{22}, \ldots, \alpha_{H2}, \sigma^2)$$

and the joint posterior distribution is

$$p(Y, \xi_3) = \prod_{h=1}^{H} p(\sigma^{-2}) p(\beta_h \mid m_\beta, \sigma_\beta^2) p(\alpha_{h0} \mid m_{\alpha_0}, \sigma_{\alpha_0}^2) p(\alpha_{h1} \mid m_{\alpha_1}, \sigma_{\alpha_1}^2) p(\alpha_{h2} \mid m_{\alpha_2}, \sigma_{\alpha_2}^2)$$
$$\prod_{t=3}^{T} p(Y_t \mid \beta_{r_t} Y_{t-1} - Y_{t-2}, \sigma^2) p(r_t \mid \xi \setminus \{r_t\}) =$$
$$p(\sigma^{-2}) \prod_{h=1}^{H} N(\beta_h \mid m_\beta, \sigma_\beta^2) N(\alpha_{h0} \mid m_{\alpha_0}, \sigma_{\alpha_0}^2) N(\alpha_{h1} \mid m_{\alpha_1}, \sigma_{\alpha_1}^2) N(\alpha_{h2} \mid m_{\alpha_2}, \sigma_{\alpha_2}^2)$$
$$\prod_{t=3}^{T} N(Y_t \mid \beta_{r_t} Y_{t-1} - Y_{t-2}, \sigma^2) \omega_{r_t}(Y_{t-1}, Y_{t-2})$$

The conditional posterior distribution is

$$p(\sigma^2 \mid \xi_3 \setminus \{\sigma^2\}) \propto p(\sigma^{-2}) \prod_{t=3}^{T} N(Y_t \mid \beta_{r_t} Y_{t-1} - Y_{t-2}, \sigma^2) =$$

$$\frac{b^a}{\Gamma(a)} \sigma^{-2(a-1)} e^{-b\sigma^{-2}} (\frac{1}{\sqrt(2\pi)})^T (\sigma^{-2})^{\frac{T}{2}} e^{\{-(\frac{1}{2})\sigma^{-2} \sum_{t=1}^{T} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2\}}$$

$$\frac{b^a}{\Gamma(a)} \sigma^{-2[(a+\frac{1}{2}T)-1]} (\frac{1}{\sqrt(2\pi)})^T e^{\{-\sigma^{-2}(b+\frac{1}{2}) \sum_{t=1}^{T} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2\}}$$

Than we have $\sigma^2 \sim IG(A, B)$ with $A = (a + \frac{T}{2})$ and $B = [b + \frac{1}{2} \sum_{t=1}^{T} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2] = \{b + \frac{1}{2} \sum_{h=1}^{H} [\sum_{t \in Q_h} (Y_t - \beta_{r_t} Y_{t-1} + Y_{t-2})^2]\}$ and precision parameter is $1/\sigma^2$.

4.4 Example

In this Section we apply the two novel theoretical models (described in Section 4.2 and 4.3) to the annual number of lynx trapped in the Mackenzie River District of North-West Canada for the year 1821 through 1934 (both years are inclusive, giving a total of 144 observations). Several authors used this dataset. Some of them studied the lynx trapping in a different interval, i.e., the observations were for the years 1749-1924. On this former dataset, Yule (1927), for example, introduced on the initial dataset an autoregressive model with two lags, but the smoothing was not satisfying. Bartlett (1966) and than Anderson (1971) improved the smoothing of the model with a simple transformation. Moran (1953) worked on the dataset in interval 1821-1934. This author is the first one who thought about an initial logarithm trasformation on the dataset. Campbell and Walker (1977) noted that an initial logarithm transformation on the data is able to give a stationary process close to the Gaussian process and they added a sinusoidal trasformation of the autoregressive process with two lags for goodness of fit of the real observations. They applied Akaike Information Criterion (AIC) in a classical statistical approach and the

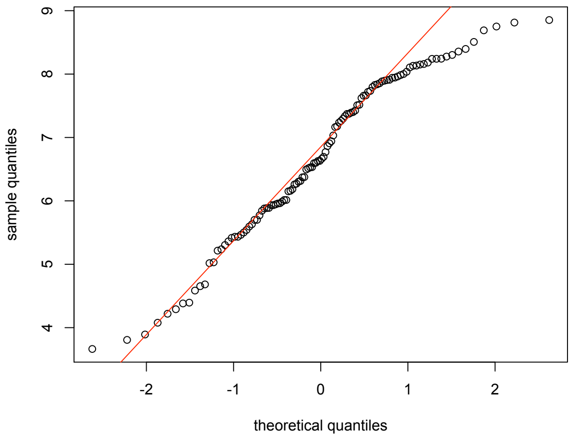

Figure 4.1: Presence of the trend for the logarithm transformation of the initial time series, using the quantile-quantile representation.

penalized maximum likelihood was preferable with two lags instead of one lag as the mentioned authors did before. We will use this dataset reported in Splus default uploaded by Becker and Chamber (1988) and then available in R software. Müller, West, MacEachern (1997) used the version of the dataset implemented in R software for a Bayesian semiparametric model, considering an autoregressive model with two lagged terms, assuming a prior only on the first lagged term.

We consider the presence of the trend in Figure 4.1, as well as Campbell and Walker (1977) observed. To remove the influence of the trend, we use the loss function and in Figure 4.2 we show the detrended series.

In Figure 4.3, we illustrate the time series that we consider for the applications. Observing the plot, we note the stationary of the series, but the presence of peaks. For the DDP-AR(2) with costant weights we propose the posterior distribution of the trajectories for the point masses β_h in a simple transformation $\lambda_h = 2\pi/acos(\beta_h/2)$ for $h = 1, 2, \ldots H$. In particular, we plot λ_h versus the initial time series splitted in two subsets: $y_{t-2} < y_{t-1}$ and $y_{t-2} > y_{t-1}$. We note two distinguished histograms.

In Figure 4.4, we ideally compare our posterior distribution considering the single-

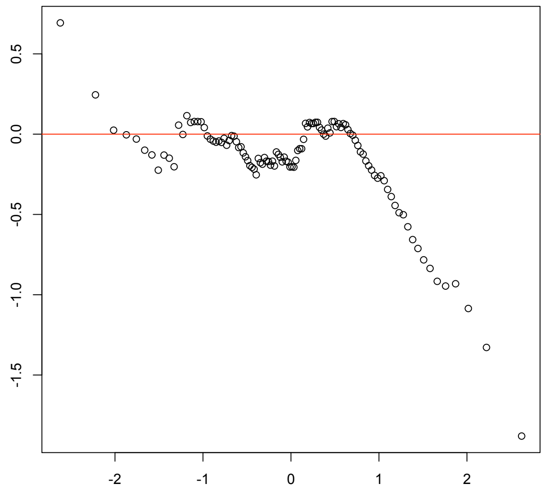

Figure 4.2: Extreme values of the trend on the logarithm transformation of the time series.

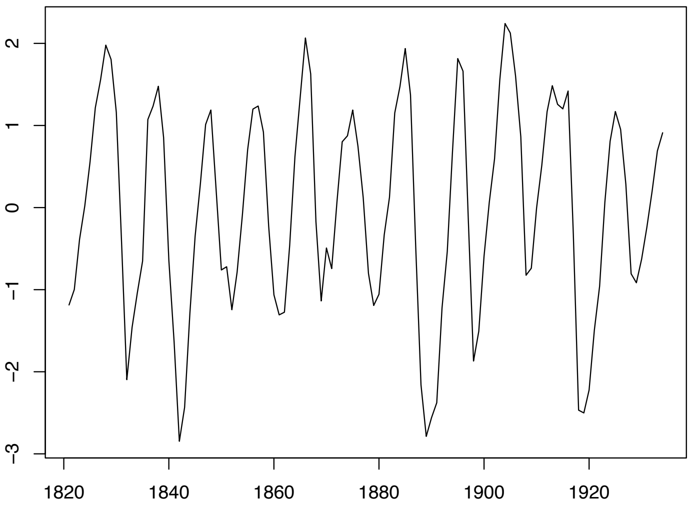

Figure 4.3: Annual Canadian lynx trappings dataset (in the logarithm scale) in the period 1821-1934. The time series was detrended.

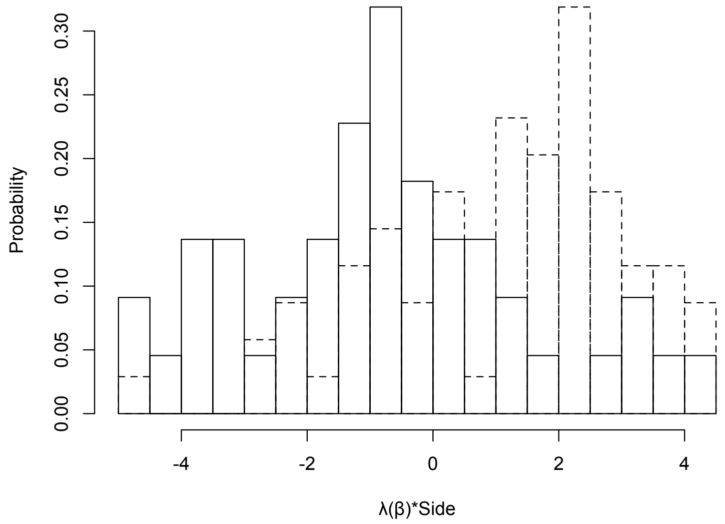

Figure 4.4: Bayesian posterior distribution for the parameter $\lambda = 2\pi/acos(\beta/2)$ assuming a single-p DDP prior model. The histogram is separated for points on the falling side (i.e. $y_{t-2} > y_{t-1}$) (solid line) and points on the rising side (i.e. $y_{t-2} < y_{t-1}$) (dashed line).

p DDP-AR(2) model with posterior computations in West *et al.* (1997). The initial log scaled dataset has extreme values and internal division. There is asymmetry in the dataset and also in our posterior distribution: the time spent on the rising side (i.e. from trough to peak) is longer than the falling side (i.e. from peak to through). We conclude that the internal division of the dataset is still present.

To appreciate the relevance of the logistic regression function introduced on the weights we show the posterior distribution for the variable weights $\alpha_{0h}, \alpha_{1h}, \alpha_{h2}$ for $h = 1, \ldots, 4$ over the two subsets $y_{t-1} < y_{t-2}$ and $y_{t-1} > y_{t-2}$. The main is the different direction of the posterior distributions and the exponential function for the variable weights.

In Figure 4.5, we illustrate posterior distributions for varying weights observing that varying weights DDP-AR(2) model is not affected by the presence of extreme values and by the falling side and the rising side.

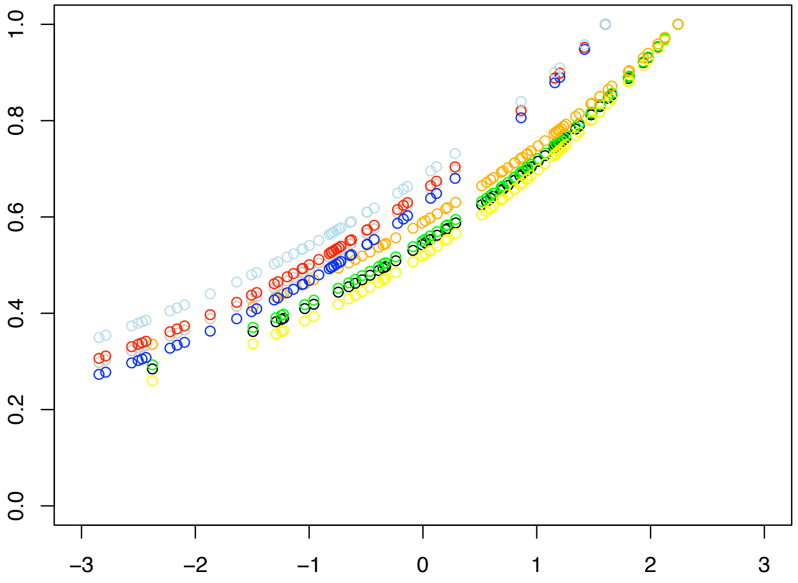

Figure 4.5: We represent lynx trappings dataset in two subsets: for $y_{t-1} < y_{t-2}$ and $y_{t-1} > y_{t-2}$ vs four posterior distributions of variable weights for the DDP-AR(2) varying weights.

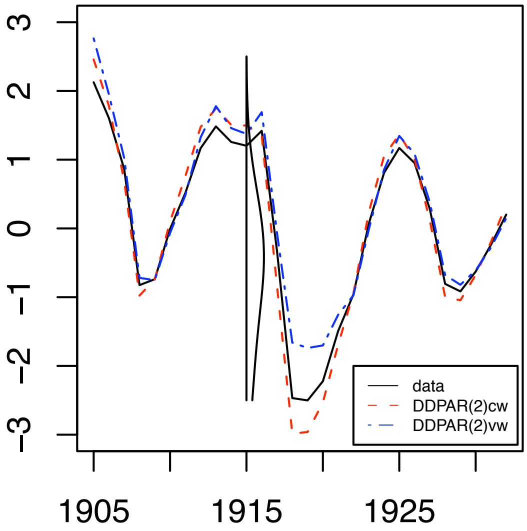

Figure 4.6: This is $\hat{y}_t = E(y_t \mid y_1, \ldots, y_{t-1})$ and the predictive distribution $p(y_{1918} \mid y_{1820}, \ldots y_{1911})$ at time 1918.

As goodness of fit of the Bayesian estimates in presence of outliers, we consider the last thirty years of the dataset. In particular, we analyze the period 1904-1934. We compare the time series with costant and varying weights DDP-AR(2) model estimates and we plot on these the predictive distribution function at 1918, see Figure 4.6.

Intuitively, in Figure 4.6, we choose the performance of varying weights DDP-AR(2) model. However, we extend our previous interesting result to all the time series using our simulated posterior distributions coming from costant weights and varying weights DDP-AR(2) (Figure 4.7). Note that for each year, the simulation of our model with variable weights holds better on the real dataset and it is less affected by the presence of the real peaks.

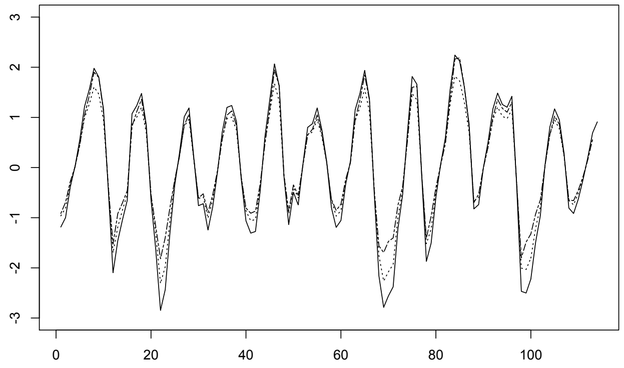

Figure 4.7: Comparisons between models. The solid line is for the observed time series the dotdashed line is for the posterior distributions with common weights and the dashed line is for our proposed model with the posterior distributions with variable weights.One step ahead forecasts for $\hat{y}_t$ is $\hat{y}_t = E(y_t \mid y_1, \ldots, y_{t-1}$and the predictive distribution $p(y_{1918} \mid y_{1820}, \ldots y_{1917})$ at time 1918.

4.5 Model adequacy

In this section, we apply model selection proceedures and we evaluete the adequacy of the choosed model. So, we calculated the Bayes factor for our two proposals. The Bayes factor is the ratio between the posterior distributions of two models (Kass and Raftery, 1995) . Specifically we have:

$$B_{12} = \frac{P(D \mid H_1)}{P(D \mid H_2)} = \frac{\int P(D \mid \xi_1, H_1)\pi(\xi_1 \mid H_1)d(\xi_1)}{\int P(D \mid \xi_2, H_2)\pi(\xi_2 \mid M_2)d(\xi_2)} \tag{4.7}$$

where H_1 and H_2 are DDP-AR(2) common weights and DDP-AR(2) varying weights, respectively; ξ_1 and ξ_2 are the two hyperparameter spaces, defined in equation 4.2 and 4.6, respectively. The Bayes factor is equal to 1.778664, which is in the interval between 1 and 3 and we conclude that it is better our DDP-AR(2) with varying weights. However, the Bayes factor is not sufficiently robust for two models with a so different number of parameters. For this reason, we use the conditional ordinated

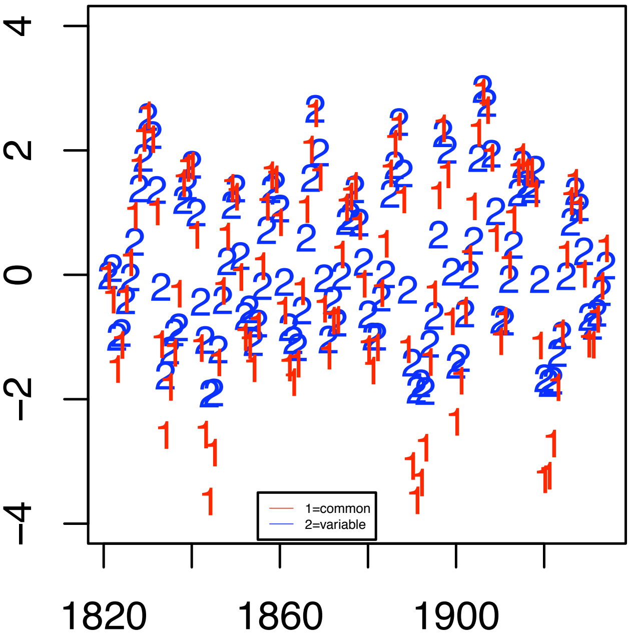

Figure 4.8: Number 1 (red color) indicates the points drawn from a single-p DDP-AR(2) model. Number 2 is for the points coming from the second model.

prediction (CPO) introduced by Gelfand *et al.* (1992). The CPO is a Bayesian diagnostic able to detect outliers. In addition, CPO can be used for model selection in the following construction:

$$p(M_j \mid Y) = \frac{f(Y \mid M_j)\omega_j}{\sum_{j=1}^{J} f(Y \mid M_j)\omega_j} \text{ for j= 1, 2, \dots, J}$$

where M_j $f(Y \mid M_j)$ is the predictive or marginal distribution of the data under model M_j. For observed y, the model, yelding the largest $p(M_j \mid y)$, is selected. In Figure 4.8, we propose the CPO plot and for each simulated point of the series, varying weights DDP-AR(2) has higher values, it means that varying weights DDPAR(2) model is preferable to costant weights DDP-AR(2) model. We analyze each possible simulated dataset drawn from our model with varying weights does

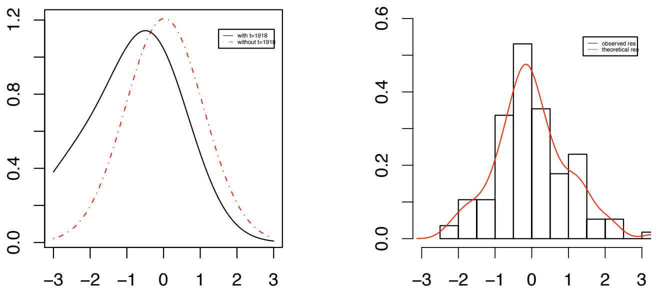

Figure 4.9: On the left hand side, the black color is for the Bayesian predictive density using variable weights model in the point $t = 1918$ and the red color is the simulated density without the same point. On the right hand side, we compare the empirical distribution of the residuals and the theorical distribution.

not contain influent observation for the year 1918. We also conducted analysis of the residuals using the CPO results. This method is robust to possible presence of outliers. In Figure 4.9, we illustrate two different aspects of the residuals. We compare Bayesian density estimates obtained by the simulation of the DDPAR(2) varying weights considering respectively with and without the year 1918. Then we observe the influence of the 1918 comparing a theoretical distribution and the distribution of the residuals without 1918. Each possible simulated dataset drawn from varying weights DDP-AR(2) does not contain the influent observation at $t = 1918$. Gelfand *et al.* (1992) suggest to interpret the residuals using the CPO from the posterior distribution of a model. So we computed the residuals on our model without the influent observation at 1918.

Through the comparison between the histogram of the residuals and the prdictive distribution for the year 1918, we can conclude that posterior residuals look like studentized residuals, which is a reasonable result, due to the CPO theory.

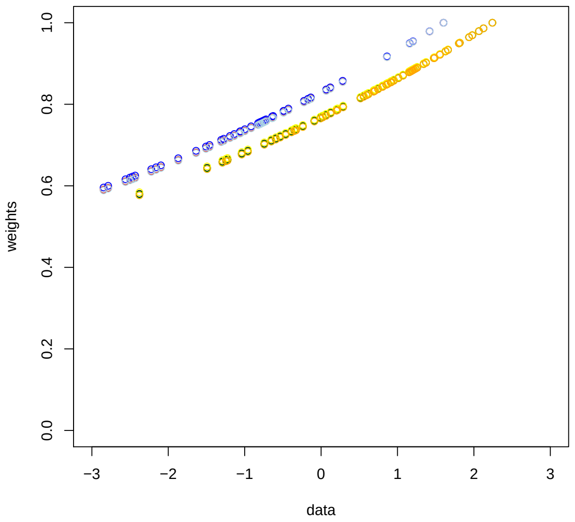

Figure 4.10: Posterior distribution for α_{0h}, α_{1h} and α_{2h} for $h = 1, \ldots 4$ on the falling and rising side of y_t.

4.5.1 The influence of a modified DDP-AR(2) Varying Weights

We introduced in Subsection 4.3.1.1 a Gamma distribution for the variance of the model σ^2. As in the second Chapter, we analyze the relevance of the prior on the observation variance in varying weights DDP-AR(2) model. In Figure 4.10, the variable weights are not influenced by two subgroups, because we should expect different directions of the points, due to the falling side and rising side, also because the two subgroups have a different sample size. The direction of this two subgroups and also the fact that the points are close to each others motivate that the variable weight prior even if is based on the data in the DDP process doesn't influence it.

In Figure 4.10, we propose for the last thirty years the real time series with the DDPAR(2) variable weights with fixed observation variance and the DDPAR(2) variable weights with random variance.

The distance between the series is not so relevant and also the performance of the DDP-AR(2) with varying weights for fixed variance is better on the influent point

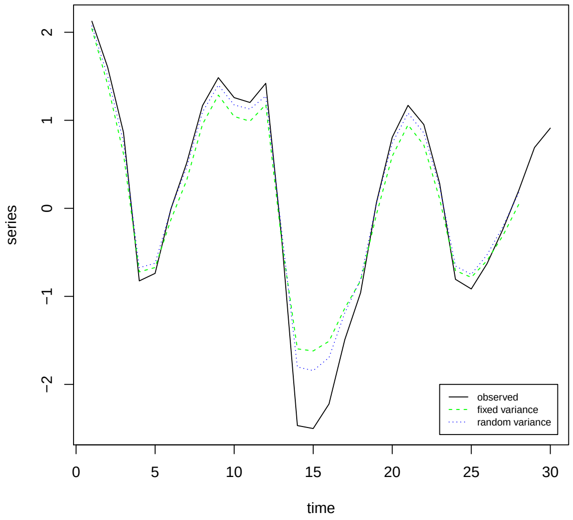

Figure 4.11: $\hat{y}_t = E(y_t \mid y_1, \ldots, y_{t-1})$ for the DDPAR(2) variable weights with fixed and random variance

at time $t = 1918$.

4.5.2 Single-p DDP prior for a complete AR(2) sampling model

We discuss in this subsection an extension of nonparametric autoregressive models to the autoregressive mean function to higher order lagged terms to allow AR(p) structure. As a motivating application, we assume an harmonic autoregression, such that it depends jointly from the first two lags, $Y_t \mid (Y_{t-1}, Y_{t-2})$. Therefore, conditional on Y_{t-1} and Y_{t-2}, we have $Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$ for $t \geq 3$, where $\{\varepsilon_t\}$ is the usual sequence of Gaussian random variables. We consider the single-p DDP prior for a complete autoregressive lag two priors. We modify the model proposed in equation (4.2) as follows:

$$\begin{aligned} Y_t \mid Y_{t-1} = y_1, Y_{t-2} = y_2, r_t = h \sim N(Y_t \mid \beta_{1h} y_1 + \beta_{2h} y_2, \sigma^2) \\ p(r_t = h) = \omega_h \\ (\beta_{1h}, \beta_{2h}) \overset{iid}{\sim} G^0 \end{aligned} \tag{4.8}$$

we suppose that $\beta_1 \overset{iid}{\sim} N(m_{\beta_1}, \sigma^2_{\beta_1}) \times \beta_2 \overset{iid}{\sim} N(m_{\beta_2}, \sigma^2_{\beta_2})$ where the averages and the variances of the priors are assumed known. We observe that the marginal distribution for the locations is: $\theta(y_1, y_2) \sim N(Y_t \mid m_{\beta_1} y_1 + m_{\beta_2} y_2; \sigma^2_{\beta_1} y_1^2 + \sigma^2_{\beta_2} y_2^2)$. In this case, the hyperparameter space has H new parameters for β_{2h} and the variance, σ^2, is costant:

$$\xi = (V_1, V_2, ..., V_{H-1}, \beta_{11}, \beta_{12}, ..., \beta_{1H}, \beta_{21}, \beta_{22}, ..., \beta_{2H}, r_1, r_2, ..., r_T)$$

The joint posterior distribution is

$$p(Y, \xi) = \prod_{t=3}^{T} P(Y_t \mid \xi) \prod_{t=3}^{T} p(r_t \mid \xi \setminus \{r_t\}) \prod_{h=1}^{H} p(V_h) p(\beta_{1h}) p(\beta_{2h}) =$$

$$\prod_{t=3}^{T} N(Y_t \mid \beta_{1,r_t} Y_{t-1} + \beta_{2,r_t} Y_{t-2}, \sigma^2) \prod_{t=3}^{T} \omega_{r_t} \prod_{h=1}^{H} Be(V_h \mid 1, M) N(\beta_{1h} \mid m_{\beta_1}, \sigma^2_{\beta_1}) N(\beta_{2h} \mid m_{\beta_2}, \sigma^2_{\beta_2})$$

given the other parameters, the posterior distribution for β_{2h} is $p(\beta_{2h} \mid \xi \setminus \{\beta_{2h}\}) \propto N(\beta_{2h} \mid m_{\beta_2}, \sigma^2_{\beta_2}) \prod_{t \in Q_h} N(Y_t \mid \beta_{1h} Y_{t-1} + \beta_{2h} Y_{t-2}, \sigma^2)$ we consider the autoregressive model as

$$Y_t = \beta_{1h} Y_{t-1} + \beta_{2h} Y_{t-2} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2).$$

and we define $Y_t^{**} = \frac{Y_t - \beta_{1h} Y_{t-1}}{Y_{t-2}}$, (where β_{h1} is a costant to respect β_{2h}) having that

$$Y_t^{**} = \beta_{2h} + \epsilon_t^{**} \qquad \epsilon_t^{**} \sim N(0, S_t)$$

it means that

$$Y_t^{**} \mid \beta_{2h} \sim N(\beta_{2h}, S_t)$$

where $S_t = \frac{\sigma^2}{Y_{t-2}^2}$. We have the same subset, as before, $Q_h = \{t : r_t = h\}$ for the link between the latent variable r_t and the trajectories, β_{2h}. We indicate the number of elements in this subset as $n_h =\mid Q_h \mid$.

$$p(\beta_{2h} \mid \xi \setminus \{\beta_{2h}\}) \propto N(\beta_{2h} \mid m_{\beta_2}, \sigma^2_{\beta_2}) \prod_{t \in Q_h} N(Y_t^{**} \mid \beta_{2h}, S_t)$$
$$p(\beta_{2h} \mid \xi \setminus \{\beta_h\}) \propto N(\beta_{2h} \mid m_1, V_1)$$
$$m_1 = V_1(\frac{m_{\beta_2}}{\sigma^2_{\beta_2}} + \sum_{t \in Q_h} \frac{Y_t^{**}}{S_t}) \;\; V_1^{-1} = (\sigma_{\beta_2}^{-2} + \sum_{t \in Q_h} \frac{\sigma^{-2}}{S_t})$$

The posterior distributions of the other parameters are similar to the obtained solutions for the single-p DDP-AR(2) model, defined in equation (4.2). The latent indicator r_t links costant weights of the stick-breaking representation (see in equation 1.1) and the trajectories for the point masses (4.2).
Therefore, the posterior distribution for the parameter V_h has the same result, as in the second Chapter and as we discussed before:

$$p(V_h \mid \xi \setminus \{V_h\}) \propto Be(1, M) \prod_{t \in S_h} (1 - V_h) \prod_{t \in Q_h} V_h =$$
$$Be(1+ \mid Q_h \mid, M+ \mid S_h \mid)$$

where $S_h = \{t : r_t > h\}$ and $Q_h = \{t : r_t = h\}$.

For the latent variable, r_t, we have

$$p(r_t = h \mid \xi \setminus \{r_t\}) \propto \omega_h N(Y_t \mid \beta_{1h} Y_{t-1} + \beta_{2h} Y_{t-2}, \sigma^2).$$

Keeping in mind that the autoregressive component has two lags, the posterior distribution is calculated on the parameter for the trajectories for the point masses, β_{1h}, for $h = 1, 2, \dots, H$. In principle, we have to consider

$$Y_t = \beta_{1h} Y_{t-1} + \beta_{2h} Y_{t-2} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2).$$

and we define $Y_t^* = \frac{Y_t - \beta_{2h} Y_{t-2}}{Y_{t-1}}$, obtaining

$$Y_t^* = \beta_{1h} + \epsilon_t \qquad \epsilon_t \sim N(0, W_t)$$

it means that

$$Y_t^* \mid \beta_{1h} \sim N(\beta_h, W_t)$$

where $W_t = \frac{\sigma^2}{y_{t-1}^2}$. We also define the subset $Q_h = \{t : r_t = h\}$ for the link between the latent variable r_t and the trajectories for the point masses, β_{1h}. We indicate the number of elements in this subset as $n_h =\mid Q_h \mid$.

$$\begin{aligned} p(\beta_{1h} \mid \xi \setminus \{\beta_{1h}\}) \propto N(\beta_{1h} \mid m_{\beta_1}, \sigma^2_{\beta_1}) \prod_{t \in Q_h} N(Y_t^* \mid \beta_{1h}, W_t) \\ p(\beta_h \mid \xi \setminus \{\beta_h\}) \propto N(\beta_{1h} \mid m_1, V_1) \\ m_1 = V_1 (\frac{m_{\beta_1}}{\sigma^2_{\beta_1}} + \sum_{t \in Q_h} \frac{y_t^*}{W_t}) \;\; V_1^{-1} = (\sigma^{-2}_{\beta_1} + \sum_{t \in Q_h} \frac{\sigma^{-2}}{W_t}) \end{aligned}$$

For these results, we could think about a more general construction for an autoregressive p-lagged terms with prior distribution considering a parameter $\boldsymbol{\beta} = (\beta_{1h}, \beta_{2h} \dots \beta_{ph})$ for $h = 1 \dots H$, obta

4.6 Discussion

The guidelines of this chapter are essentially two possible extentions of the former models presented in the second and in third Chapter, respectively. We introduced on the trajectories more than one lag ($p \geq 1$) as a simple extension of the usual nonparametric autoregressive models, just for example we used only two lags. On the lags we inserted only one prior for an applied comparison with West *et al.* result for the lynx trappings. However, it is interesting to check how the solutions will change if we have more than one prior on th lags. In Bayesian nonparametric prospective we moved from a single-p DDP to varying weights DDP, which is the core of this chapter. Specifically, we illustrated variable weights DDP such that the distribution of latent variable, r_t is connected to the lags of the trajectories Y_{t-1} and Y_{t-2} and the utility of a logistic function. Many applications of Bayesian nonparametric models use the stick-breaking representation with costant weights. The reason is the conjugacy property of a Beta distribution as prior allowing fast results in a Gibbs sampler algorithm. When the weights of the stick-breaking are varying, the model presents more parameters and more priors and it is not possible to apply the conjugacy even if the prior is a Gaussian distribution, as we showed in Section 4.3, it is necessary to introduce Metropolis-Hastings steps in the Gibbs sampler. Here we also skip an other possible point of research. We could introduce dependence only on varying weights and fix trajectories. Nieto-Barajas *et al.* (2011) affirmed that for time series models can be sufficient introduce varying weights in a DDP prior model. However, we introduced also two priors in the harmonic autoregression for the single-p DDP. The core of this chapter is the influence and the effects of the assumptions on the sticks. For the application on the lynx trappings, we did not consider possible variations inserting two priors on the lags. This is an open point of research that we illustrated as methodological framework. It could be interesting for the same application, apply and compare the performance on the peaks of the time series assuming a single-p DDP-AR(2) model with two lagged priors and varying weights DDP-AR(2) with two lagged prior terms. An other direction can be use bsplines. About this point we will remand to the next chapter.

Bibliography

Anderson, T. W. (1971). *The Statistical nalysis of Time Series*, New York: Wiley.

Antoniak, C. E. (1974). ”Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems”. *Annals of Statistics*, **2**, 1152-1174.

Bartlett, M. S. (1954). ”Problmes de l’analyse spectrale des sries temporelles stationnaires.” *Publ. Inst. Statist. (Univ. de Paris)*. **3**, 119-134.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). ”The New S Language”. *Wadsworth Brooks/Cole.*

Campbell, M. J. and Walker, A. M. (1977). ”A Survey of Statistical Work on the Mackenzie River Series of Annual Canadian Lynx Trappings for the Years 1821-1934 and a New Analysis.” *Journal of the Royal Statistical Society*. Series A. **140**, 4. 411-431.

De Iorio, M., M$\ddot{u}$ller, P., Rosner, G., MacEachern, S. N., (2004).”An ANOVA Model for Dependent Random Measures”. *Journal of the American Statistical Association*, **99**, 205-215.

Doss, H. (1994). ”Bayesian Nonparametric Estimation for Incomplete Data via Successive Substitution Sampling”. *Annals of Statistics*. **22** vol.4. 1763-1786.

Escobar, M. (1988). ”Estimating the means of several normal populations by estimating the distribution of the means”. Ph.D. dissertation, Dept. Statistics, Yale Univ.

Escobar, M. D. West, M. (1995). ”Bayesian density estimation and inference using mixtures”. *Journal of American Statistical Association*, **90**, 577-588.

Ferguson, T. S. (1973). ”A Bayesian analysis of some nonparametric problems”. *Annals of Statistics*, **1**, 209-230.

Griffin, J.E., Steel, M., (2006). ”Order-Based Dependent Dirichlet Processes”. *Journal of the American Statistical Association* THEORY AND Methods, **101**, 179-194.

Gasparini, M., (1996). ”Bayesian density estimation via mixture of Dirichlet processes”. *Journal Non Parametric Statistics*. **6**. 355-366.

Kottas, A., Raftery, A.E. (1995). Bayes Factors. *Journal of American Statistical Association*, **90**, 773–795.

Ishwaran, H., James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

Ishwaran, H., and Zarepour, M. (2002). ”Exact and approximate sum representations for the Dirichlet process”. *The Canadian Journal of Statistics.* **30**, 2. 269-283.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm.Statist. Simulation Comput.* **23**, 727-741.

MacEachern, S.N., M$\ddot{u}$ller, P., (2000). ”Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichelt Process Mixture Models”. *Robust Bayesian Analysis. Lecture Notes in Statistics*. **152** 295-316.

Moran, P. A. P. (1953a). ”The statistical analysis of the Canadian lynx cycle: I”. *Austr. Zool.***1**, 163-173.

Moran, P. A. P. (1953b). ”The statistical analysis of the Canadian lynx cycle: II”. *Austr. Zool.***1**, 291-298.

Nieto-Barajas, L., Müller, P., Ji, Y., Lu, Y. and Mills, G. (2011). A time-Series DDP for Functional Protemics Profiles. *Submitted.*

Sethuraman J. (1994), ”A constructive definition of the Dirichlet process prior”, *Statistica Sinica*. **2**, 639-650.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* Second Edition. Springer, New York.

West M., M$\ddot{u}$ller P. and MacEachern N. S., (1997). ”Bayesian Models for Non-Linear Autoregressions”. *Journal of Time Series Analysis*. **18**, 593-614.

Wood, S., Rosem, O. and Kohn, R. (2011). Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics*, **20**, 174-195.

Yule, G. U. (1927). ”On the method of investigating periodicities in disturbed series, with special reference to Wolferś sunspot numbers.” *Phil. Trans.* A, **226**, 267-298.

Chapter 5

Further Research

Abstract. *In this chapter, we propose an overview of the proposed models. We discuss about further possible developments of research.*

Several authors wrote about dependent Dirichlet processes and Bayesian time series analysis, i.e., autoregressive models and conditional autoregressive models. In principle, structural differences are in the assumptions of dependent prior distributions. The dependent Dirichlet process introduces dependence on the base measures such that in some way marginally is a DP. The DDP has three possible different structures: common weights and dependent locations, known as single-p DDP; dependent weights and common locations; dependent weights and locations. Griffin and Steel (2006; 2011) and Rodriguez and ter Horst (2008) considered dependence on the weights and common locations as form of dependent Dirichlet processes for dynamic linear models as well, involving state space problems and not traditional autoregressive linear models. Walker and Mena (2005; 2011) focused on Dirichlet process prior, instead of DDP and resolving the discretness introducing a Gaussian process construction for a stationary autoregressive lag one model, the main difference is the presence, in our proposals, of one more dependent level in the hierarchy.

As methodological statement, we assumed homogeneity over time for random probability measures. This assumption is not trivial: in particular, for a single-p DDP we do not need stationary assumption. As a consequence, we excluded the Gamma process over time. We analyzed a generic sequence of increasing conditionals $Y_t \mid Y_{t-1}, \ldots, Y_0$ such that the first lag was sufficient to describe asymmetric dependence. A similar process was proposed for a real DNA-sequencing dataset. We considered more real problems as the noises due to the Base Calling method. In the third Chapter, we discussed a novel methodological improvement. We introduced Bayesian mixtures of dependent locations. The sample size of the DNA-sequencing is equal to 36,000 intensity measurements and 4 nucleotides. Gibbs sampler algorithm was implemented and Bayesian predictive distributions obtained. However, more robust analysis and massive iterations can be done for the submission of the related paper.

In the literature, there are different papers concerning about single-p DDP prior models based on the dependence of the trajectories and others were considering varying weights DDP prior and fixed locations. The contribution of our fourth Chapter is to study the relevance of costant and varying weights DDP in an autoregressive fashion-type. We introduce dependence on the locations and on the weights of dependent Dirichlet processes. Previous papers were focused only on DDP structures with common locations and dependent weights or on dependent locations and common weights. The use of dependent weights and jointly dependent locations is the main theme of our contribution focused on time series analysis. For the first time we introduce a different proposal of the stick-breaking representation. We abandoned traditional rescaled Beta distribution for the weights, and we added a logistic function for varying weights and Gaussian distributions for the prior of the parameters. The other relevant assumption in all our proposals and more clearly discussed in the fourth Chapter is the finite Dirichlet process, (Ishwaran and Zarepour, 2008). In a Pólya urn schemes is unknown the number of sticks in the stick-breaking representation of the DP, we infer for a solution. An other open problem is a random functional for the choice of the number of sticks. MacEachern

(1999) proved that a DDP holds Muliere and Tardella (1998) result.

In this thesis we based on continous outcomes and equal spaced over time for autoregressive models. This is one of our points of force, but it is also a limit of our research: we discussed about only continous outcomes, we could extend to sequences of binary responses or ordinal outcomes involving the convolution of a continous kernel, i.e. Gaussian distributions with a countable mixture of regressions on lagged terms.

We proposed the single-p DDP through linear or quadratic trajectories. These are only two examples of a large class of models that it can adopt different forms. Other nonlinear functions of lagged terms can be accomodated under the general framework, for instance b-splines (Eilers and Marx, 1996). We assumed linearity on the autoregressive coefficients, but the mixture component means can be arbitrarly specified as we adapted for the DNA-sequencing dataset.

The simplest variation is for the single-p DDP-AR(1) model: in the DP mixture modelling we used a Gaussian kernel distribution function. We could introduce a Skew Normal distribution (Azzalini, 1985) for the errors of the autoregression lag one keeping a single-p dependent Dirichlet process prior. This new model will change the distribution for the depending locations. The Skew Normal distribution has the same kernel distribution of the normal distribution and it considers in the exponential function the parameter of skewness. In that case, we could compare if one more parameter as the skewness is useful for real problems or not. Most probably the applicability of this type of processes could be useful for financial problems, where there are stochastic volatility problems and not symmetric distributions can be assumed. This type of modeling can provide a better prediction of the phenomenon and can be a new solution for high volatiliy problems usually resolved using ARCH or GARCH models.

In the second Chapter, we discussed abut a multivariate unseparable equations underling the relevance of the ANOVA structure. However, we think about to introduce a product of independent variables in the model, and in a classical point of view it coud be constructed as a Generalized Skew Normal, in that case, our assumptions have to be reviewed.

An other methodological variation for our last proposal is a Gaussian process. Gelfand, Kottas and MacEachern (2005) introduced the Gaussian process for spatial problems. They proposed an alternative process at the DDP. We could introduce this type of process for the DDP structure. The idea is essentially to use a Gaussian process on dependent locations and varying weights, or for a single-p DDP.

Bibliography

Contreras-Christan, A., Mena, R. H. and Walker, S. G. (2009). ”On the construction of stationary AR(1) models via random distributions”. *Statistics*, **43**, 3, 227-240.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing with B-splines and penalties. *Statistical Science*, **11(2)**, 89–121.

Jain, S. and Neal, R. M. (2000). ”A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model”. *Technical Report* **2003**. University of Toronto.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). *Journal of American Statistical Association*. **100**. 1021-1035

Görür, D. and Rasmussen, E. (2009). ”Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution”. *Journal of Computer Science and Technology.* **25**,4. 615-626.

Mena, R. H., Walker, S. G. (2007). ”Stationary mixture transition distribution (MTD) models via predictive distributions”. *Journal of Statistical Planning and Inference*. **137**. 3103-3112.

Neal, R. M. (1997). ”Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification.” *Technical Report* **9702**. University of Toronto.

Neal, R. M. (1998). ”Regression and Classification Using Gaussian Prosses Priors.” *Bayesian Statistics.* Oxford University Press.

Zucchini, W. and Macdonald, I.L. (2009). *Hidden Markov models for time series*. Taylor & Francis.