UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

BAYESIAN INFERENCE USING LOSS FUNCTIONS FOR SOME DATA ANALYSIS PROBLEMS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTORA EN CIENCIAS

PRESENTA:
GUADALUPE EUNICE CAMPIRÁN GARCÍA

TUTOR PRINCIPAL:
EDUARDO ARTURO GUTIÉRREZ PEÑA
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS, UNAM

MIEMBROS DEL COMITÉ TUTOR:
RAMSÉS HUMBERTO MENA CHÁVEZ
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS, UNAM
LUIS ANTONIO RINCÓN SOLIS
FACULTAD DE CIENCIAS, UNAM

CIUDAD UNIVERSITARIA, CD. MX. JUNIO DE 2018.

A mi mami

Difícil lograrlo sin su apoyo y difícil no llegar con él.

Agradecimientos

Agradezco profundamente a mi tutor, el Dr. Eduardo Gutiérrez-Peña, por su invaluable guía tanto a nivel académico como a nivel personal. Su infinita paciencia y gran ecuanimidad me permitieron concluir este trabajo. ¡¡MIL GRACIAS EDUARDO!!

Quiero agradecer a mis sinodales, Ramsés Mena, Luis Enrique Nieto, Grabriel Nuñez y Lizbeth Naranjo por el tiempo que dedicaron a la lectura de la tesis, así como sus comentarios y correcciones.

A Ramsés, quien siempre me ha proporcionado una visión muy amplia y completa del área de estadística bayesiana no paramétrica. También me ha contagiado de su inagotable energía.

A Luis Enrique Nieto por sus comentarios precisos y atinados que me forzaron a comprender mejor los conceptos. Gracias por ser un ejemplo de equilibro entre trabajo académico y vida personal.

A Gabriel Nuñez, que apesar de su carga de trabajo, me dedicó un día completo para explicarme con cuidado sus comentarios y correcciones. También te agradezco tus palabras de aliento que me hiciste en diferentes eventos académicos. Me motivaron a seguir adelante.

A Lizbeth Naranjo, quien revisó el desarrollo algebráico de los apéndices B y C en una etapa temprana de la elaboración de la tesis.

Quiero aprovechar este espacio para reconocer y agradecer el trabajo continuo y constante de todos los académicos del departamento de probabilidad y estadística del IIMAS por impulsar la estadística y la probabilidad en México. Espero estar a la altura para unirme a su labor.

A Silvia Ruiz, Alfredo, Alexia, Coco, Élida , Lucía, María Inés y Tere, que siempre están en la mejor disposición por apoyarnos y facilitarnos los trámites del posgrado, conacyt y asistencia a congresos. Sin su disposición y alegría, el proceso sería muy arduo.

A todo el personal adminstrativo, de la biblioteca y de intendencia, por su trabajo silencioso pero escencial para el buen funcionamiento de la UNAM.

Gracias a todas aquellos compañeros con los que coincidí en algún momento durante mis estudios en el IIMAS, por sus consejos, por darme momentos inolvidables tanto felices como tristes, pero que al final son parte de mí; por compartir un poco de su tiempo, de su vida, sus alegrías, sus sufrimientos y sus logros.

Estoy agradecida por la **beca (No.\ 205515)** del Consejo Nacional de Ciencia y Tecnología (CONACYT, Mexico) . Este trabajo fue apoyado por el **proyecto IN106114-3** del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (DGAPA-UNAM, Mexico).

A mi mamá por su amor, sacrificios y ayuda. Por darme el mejor regalo del mundo: el Dharma.

A mi papá por su cariño y apoyo. Por ser un ejemplo para salir adelante ante las tragedias.

A mi abuelita por ser un ejemplo de fortaleza, perdón, compasión y amor a la vida.
¡No sabes como te extraño y cuanta falta me haces!

A mi familia y amigos por apoyarme y ayudarme en tiempos difíciles.

A la familia Calderas Sánchez Marín, por su cariño y comprensión
¡¡Gracias Pedro, Vale, Patito, Sergio y Paty!!

A la familia Descentis Mulia, por ser un apoyo incondicional para mi mamá y para mí.
¡¡Muchas gracias Madrina!! ¡¡Muchas gracias Roberto y Eva!!

A la familia Descentis Santa María, por contagiarme de su optimismo y amor a la vida.

A la familia Descentis Giraldo por su cariño y apoyo.

A la familia Lim Reyo por su apoyo y compañía, Gracias Madrina Rosita, Gracias Padrino Juan, Gracias Daya, Yatzi y Juan.

A la familia Villegas, por su hospitalidad y solidaridad en todo momento.

A la familia Calderas Lim por las historias familiares vividas.

A la familia González, por su amistad y por ser una extensión de la familia.

A Mario, por su constante apoyo. Indisctutiblemente eres un miembro más de la familia.

A la familia Rello, por su calidez.

A Luis Gómez, Élsa López Bravo y Érika Vértiz por ayudarme a recobrar la confianza en mí.

A Rocío, por ser mi mejor amiga y confidente.

A Angeles, por su cariño y amistad. Por su apoyo en el fallecimiento de mi abuelita.

A Gonz!!, por tantas vivencias de la maestría, por iniciarme con el Proceso de Dirichlet, por ser mi cuate y amigo.

A Mónica Tinajero, por su amistad, con quien compartí muchas vivencias del doctorado, por enseñarme los encantos del muestro y con quien tuve discusiones estimulantes que contribuyeron al capítulo 3 de este trabajo.

A todos mis amigos que por falta de espacio no son mencionados personalmente pero ocupan un lugar importante en mi corazón.

Contents

List of Figures

List of Tables

LIST OF TABLES

List of Algorithms

Glossary

$\boldsymbol{\mu}$	$(\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n)$, page 30
$\boldsymbol{\pi}_{S_j}$	$(\pi_i \vert i \in S_j)$, page 13
$\boldsymbol{\theta}$	$(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$, page 9
$\boldsymbol{\theta}^S$	vector parameter of $f_{\boldsymbol{Y}_S}$, see equation (1.3), page 9
$\boldsymbol{F}$	$(F_1, ..., F_n)$, page 25
$\boldsymbol{X}_S$	$\left(X^i \vert i \in S\right)$, page 83
$\boldsymbol{y}$	$(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, page 6
$\boldsymbol{y}_S$	$(\boldsymbol{y}_i, i \in S)$, page 6
$\boldsymbol{y}_S^X$	$(y_i \vert x_i \in X_S)$, page 83
$\delta_A(x)$	characteristic function of a set. $\delta_A(x) = 1$ if $x \in A$ and $\delta_A(x) = 0$ otherwise
$\hat{\boldsymbol{\mu}}_\rho(\boldsymbol{x})$	$E(\boldsymbol{\mu} \vert \rho, \boldsymbol{x})$, page 31
$\hat{\boldsymbol{\mu}}_B(\boldsymbol{x})$	$E(\boldsymbol{\mu} \vert \boldsymbol{x})$, page 31
$\mathbb{N}$	the set of natural numbers
$\mathbb{N}^+$	the set of strict positive natural numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^+$	the set of strict positive real numbers
$\mathbb{R}^p$	p-dimensional Euclidean space
$\mathcal{P}$	the space of all partitions of S_0
$\operatorname{cov}(\mathbf{X}, \mathbf{Y})$	covariance of $\mathbf{X}$ and $\mathbf{Y}$
$\min\{A\}$	minimum value of the finite set A
$\overline{\boldsymbol{y}}$	mean of $\boldsymbol{y}$
ρ	partition of a set, page 5
$\underline{\mathbf{0}}_{n\times p}$	$n \times p$ matrix with all entries equal to zero.
$\vert A \vert$	determinant of matrix A
$\vert S \vert$	cardinality of set S
B_n	Bell number, see equation (1.1), page 6
$c(A)$	cohesion function of set A, see equation (1.2), page 8
$DP(\alpha, \nu)$	Dirichlet process with dispersion parameter α and base measure ν, page 10
$F_{\boldsymbol{X}}^S$	joint distribution function of $\boldsymbol{x}_S$ given $S \in \rho$, see equation (2.3), page 25
$f_{\boldsymbol{Y}_S}$	joint probability density of $\boldsymbol{y}_S$ given $S \in \rho$, see equation (1.3), page 9
F_X^S	the common marginal distribution for the x_i's with $i \in S$ given that $S \in \rho$, see equation (2.3), page 25
f_{Y_S}	the common marginal density for the $\boldsymbol{y}_i$ when $i \in S$ given that $S \in \rho$, see equation (1.3), page 9
$N(\boldsymbol{\mu}, \boldsymbol{\lambda})$	normal distribution with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\lambda}$, page 117
$N_n(\boldsymbol{\mu}, \boldsymbol{\lambda})$	multivariate normal distribution with vector mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\lambda}$, page 117
$Ng(\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta)$	normal-gamma distribution, page 118
$Ng_n(\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta)$	multivariate normal-gamma distribution, page 118
$RandI$	Rand index , page 48
$RP(S_0)$	random partition of S_0, page 6
s_i	$s_i = j$ if $i \in S_j, i = 1, \ldots, n$, page 84

$sN(\xi, \omega, \alpha)$ skew normal distribution , page 118

t_ν univariate t-distribution with ν degrees of freedom , page 118

$t_\nu(\boldsymbol{\mu}, \boldsymbol{V})$ multivariate non-standardized Student's t-distribution with ν degrees of freedom and vector mean $\boldsymbol{\mu}$, page 118

$tr(A)$ the trace of matrix A

aCGH array comparative genomic hybridization , page 60

ECP E-divisive change-point analysis , page 47

EPPF exchangeable product partition function , see equation (1.1), page 7

i.i.d. independent and identically distributed

ind independent

MCMC Markov chain Monte Carlo

NDP nested Dirichlet process , page 27

NGPPM normal-gamma Product Partition Model , see equation (1.3), page 9

NPPM(s) nonparametric product partition model(s), page 24

PELT pruned exact linear time , page 47

posterior cohesion $c(S_j)p_{S_j}(\boldsymbol{x}_{S_j})$, page 25

PPM(s) product partition model(s), page 8

Robust estimator estimator that is resistant to errors that can be produced by deviations from assumptions; it will still have a reasonable efficiency, reasonably small bias, as well as being asymptotically unbiased.

Introduction

We have bigger houses, but smaller families
More conveniences, but less time.
We have more degrees, but less sense.
More knowledge, but less judgement.
More experts, but more problems.
More medicines, but less wellness.
We have conquered outer space, but not inner space.
We've cleaned up the air, but polluted our soul.
We've split the atom, but not our prejudice.

Dalai Lama XIV extract from THE PARADOX OF OUR TIMES

Loss functions play a key role in Bayesian inference; they lie at the foundations of Bayesian decision theory. Nevertheless, loss functions do not seem to have been extensively and systematically used in applied problems such as cluster analysis, change-point analysis, finite population sampling or variable selection. Binder (1978, 1981) provides a first approach to this issue, placing emphasis on the loss incurred when we cluster two subjects that do not belong together or when we do not cluster two observations that belong to the same cluster. Lau and Green (2007) developed this ideas in a Bayesian nonparametric framework, whereas Hurn et al. (2003) applied Binder's loss function to cluster linear regression curves. Later, Killick et al. (2012) introduced a change-point analysis with a linear computational cost in a Bayesian parametric framework. A modern approach is provided by Yau and Holmes (2013) using Markovian loss functions and dynamic programming algorithms in order to make inference. Their results can be applied to change-point analysis or product partition models. In variable selection, Hahn and Carvalho (2015) and Puelz et al. (2016) used a loss function to

compare the various models with the full one. Quintana and Iglesias (2003) provided a new approach, proposing a loss function for estimating the parameters of a product partition model and penalizing the number of clusters. Quintana et al. (2005a) and Bormetti et al. (2012) applied this approach to outlier detection and change-point analysis.

In this thesis, we exploit the ideas presented in Quintana and Iglesias (2003) and extend their methodology in three relevant problems in statistics. In the case of cluster and change-point analysis, we extend parametric product partition models to the nonparametric case and generalize their loss function to deal with arbitrary distributions functions instead of parameters. For finite population sampling, we apply them to find the optimal post-stratification. Finally, in variable selection, we include a term in the loss function which penalizes correlated variables in addition to model complexity.

This thesis is organized as follows. The first chapter provides a brief introduction to parametric product partition models (PPMs) as defined by Barry and Hartigan (1992), as well as to the Dirichlet process introduced by Ferguson (1973) and their relationship. We also describe how PPMs are applied to change-point detection and provide simulated examples where we show the limitations of using the marginal probability criterion of Loschi and Cruz (2005) to detect change points. We also show how this detection can be improved using a loss function.

In Chapter 2, we define nonparametric product partition models (NPPMs), which use the Dirichlet process, and describe how to model the distribution of the data within clusters. We also discuss some important properties such as its relationship with the nested Dirichlet process (Rodríguez et al., 2008). We then propose an inference procedure using suitable loss functions for distribution functions, exploiting the ideas presented in Quintana and Iglesias (2003). We apply NPPMs to nonparametric change-point analysis and take advantage of the random partition structure to deal with missing values. We also compare our methodology through simulations with other models recently discussed in the literature, and apply it to financial and genetic data. This chapter represents the main contribution of the thesis and the main results are published in Campirán García and Gutiérrez-Peña (2018).

In Chapter 3, we begin with a brief review of finite population sampling; then, we explore a new framework for Bayesian post-stratification sampling using random partition models and propose a suitable loss function for estimating the parameters of

interest and finding the optimum post-stratification. This would be the first model for sampling design. We also discuss a new methodology, based on the Bayesian learning process, that allows us to use previous surveys to obtain better estimates of the parameters of interest.

In Chapter 4, we present a novel approach to variable selection in regression models which can also be used in logit models and other generalized linear models. We propose a loss function that penalizes high correlations between the explanatory variables, in addition to the model complexity. Furthermore, we provide an algorithm to find the subset of variables with minimum expected loss. Our approach is similar to that of Hahn and Carvalho (2015). They use a loss function which compares each model with the full one; however, this approach has several limitations that we will point out later in that chapter.

Finally, in Chapter 5 we offer some concluding remarks and discuss further work. In Appendix A we provide a more detailed proof, clarifying all the elements of the sequential decision problem presented in Quintana and Iglesias (2003). In Appendices B and C we present the calculations of the prior predictive distributions of the models used in this thesis.

We used R Core Team (2016) and Gfortran in the simulations and numeric examples presented in this work.

Chapter 1

Product Partition Models and Dirichlet Process

My joy is like Spring, so warm it makes flowers bloom all over the Earth.
My pain is like a river of tears, so vast it fills the four oceans.
Please call me by my true names, so I can hear all my cries and my laughter at once, so I can see that my joy and pain are one.

Thich Nhat Hahn

In this chapter, we provide a brief introduction to product partition models (PPMs) defined by Hartigan (1990), the Dirichlet process (Ferguson, 1973) and their relationship established by Quintana and Iglesias (2003). We describe how PPMs can be applied to change-point analysis (Barry and Hartigan, 1993). Finally, through simulations, we compare marginal-probability and loss-function criteria to select change points in PPMs.

1.1 Random partitions

In this section, we review a class of models that induce probability distributions on the space of partitions of a finite set of objects. This kind of models are used in cluster analysis and model comparison, among other applications (Quintana 2006; Tarantola et al. 2008). Let $S_0 = \{1, \ldots, n\}$ be a index set of observations. Let $\rho = \{S_1, \ldots, S_k\}$ denote a partition of S_0 into k subsets with $S_i \subseteq S_0$, and $S_i \cap S_j = \varnothing$ for all $i \neq j$. For

example, if $S_0 = \{1, 2, 3\}$ then we obtain five partitions:

$$[\{1,2,3\}]\ [\{1,2\},\{3\}]\ [\{1,3\},\{2\}]\ [\{1\},\{2,3\}]\ [\{1\},\{2\},\{3\}]$$

To avoid any confusion in the following definitions, we will assume that the elements of S, with $S \in \rho$, are sorted in ascending order and that $\min\{s|s \in S_1\} < \min\{s|s \in S_2\} < \cdots < \min\{s|s \in S_k\}$. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be a vector of n observations with $\boldsymbol{y}_i \in \mathbb{R}^p$ for $1 \leq i \leq n$. As in Crowley (1997), we define the vector

$$\boldsymbol{y}_S = (\boldsymbol{y}_i, i \in S).$$

Obviously, a partition over S_0 induces a partition over the entries of $\boldsymbol{y}$. The number of possible partitions of n objects is the Bell number, B_n, which satisfies the recursive equation

$$B_{n+1} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) B_k \tag{1.1}$$

with $B_0 = 1$. Note that this number grows exponentially; for example, $B_{200} > 10^{275}$!!! (Dahl, 2009). This represents a challenge when we are dealing with partitions. Despite this computational difficulty, random partitions models have been applied in medicine (Leon-Novelo et al. 2012; Müller and Quintana 2010; Müller et al. 2011), finance (Bormetti et al. 2012; De Giuli et al. 2010; Quintana et al. 2005b), the analysis of contingency tables (Tarantola et al. 2008), among other problems.

In some applications, it is enough to consider a particular class of partitions. For example, if $\boldsymbol{y}$ is a real vector for which all the entries are ordered, it seems natural to consider the partitions $\rho = \{S_1, \ldots, S_k\}$ with the following form:

$$S_1 = \{1, \ldots, m_1\}, S_2 = \{m_1 + 1, \ldots, m_2\}, \ldots, S_k = \{m_{k-1} + 1, \ldots, n-1, n\}$$

with $m_i < m_j$ if $i < j$. Note that there are 2^{n-1} partitions of n points into blocks of consecutive segments.

This kind of partitions are useful, for example, in change-point detection problems (Barry and Hartigan 1992, 1993), multiple change-point analysis for linear regression (Loschi et al. 2010) and text segmentation (Kehagias et al. 2004).

A random partition, denoted by $RP(S_0)$, is a probability distribution over all the partitions of an $n-$element set $S_0 = \{1, \ldots, n\}$. Clearly, this induces a random partition over the entries of $\boldsymbol{y}$, denoted by $RP(\boldsymbol{y})$.

For several applications, two basic properties are desirable for a random partition model. The model should be *exchangeable* with respect to permutations of the indices of the experimental units (symmetry property): Let $\sigma = (\sigma_1, ..., \sigma_n)$ denote a permutation of S, and let $\boldsymbol{s}_\sigma = (s_{\sigma_1}, ..., s_{\sigma_n})$ describe the clusters implied by re-labeling experimental unit i by $h = \sigma_i^{-1}$, i.e., $\sigma_h = i$. We require

$$p(\boldsymbol{s}) = p(\boldsymbol{s}_\sigma)$$

for all partitions. A second important property, known as *scalability*, is that the model should scale across sample sizes. We want

$$p(\boldsymbol{s}_n) = \sum_{j=1}^{|\rho|+1} p(\boldsymbol{s}_n, s_{n+1} = j).$$

where $|\rho|$ is the number of clusters on the partition of S and $s_i = j$ if $i \in S_j$ denote the cluster memberships. A probability model on ρ that satisfies these two conditions is called an exchangeable product partition function (EPPF) which can be written as $p\left(|S_1|, ..., |S_n|\right)$ (Pitman, 1996). In words, $p(\rho)$ depends on the specific partition only indirectly through the sizes $|S_k|$ of the partitioning subsets S_k. Several probability models $p(\rho)$ have been used in the recent literature, including product partition models (PPM), species sampling models (SSM) and model-based clustering (MBC). The SSM and MBC satisfy the requirements of symmetry and scalability by definition, but not all PPMs do. We shall discuss this issue in Subsection 1.4.1. For an extensive review of random partitions, see for example, Quintana (2006).

Note that in clustering applications, each element of the partition represents a cluster. In model comparison, each partition and the associated probability $p_S(\boldsymbol{y}_S)$ represent a model for the data. For example, let $S_0 = \{1, 2, 3, 4, 5\}$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_5)$. Considering the partitions $\rho_1 = [\{1, 2, 3\}, \{4, 5\}]$ and $\rho_2 = [\{1, 2, 3, 4, 5\}]$. To the first partition we associate the model

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_5|\rho_1) = p_{\rho_{1,1}}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3) p_{\rho_{1,2}}(\boldsymbol{y}_4, \boldsymbol{y}_5)$$

and to the second one

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_5|\rho_2) = p_{\rho_2}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4, \boldsymbol{y}_5)$$

where $\rho_{1,1} = \{1, 2, 3\}$, $\rho_{1,2} = \{4, 5\}$ and $\rho_2 = \{1, 2, 3, 4, 5\}$.

Quintana (2006) gives emphasis to PPMs as a relevant random-partition model. We describe this class of models concisely in the following subsection.

1.2 Product partition models

Product partition models (PPMs) were defined by Hartigan (1990) and Barry and Hartigan (1992); these models are a particular case of random partitions. PPMs induce a probability distribution over all possible partitions of a finite set of distinct observations: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ with $\boldsymbol{y}_i \in \mathbb{R}^p$. For each partition, the n data points are divided into k subsets, and each data point $\boldsymbol{y}_i$ belongs only to one subset. Data points of distinct subsets are assumed independent and data points belonging to the same subset are assumed exchangeable. For any partition $\rho = \{S_1, \ldots, S_k\}$ of S_0 and data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, it is assumed that

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n|\rho) \propto \prod_{j=1}^{k} p_{S_j}(\boldsymbol{y}_{S_j}), \tag{1.2}$$

where $p_S(\boldsymbol{y}_S)$ is the conditional density for observations in S given that $S \in \rho$. This density depends only on S and not on other subsets in the partition. The partition ρ is in turn assigned a prior probability model

$$P(\rho = \{S_1, \ldots, S_k\}) \propto \prod_{j=1}^{k} c(S_j), \tag{1.3}$$

where, for $A \subseteq \{1, 2, \ldots, n\}, c(A) \geq 0$ is called the *cohesion function* of the subset A. It is well known that the posterior distribution of ρ is again a PPM with cohesions given by $c(S)p_S(\boldsymbol{y}_S)$. In the literature, the joint probability density of $\boldsymbol{y}_S$, denoted by $f_{\boldsymbol{Y}_S} = p_S(\boldsymbol{y}_S)$, belongs to a parametric family and the previous hierarchical model defined by equations (1.2) and (1.3) includes a level with a prior for the parameters (see, for example, Barry and Hartigan 1992, 1993; Bormetti et al. 2012; Crowley 1997; Dahl 2009; Fearnhead 2006; Hartigan 1990; Hegarty and Barry 2008; Jordan et al.

2007; Kehagias et al. 2004; Loschi 2002; Loschi and Cruz 2005; Loschi et al. 2003, 2010; Monteiro et al. 2011; Müller et al. 2011; Quintana and Iglesias 2003; Tarantola et al. 2008). Barry and Hartigan (1992) referred to these models as parametric PPMs.

Remark 1.1. *Recall that the entries of $\boldsymbol{y}_S$ are exchangeable. Then, by de Finetti's theorem,*

$$\begin{aligned} p_S(\boldsymbol{y}_S) &= \int f_{\boldsymbol{Y}_S}(\boldsymbol{y}_S|\boldsymbol{\theta}^S)p(\boldsymbol{\theta}^S)d\boldsymbol{\theta}^S \\ &= \int \prod_{i\in S} f_{Y_S}(\boldsymbol{y}_i|\boldsymbol{\theta}^S)p(\boldsymbol{\theta}^S)d\boldsymbol{\theta}^S \end{aligned}$$

where f_{Y_S} is the common marginal density for the $\boldsymbol{y}$'s belonging to S given that $S \in \rho$. We can define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$ where $\boldsymbol{\theta}_i = \boldsymbol{\theta}^S$ when $i \in S$, the common parameter of f_{Y_S} for $\boldsymbol{y}$'s belonging to S.

Example 1.2. *Normal-gamma PPM*

$$\begin{aligned} p(y_1, ..., y_n|\rho, \boldsymbol{\mu}, \boldsymbol{\tau}) &\propto \prod_{i=1}^{n} N(y_i|\mu_i, \tau_i) \\ \left(\mu^S, \tau^S\right)_{S\in\rho}|\rho &\overset{ind}{\sim} Ng\left(\cdot|a_0^S, b_0^S, \alpha_0^S, \beta_0^S\right) \\ p(\rho|c(S) \text{ with } S \subseteq S_0) &\propto \prod_{S\in\rho} c(S) \end{aligned} \tag{1.4}$$

where Ng is the normal-gamma distribution (see Appendix B for more details), $\left(a_0^S, b_0^S, \alpha_0^S, \beta_0^S\right)$ are hyperparameters associated to S given that $S \in \rho$. We name this model the normal-gamma product partition model (NGPPM).

In many applications, covariates are available and can be used in clustering. In Müller and Quintana (2010) and Müller et al. (2011) the authors propose a generalization of the PPM, introducing covariates as follows. Let

$$P(\rho = \{S_1, \ldots, S_k\}) \propto \prod_{i=1}^{k} g(\boldsymbol{x}_{S_i})c(S_i), \tag{1.5}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ denotes the entire set of recorded covariates (x_i is the covariate of the *i-th* observation) and $\boldsymbol{x}_S = \{x_i, i \in S\}$. The function $g(\boldsymbol{x}_S)$ denotes a non-negative function of $\boldsymbol{x}_S$ that formalizes the similarity of the x_i's, with larger values $g(\boldsymbol{x}_S)$ for sets of covariates that are judged to be similar. Another extension is presented by Park and Dunson (2010).

1.3 PPMs and Dirichlet process

Definition 1.3. *The Dirichlet process, introduced by Ferguson (1973), is a stochastic process that can be thought of as a probability whose domain is the space of probability measures on $\mathcal{X}$.*
Let $(\mathcal{X}, \mathcal{F}, \nu)$ an space with $\nu : \mathcal{F} \to [0,1]$ be a probability measure and α be a positive real number; then a stochastic process P indexed by elements $B \in \mathcal{F}$ is said to be a Dirichlet process on $(\mathcal{X}, \mathcal{F})$ with parameter ν if, for any partition $(B_1, ..., B_n)$ with $B_i \in \mathcal{F}$, the random vector $(P(B_1) \cdots P(B_n))$ has a Dirichlet distribution with parameter $(\alpha\nu(B_1), \cdots, \alpha\nu(B_n))$. We will denote such process by $DP(\alpha, \nu)$, with base measure ν and dispersion parameter α.

Remark 1.4. *When we deal with $\mathcal{X} = \mathbb{R}^p$ we will use the notation G for the base measure; when dealing with more complex spaces, we will retain the notation ν.*

We state some useful properties of the DP in the following proposition:

Proposition 1.5. *(Ferguson, 1973)*
Let $P \sim DP(\alpha, \nu)$ then
a) *$E(P) = \nu$*
b) *If ψ is a P-integrable function, then $E\left(\int \psi dP\right) = \int \psi d\nu$. This holds for indicator functions from the relation $E(G(A)) = \nu(A)$, and then standard measure theoretic arguments extend this sequentially to simple measurable functions, nonnegative measurable functions and finally to all integrable functions.*
c) Conjugacy: *Let $\pi_1, ..., \pi_n$ be a sequence of independent draws from P then*

$$P|\pi_1, ..., \pi_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n}\nu + \frac{n}{\alpha + n}\frac{\sum_{i=1}^{n} \delta_{\{\pi_i\}}}{n}\right).$$

Another approach to the Dirichlet process is using the stick breaking representation of the elements of the $DP(\alpha, G)$ provided by Sethuraman (1994).

Proposition 1.6. *Let $H \in DP(\alpha, G)$, then, with probability one, H has the following form:*

$$H(\bullet) = \sum_{i=1}^{\infty} \pi_i \delta_{X_i}(\bullet) \tag{1.6}$$

where $\pi_i = \beta_i \prod_{k=1}^{i-1}(1 - \beta_k)$ with $\beta_k \overset{iid}{\sim} Beta(1, \alpha)$

We now describe the generalized Pólya urn scheme proposed by Blackwell and MacQueen (1973). Let $\pi_1, \pi_2, ...$ be a infinite sequence of random elements on (the Borel sets of) a complete separable metric space $\mathcal{X}$ with

$$\mathbb{P}(\pi_1 \in \cdot) = \nu(\cdot)$$

and

$$\mathbb{P}(\pi_{i+1} \in \cdot | \pi_1, ..., \pi_i) = \nu_i(\cdot) \text{ with } i \geq 1,$$

where

$$\nu_i(\cdot) = \frac{\alpha}{\alpha + i}\nu(\cdot) + \frac{i}{\alpha + i}\frac{\sum_{j=1}^{i} \delta_{\{\pi_j\}}(\cdot)}{i} \tag{1.7}$$

$\alpha \in \mathbb{R}^+$, and ν is a probability measure on $\mathcal{X}$. Blackwell and MacQueen (1973) showed that

a) $\nu_i(\cdot)$ converges almost surely to a discrete random measure ν^*.

b) ν^* is the Dirichlet process with base measure ν and concentration parameter α.

c) Given ν^*, $\pi_1, \pi_2....$ are independent with distribution ν^*.

Conversely :

Let $\pi_1, \pi_2, ...$ be a random sample from P with P is taken from a the Dirichlet process with parameters ν and α, i.e.

$$\begin{aligned} (\pi_i|P) &\overset{\text{i.i.d.}}{\sim} P, \\ P &\sim \nu^*. \end{aligned}$$

By Proposition 1.5 c), the posterior of ν^* based on the first i observations $\pi_1, ..., \pi_i$ is also a Dirichlet process, but with an updated parameters $\alpha + i$ and ν_i. Therefore

$$\begin{aligned} \mathbb{P}(\pi_{i+1} \in \cdot | \pi_1, ..., \pi_i) &= \int \mathbb{P}(\pi_{i+1} \in \cdot | \pi_1, ..., \pi_i, P)\, \nu^*(dP | \pi_1, ..., \pi_i) \\ &= \int P(\cdot) \nu^*(dP | \pi_1, ..., \pi_i) \\ &= \nu_i(\cdot) \end{aligned}$$

Note that $\mathcal{X}$ can be the space of distributions functions on $\mathbb{R}^p$ or even a more complex space of objects such as stochastic processes, provided with the L_2 metric, or random measures such as the Dirichlet process. The nested Dirichlet process (NDP) defined by Rodríguez et al. (2008) is an example to such construction.

1.3.1 The cluster property of the DP

Let $\pi_1, ..., \pi_n$ a random sample of the Dirichlet process with parameters (α, ν). Using equation (1.7), we can see that, with positive probability, we will have repetitions. Matching the indices of unique values of $\pi_1, ..., \pi_n$ we induce a partition of $S_0 = \{1, ..., n\}$. For instance, if we obtain $\pi_1 = 0.28$, $\pi_2 = 4.3$, $\pi_3 = 0.28$, $\pi_4 = 3.1$ and $\pi_5 = 4.3$, since $\pi_1 = \pi_3$, $\pi_2 = \pi_5$ we obtain the partition $\{\{1, 3\}, \{2, 5\}\{4\}\}$. Given that $\pi_1, ..., \pi_n$ are random, this induces a random partition of S_0. This random partition, in fact, encapsulates all the properties of the DP. To see this, we simply invert the generative process; starting from the distribution over partitions, we can reconstruct the joint distribution $\mathbb{P}(\pi_1, ..., \pi_n) = \prod_{i=1}^{n} \mathbb{P}(\pi_i | \pi_{i-1}, ..., \pi_1)$ by first drawing a random partition of S_0, then for each cluster j in the partition, draw a $\pi_k^* \sim \nu$, and finally assign $\pi_i = \pi_k^*$ for each i in cluster j.

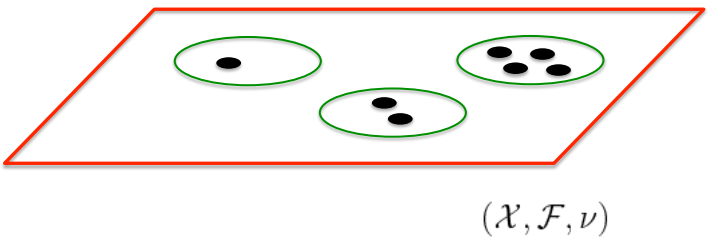


Figure 1.1: Cluster structure induced by the Dirichlet process on a random measure space $(\mathcal{X}, \mathcal{F}, \nu)$.

In Figure 1.1 we show an example of the cluster structure induced by the Dirichlet process on a random measure space.

Quintana and Iglesias (2003) pointed out an interesting connection between parametric PPMs and the Dirichlet process. Because of its clustering properties, it is not surprisingly that we can obtain a particular case of the PPM by integrating out the DP. This result is formally presented in the following proposition for a general space $\mathcal{X}$.

Proposition 1.7. *Assume the following model*

$$\begin{aligned} \pi_1, ..., \pi_n | P &\overset{i.i.d.}{\sim} P \\ P &\sim DP(\alpha, \nu), \end{aligned}$$

where α is a dispersion parameter and ν is the base measure of the DP. By integrating out the DP, the Pólya urn representation of Blackwell and MacQueen (1973) implies

$$\mathbb{P}(\pi_1, ..., \pi_n) = \prod_{i=1}^{n} \left\{ \frac{\alpha\nu(\pi_i) + \sum_{j<i} \delta_{\{\pi_i\}}(\pi_j)}{\alpha + i - 1} \right\}, \tag{1.8}$$

where $\delta_A(x) = 1$ if $x \in A$ and $\delta_A(x) = 0$ otherwise.

This joint marginal distribution can be expressed alternatively as follows. For a given partition $\rho = \{S_1, ..., S_k\}$, let $e_{j,1} < ... < e_{j,|S_j|}$ denote the elements of S_j, assumed to be sorted in ascending order. By the combinatorial arguments developed in Lo (1984) it follows that equation (1.8) can be expressed as

$$\begin{aligned} \mathbb{P}(\pi_1, ..., \pi_n) &= \sum_{\rho \in \mathcal{P}} \frac{\alpha^{|\rho|}}{\prod_{l=1}^{n}(\alpha + l - 1)} \prod_{j=1}^{|\rho|} (|S_j| - 1)! \nu(\pi_{e_{j,1}}) \prod_{i=2}^{|S_j|} \delta_{\{\pi_{e_{j,i}}\}} \left(\pi_{e_{j,1}}\right) \\ &= K \sum_{\rho \in \mathcal{P}} \prod_{j=1}^{|\rho|} c(S_j) p_{S_j}(\boldsymbol{\pi}_{S_j}), \end{aligned} \tag{1.9}$$

where $\mathcal{P}$ is the set of all partitions, $K = \prod_{l=1}^{n}(\alpha + l - 1)$, $c(S) = \alpha \times (|S| - 1)!$, $\boldsymbol{\pi}_{S_j} = (\pi_i | i \in S_j)$, the blocks $\boldsymbol{\pi}_{S_1}, ..., \boldsymbol{\pi}_{S_{|\rho|}}$ are independent and $p_{S_j}(\boldsymbol{\pi}_{S_j})$ is defined as the probability such that all the elements in $\boldsymbol{\pi}_{S_j}$ are identical to a value drawn from ν. Because equation (1.9) is identical to the marginal distribution that is obtained through equations (1.2) and (1.3) for the choices just described, the integrated-out nonparametric model can be seen as a special case of a PPM. This proposition is very important for application purposes because it allows us to simulate this particular case of PPMs more efficiently.

1.4 PPMs for change-point analysis

There is an increasing interest among statisticians in the area of change-point analysis; this has been triggered by an awareness of important applications such as text segmentation, detection of genes causing an abnormality, change-points in economic models,

detection of discontinuities in geophysics time series, etc. New theoretical and computational methods also contributed to the development of this area. This is the case of product partition models, since one of their main application has been in change-point analysis. In this section, we will describe the model for such purpose and discuss the limitations of the marginal probability criterion to detect changes point. We also provide an alternative approach using loss functions.

In one-dimensional change-point problems, the goal is to partition the sequence of observations $x_1, ..., x_i, ..., x_n$ (ordered by index i) into b contiguous subsequences or blocks,

$$[x_1, ..., x_{i_1}], [x_{i_1+1}, ..., x_{i_2}], ..., \left[x_{i_{b-1}+1}, ..., x_{i_b}\right].$$

Let f_i be the density function of x_i, parametrized by $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$ (whose value may change from one observation to the next). We suppose that there exists a partition ρ of the set $\{1, ..., n\}$ into contiguous sets or blocks such that the sequence $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ is constant within blocks; that is, there exists a partition $\rho = (i_0, i_1, ..., i_b)$ of the set $\{1, 2, ..., n\}$ such that

$$0 = i_0 < i_1 < i_2 < \cdots < i_b = n$$

and

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i_r} \text{ with } i_{r-1} < i \leq i_r$$

for $r = 1, 2, ..., b$. The parameter values change at the change points $i_1 + 1, i_2 + 1, ..., i_{b-1} + 1$. We denote the observations $x_{i+1}, ..., x_j$ by $\boldsymbol{x}_{ij}$. Let $f_{ij}\left(\boldsymbol{x}_{ij}|\boldsymbol{\theta}^{ij}\right)$ be the joint density of $\boldsymbol{x}_{ij}$ given $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_{i+2} = ... = \boldsymbol{\theta}^{ij}$. (The notation $f(\cdot)$ will be used for densities, and $f(\cdot|\cdot)$ for conditional densities.) The observations $x_1, ..., x_n$ are assumed independent between different blocks, given the partition and given the parameters. Then we have

$$f_{0n}\left(\boldsymbol{x}_{0n}|\rho, \boldsymbol{\theta}\right) = \prod_{j=1}^{b} f_{i_{j-1}i_j}\left(\boldsymbol{x}_{i_{j-1}i_j}|\boldsymbol{\theta}^{i_{j-1}i_j}\right)$$

where, as before, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$. The partition is selected randomly according to a product partition distribution. The probability of a partition $\rho = (i_0, i_1, i_2, ..., i_b)$ is

$$p(\rho) \propto c_{i_0 i_1} c_{i_1 i_2} \cdots c_{i_{b-1} i_b},$$

where c_{ij} is known as a cohesion function and is specified for each possible block ij. We will define a new level of this hierarchical model, a prior distribution of $\boldsymbol{\theta}$.

Given the partition ρ with b blocks, $\boldsymbol{\theta}_{i_1}, \boldsymbol{\theta}_{i_2}, ..., \boldsymbol{\theta}_{i_b}$ are independent. Let $f_{i_{j-1}i_j}\left(\boldsymbol{\theta}_{i_{j-1}i_j}\right)$ be the density of $\boldsymbol{\theta}_{i_{j-1}i_j} = (\theta_{i_{j-1}+1}, ..., \theta_{i_j})$. Because all parameters in the block $i_{j-1}i_j$ are equal to $\boldsymbol{\theta}^{i_{j-1}i_j}$, the joint distribution of all the parameters is now determined. The density $f_{ij}\left(\boldsymbol{\theta}^{ij}\right)$ will be called the block ij prior density.

For a given block ij, the predictive density of the observations $\boldsymbol{x}_{ij}$ is

$$f_{ij}\left(\boldsymbol{x}_{ij}\right) = \int f_{ij}\left(\boldsymbol{x}_{ij}|\boldsymbol{\theta}^{ij}\right) f_{ij}\left(\boldsymbol{\theta}^{ij}\right) d\boldsymbol{\theta}^{ij}.$$

The joint probability density of $\boldsymbol{x}_{ij}$, denoted by $f_{\boldsymbol{x}_{ij}}$, belongs to a parametric family and the previous hierarchical model defined above includes a level specifying a prior for the parameters. For a graphical representation of the PPM, see Figure 1.2.

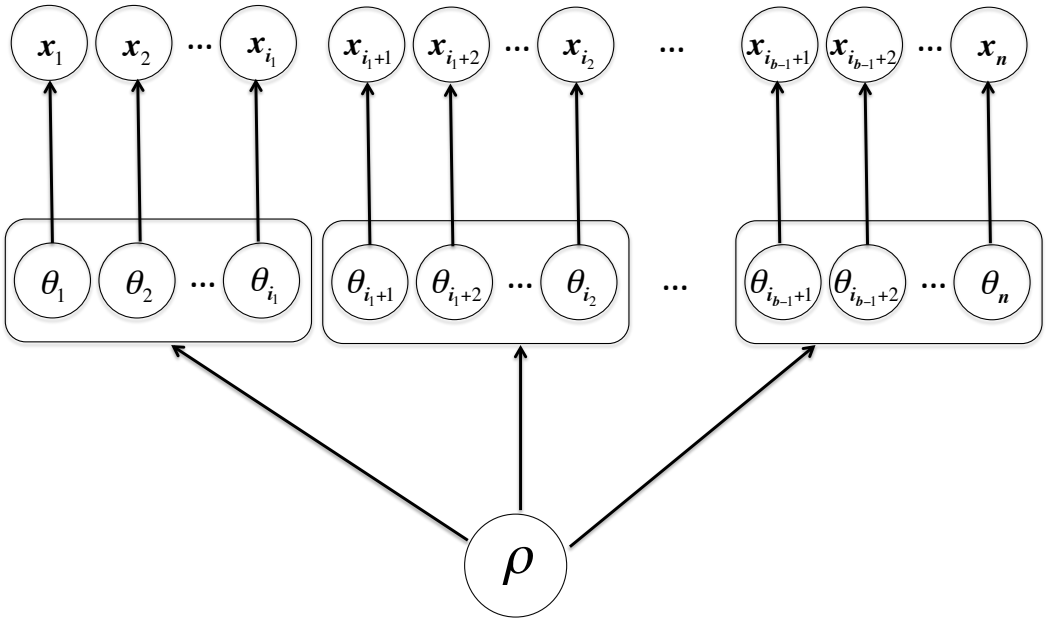


Figure 1.2: Graphical representation of the PPM for change-point detection.

We now describe the normal-gamma example studied by Loschi et al. (2003). We will compare it later with the NPPMs for change-point analysis.

Example 1.8. *Normal-Gamma model for change-point analysis.*

$$\begin{aligned} x_1,...,x_n|\rho,\mu_1,...,\mu_n,\tau_1,...,\tau_n &\sim \prod_{i=1}^{n} N(x_i|\mu_i,\tau_i) \\ \left(\mu^{i_{j-1}i_j},\tau^{i_{j-1}i_j}\right)_{j=1,...,b}|\rho &\overset{ind}{\sim} Ng\left(\cdot|a_0^{i_{j-1}i_j},b_0^{i_{j-1}i_j},\alpha_0^{i_{j-1}i_j},\beta_0^{i_{j-1}i_j}\right) \\ p(\rho=\{i_0,i_1,...,i_b\}) &\propto c_{i_0i_1}c_{i_1i_2}\cdots c_{i_{b-1}i_b} \end{aligned} \tag{1.10}$$

where $\left(a_0^{i_{j-1}i_j},b_0^{i_{j-1}i_j},\alpha_0^{i_{j-1}i_j},\beta_0^{i_{j-1}i_j}\right)$ *are hyperparameters associated with block* $i_{j-1}i_j$. *A well known result gives*

$$\begin{aligned} p(\mu^{ij},\tau^{ij}|\rho,\boldsymbol{x}_{ij}) &= Ng(\mu^{ij},\tau^{ij}|a^{ij},b^{ij},\alpha^{ij},\beta^{ij}) \\ a^{ij} &= \frac{b_0^{ij}a_0^{ij}+n_{ij}\overline{\boldsymbol{x}}_{ij}}{b_0^{ij}+n_{ij}} \\ b^{ij} &= b_0^{ij}+n_{ij} \\ \alpha^{ij} &= \alpha_0^{ij}+\frac{n_{ij}}{2} \\ \beta^{ij} &= \beta_0^{ij}+\frac{1}{2}\sum_{h=i+1}^{j}(\boldsymbol{x}_h-\overline{\boldsymbol{x}}_{ij})^2+\frac{b_0^{ij}n_{ij}(\overline{\boldsymbol{x}}_{ij}-a_0^{ij})^2}{2(b_0^{ij}+n_{ij})}, \end{aligned} \tag{1.11}$$

where n_{ij} *is the number of data points in block* ij *and* $\overline{\boldsymbol{x}}_{ij}$ *is the mean of the data points belonging to block* ij. *The posterior marginals are*

$$\begin{aligned} p(\tau^{ij}|\rho,\boldsymbol{x}_{ij}) &= Ga(\tau^{ij}|\alpha^{ij},\beta^{ij}) \\ p(\mu^{ij}|\rho,\boldsymbol{x}_{ij}) &= t_{2\alpha^{ij}}(\mu^{ij}|a^{ij},\beta^{ij}/(\alpha^{ij}b^{ij})), \end{aligned} \tag{1.12}$$

where $t_\nu(\boldsymbol{\mu},\boldsymbol{V})$ *denotes the non-standardized Student's t-distribution. Hence*

$$\begin{aligned} E(\tau^{ij}|\rho,\boldsymbol{x}_{ij}) &= \frac{\alpha^{ij}}{\beta^{ij}} \\ E(\mu^{ij}|\rho,\boldsymbol{x}_{ij}) &= a^{ij}. \end{aligned} \tag{1.13}$$

The predictive distribution is given by $\mathbf{x}\sim t_{2\alpha_0^{ij}}\left(\mathbf{a}_0^{ij},(\alpha_0^{ij})^{-1}\beta_0^{ij}\left(\boldsymbol{I}_n+\frac{H}{b_0^{ij}}\right)\right)$ *with* $\mathbf{a}_0^{ij}\in\mathbb{R}^{n_{ij}}$ *where* $\mathbf{a}_0^{ij}=(a_0^{ij},...,a_0^{ij})$, $\boldsymbol{I}_n$ *denotes the* $n\times n$ *identity matrix and* H *is an* $n\times n$ *matrix such that* $H_{ij}=1\ \forall i,j$ *(see Appendix B for more details).*

1.4.1 Cohesions functions

The selection of adequate cohesions functions is an important issue in PPMs. In this subsection, we will review several alternatives proposed in the literature. Barry and Hartigan (1992) considered a model in which the partition distribution used cohesions of the form:

$$\begin{aligned} c_{ij} &= (j-i)^{-3} \text{ for } 0 < i < j < n, \\ c_{ij} &= (j-i)^{-2} \text{ for } i = 0 \text{ or } j = n, \\ \text{and } c_{0n} &= n^{-1}. \end{aligned}$$

They also proved some desirable consistency properties of this choice of cohesions.

Inspired by previous work of Yao (1984), Barry and Hartigan (1993) used the following cohesion functions, which imply that the sequence of change points forms a discrete renewal process with inter-arrival times identically and geometrically distributed:

$$\begin{aligned} c_{ij} &= (1-p)^{j-i-1}p \text{ if } j < n, \\ c_{ij} &= (1-p)^{j-i-1}, \text{ if } j = n, \end{aligned}$$

where p denotes the probability that a change occurs at any instant in the sequence. Note that c_{ij} corresponds to the probability that a new change takes place after $j-i$ instants, given that a change has taken place at the instant i. Such cohesions are appropriate when it is reasonable to assume that the past change points are noninformative about the future change points, which is sensible in many practical applications (Loschi and Cruz, 2005).

Monteiro et al. (2011) studied some properties of the number of clusters (denoted by C) in the partition ρ when $p \sim Beta(a, b)$. Quintana et al. (2005a) profit from the relationship between the Dirichlet process and cohesions functions of the form $c(S) \propto \alpha \times |S-1|!$. More generally, cohesions functions depending only on the cardinality of the set $j-i$ can be induced by Gibbs-type priors if and only if $c_{ij} = \frac{(1-\sigma)_{j-i-1}}{(j-i)!}$ for $0 \leq i < j \leq n$ and $\sigma \in [-\infty, 1]$ with $(1-\sigma)_{j-i-1} = 1$ if $\sigma = -\infty$ and $\rho = (0, n)$ when $\sigma = 1$ (Blasi et al., 2015). This choice of cohesions leads to exchangeable blocks, which can be very useful in many applications such as detecting gain or loss of material in DNA sequences. However, there are situations in which the exchangeability assumption is not adequate. For instance, in the context of Global warming, if we want to detect

change of behaviour in the average temperature since 1900, then it is not reasonable to suppose that the probability of partition $[1900, 1980][1981, 2014]$ is the same that $[1900, 1933][1934, 2014]$ because there were more factories, cars, etc. in 1981 than in 1934.

For a non informative prior, we can use the uniform distribution giving equal weights to all possible partitions.

1.4.2 Inference

Although Barry and Hartigan (1992, 1993) introduced PPMs for change-point analysis and studied some consistency properties, they did not provide any methodology to detect the points in time where a shift occurs. Indeed, they introduced their model only for prediction purposes. Later, Loschi and Cruz (2005) used the marginal probability of change point to decide whether a point in time is a changing point or not. In this criterion, we fixed a probability p_0, for each point i we calculate the marginal probability of being a change point p_i. If $p_0 > p_i$ then x_i is a change point.

Quintana et al. (2005a) provide a different approach using a weighted loss function

$$L\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\hat{\rho}}\right) = \gamma \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\hat{\rho}}\|^2 + (1-\gamma)|\hat{\rho}|,$$

where $\hat{\boldsymbol{\theta}}_{\hat{\rho}}$ is an estimate of $\boldsymbol{\theta}$ associated with the selected partition $\hat{\rho}$ and $0 \leq \gamma \leq 1$. Quintana and Iglesias (2003) show that the expected loss minimization criterion leads to the choice $\hat{\rho}^*$ that minimizes

$$SC(\hat{\rho}) = \gamma \|\hat{\boldsymbol{\theta}}_B(\boldsymbol{x}) - \hat{\boldsymbol{\theta}}_{\hat{\rho}}(\boldsymbol{x})\|^2 + (1-\gamma)|\hat{\rho}|, \tag{1.14}$$

where $\boldsymbol{x} = (x_1, ..., x_n)$, $\hat{\boldsymbol{\theta}}_B(\boldsymbol{x}) = E(\boldsymbol{\theta}|\boldsymbol{x})$ and $\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{x}) = E(\boldsymbol{\theta}|\boldsymbol{x}, \rho)$ (see Appendix A). The above result shows that the optimal choice $\hat{\rho}^*$ will be the partition for which the resulting estimate $\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{x})$ is closest to $\hat{\boldsymbol{\theta}}_B(\boldsymbol{x})$, penalized by the number of clusters. To identify the change points, Quintana et al. (2005a) propose the following strategy with SC defined by equation (1.14).

Algorithm 1.1 Change-point detection

The basic procedure consists of recursively assessing subsequences of the set $\{1, \ldots, n\}$ and identifying change points by splitting each subsequence into two parts.

Step 1: Set $\mathcal{C} = \varnothing$, $l = 1$, $u = n$, $\hat{\rho}^* = \{1, \ldots, n\}$.

Step 2: In the current partition $\hat{\rho}^*$, split the set $\{l, \ldots, u\}$ into $\{l, \ldots, j-1\}$ and $\{j, \ldots, u\}$ for $j = l+1, \ldots, u$. Denote by $\overline{\rho}_j$ the corresponding partition. Let k^* be defined as $SC\left(\overline{\rho}_{k^*}\right) = \min_{l+1 \leq j \leq u-1} SC(\overline{\rho}_j)$.

Step 3: If $SC\left(\overline{\rho}_{k^*}\right) < SC\left(\hat{\rho}^*\right)$ then add k^* to $\mathcal{C}$, replace $\hat{\rho}^*$ by $\overline{\rho}_{k^*}$, and recursively repeat Step 2 for $l = l$, and $u = k^* - 1$ and for $l = k^*, u = u$. Otherwise, stop.

Yau and Holmes (2013) exposed the limitations of using the marginal probability of change point or the most probable state sequence in hidden Markov models.

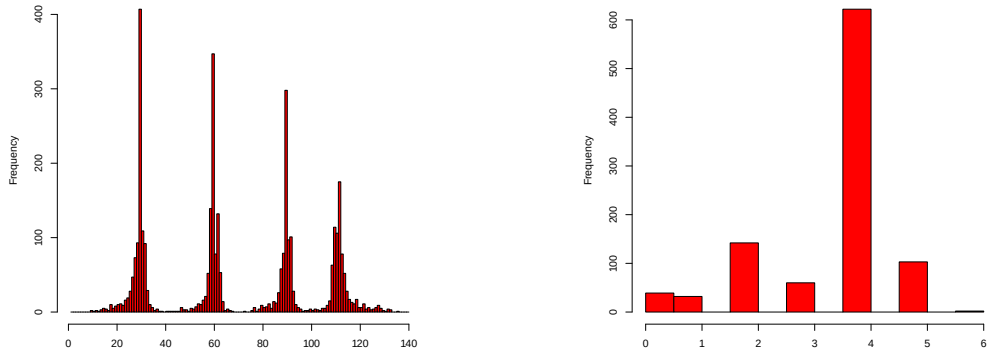
To see this, we analyze the following example using PPMs with cohesions functions of the form: $c_{ij} = (1 - p_0)^{j-i-1} p_0$.

Example 1.9. *We simulated the following sequence of random variables* 1000 *times:*

$$X_i \sim \begin{cases} Bernoulli(0.1), & \text{if } 1 \leq i \leq 30 \\ Bernoulli(0.8), & \text{if } 31 \leq i \leq 60 \\ Bernoulli(0.1), & \text{if } 61 \leq i \leq 90 \\ Bernoulli(0.8), & \text{if } 91 \leq i \leq 110 \\ Bernoulli(0.2), & \text{if } 111 \leq i \leq 140. \end{cases}$$

This sequence of data points contains four change points at $\{31, 61, 91, 111\}$*. At each iteration, we used the loss function criterion with* $\gamma = 0.7$ *and the probability of change point for several values of* p_0 *(*$p_0 = 0.67, 0.69, 0.70$*). The results are shown in Figures 1.3 and 1.4 .*

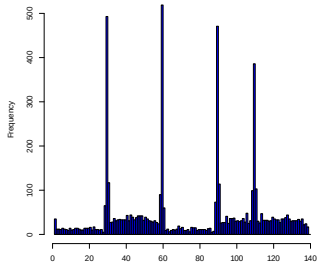
Comparing the histograms of the number of change points detected by the different procedures with different parameters in Figures 1.3b, 1.4b, 1.4d and 1.4f, we obtain more accurate results using a loss function. Indeed, the results obtained using the probability criterion are quite poor. The histograms of change points locations are presented in Figures 1.3a, 1.4a, 1.4c and 1.4e. As before, results obtained using a loss function are better.



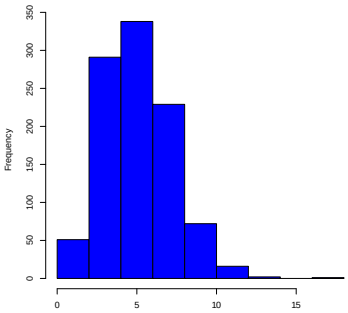
(a) Change points detected.

(b) Number of change points.

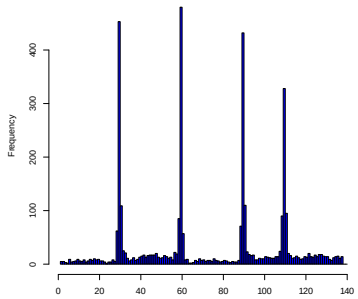
Figure 1.3: Change point analysis of Example 1.9 using the loss function criterion with $\gamma = 0.7$.



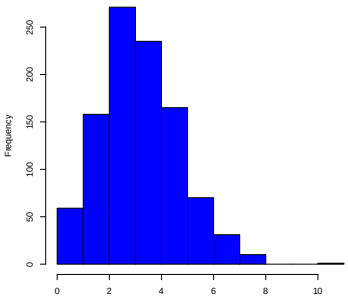
(a) Change points detected with $p_0 = 0.67$.



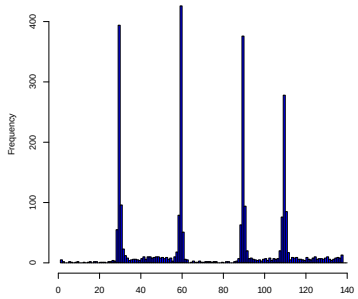
(b) Number of change points with $p_0 = 0.67$.



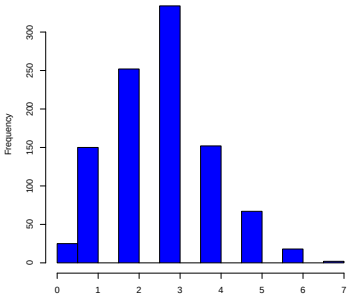
(c) Change points detected with $p_0 = 0.69$.



(d) Number of change points with $p_0 = 0.69$.



(e) Change points detected with $p_0 = 0.7$.



(f) Number of change points with $p_0 = 0.7$.

Figure 1.4: Change point analysis of Example 1.9 using the change point probability criterion with different values of p_0.

.

Chapter 2

Nonparametric Product Partition Models

The woods would be very silent if no birds sang except those that sang best.

Rabindranath Tagore

In this chapter we propose an extension of parametric product partition models (PPMs) introduced by Hartigan (1990) and Barry and Hartigan (1992), using ideas presented by Quintana and Iglesias (2003). We name our proposal nonparametric product partition models (NPPMs) because we associate a random measure instead of a parametric kernel to each set within a random partition. Our methodology does not impose any specific form on the marginal distribution of the observations, allowing us to detect shifts of behaviour even when dealing with heavy-tailed or skewed distributions.

We propose a suitable loss function and find the partition of the data having minimum expected loss. We then apply our nonparametric procedure to multiple change-point analysis and compare it with PPMs and with other methodologies that have recently appeared in the literature. Also, in the context of missing data, we exploit the product partition structure in order to estimate the distribution function of each missing value, allowing us to detect change points using the loss function mentioned above. Finally, we present applications to financial as well as genetic data.

2.1 Definition of nonparametric product partition models.

2.1.1 Motivation

Let Y be a discrete finite random variable. In the context of finite population sampling, we want to find the stratification or partition $\rho = \{H_1, \ldots, H_k\}$ of the range of Y such that

$$f(x) = f(x|Y \in H_1)P(Y \in H_1) + \ldots + f(x|Y \in H_k)P(Y \in H_k) \tag{2.1}$$

minimizes the loss function

$$L(\mu_\rho, \mu) = \beta\|\mu_\rho - \mu\|^2 + (1-\beta)|\rho|$$

where $\mu = E(X)$ and μ_ρ is the estimate of μ associated with the partition or stratification ρ, i.e.

$$\mu_\rho = E(X|Y \in H_1)P(Y \in H_1) + \ldots + E(X|Y \in H_k)P(Y \in H_k). \tag{2.2}$$

Parametric product partition models are not flexible enough to represent the situation of equation (2.1) because it is assumed that for all strata H, $f(x|Y \in H)$ belongs to the same parametric family of distributions. Therefore, we need a more flexible family of product partition models. It would be convenient to model $f(x|Y \in H)$ non-parametrically.

Now, we define nonparametric product partition models. For any partition $\rho = \{S_1, ..., S_k\}$ of $S_0 = \{1, ..., n\}$ and data $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, it is assumed that

$$\begin{aligned} p\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_n \mid \rho, F_X^{S_j}\right) &\propto \prod_{i=1}^{n} F_i(\boldsymbol{x}_i) \\ F_X^{S_j} \mid \rho &\overset{ind}{\sim} DP\left(\alpha^{S_j}, G^{S_j}\right) \\ P\left(\rho = \{S_1, ..., S_k\}\right) &\propto \prod_{i=1}^{k} c(S_i) \end{aligned} \tag{2.3}$$

where $\alpha^{S_j} > 0$ and G^{S_j} is a probability distribution function. In the same manner as the PPMs, $c(S)$ is a nonnegative set function called cohesion function. .

Remark 2.1. *We can write* $\prod_{i=1}^{n} F_i(\boldsymbol{x}_i) = \prod_{j=1}^{k} F_X^{S_j}(x_{S_j})$ *where* k *is the number of clusters.*

NPPMs, just as PPMs, induce a probability distribution over all possible partitions of a finite set of observations: $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ with $\boldsymbol{x}_i \in \mathbb{R}^p$. For each partition, the n data points are divided into k subsets, and each data point $\boldsymbol{x}_i$ belongs to only one subset. As before, data points of distinct subsets are independent and data points belonging to the same subset are exchangeable.

Remark 2.2. *Since the entries in $\boldsymbol{x}_S$ are exchangeable, if $F^S_{\boldsymbol{X}}$ is the joint distribution function of $\boldsymbol{x}_S$ given ρ, applying de Finetti's theorem (de Finetti 1972), we obtain*

$$\begin{aligned} p_S(\boldsymbol{x}_S) &= \int \left\{ \prod_{i \in S} p_S(\boldsymbol{x}_i | F^S_{\boldsymbol{X}}) \right\} d\mathbb{P}_S(F^S_{\boldsymbol{X}}) \\ &= \int \left\{ \prod_{i \in S} F^S_{\boldsymbol{X}}(\boldsymbol{x}_i) \right\} d\mathbb{P}_S(F^S_{\boldsymbol{X}}) \end{aligned}$$

where $F^S_{\boldsymbol{X}}$ is the common marginal distribution for the $\boldsymbol{x}$'s such that $i \in S$ given that $S \in \rho$. and $\mathbb{P}_S(F^S_{\boldsymbol{X}})$ is the probability measure over the space of distribution functions induced by the Dirichlet Process associated to the set S. We can define $\boldsymbol{F} = (F_1, ..., F_n)$ where $F_i = F^S_{\boldsymbol{X}}$ when $i \in S$.

We now show that the posterior distribution is again a NPPM.

Proposition 2.3. *The corresponding posterior distribution of ρ is again of the form of equation (2.3), with cohesions given by $c(S_j)p_{S_j}(\boldsymbol{x}_{S_j})$ where p_{S_j} is the predictive distribution given in remark 2.2. We define $c(S_j)p_{S_j}(\boldsymbol{x}_{S_j})$ as the posterior cohesions.*

Proof.

$$\begin{aligned} p(\rho | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) &\propto p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \rho) p(\rho) \\ &\propto p_{S_1}(\boldsymbol{x}_{S_1} | \rho) \ldots p_{S_k}(\boldsymbol{x}_{S_k} | \rho) \times \prod_{j=1}^{k} c(S_j) \\ &\propto \int p_{S_1}(\boldsymbol{x}_{S_1} | \rho, F^{S_1}_{\boldsymbol{X}}) d\mathbb{P}(F^{S_1}_{\boldsymbol{X}}) \times \ldots \\ &\quad \times \int p_{S_k}(\boldsymbol{x}_{S_k} | \rho, F^{S_k}_{\boldsymbol{X}}) d\mathbb{P}(F^{S_k}_{\boldsymbol{X}}) \times \prod_{j=1}^{k} c(S_j) \\ &\propto c(S_1) \int \left\{ \prod_{i \in S_1} p_{S_1}(\boldsymbol{x}_i | F^{S_1}_{\boldsymbol{X}}) \right\} d\mathbb{P}(F^{S_1}_{\boldsymbol{X}}) \times \ldots \end{aligned}$$

$$
\begin{aligned}
& \times c(S_k) \int \left\{ \prod_{i \in S_k} p_{S_k}(\boldsymbol{x}_i | F_{\boldsymbol{X}}^{S_k}) \right\} d\mathbb{P}(F_{\boldsymbol{X}}^{S_k}) \\
\propto \quad & c(S_1) \int \left\{ \prod_{i \in S_1} F_{\boldsymbol{X}}^{S_1}(\boldsymbol{x}_i) \right\} d\mathbb{P}(F_{\boldsymbol{X}}^{S_1}) \times \ldots \\
& \times c(S_k) \int \left\{ \prod_{i \in S_k} F_{\boldsymbol{X}}^{S_k}(\boldsymbol{x}_i) \right\} d\mathbb{P}(F_{\boldsymbol{X}}^{S_k})
\end{aligned}
$$

We can recognize that $\int \left\{ \prod_{i \in S_k} F_{\boldsymbol{X}}^{S_k}(\boldsymbol{x}_i) \right\} d\mathbb{P}(F_{\boldsymbol{X}}^{S_k})$ is the predictive distribution evaluated at $\boldsymbol{x}_{S_j}$. As in the parametric case, the posterior distribution has cohesion functions $c(S_j)p(\boldsymbol{x}_{S_j})$, where $p(\boldsymbol{x}_{S_j})$ is the predictive distribution with respect to the Dirichlet process (with concentration parameter α^{S_j} and base measure G^{S_j}). □

Using the relationship between PPM and the Dirichlet process established by Quintana and Iglesias (2003), stated in Proposition 1.7, we will show that using the NDP in Definition 2.4, we can obtain a special case of NPPMs with cohesions functions $c(S) = \alpha \times (|S| - 1)!$. This choice of cohesions functions promotes few clusters with large amounts of data in each one. This can be very restrictive in many applications, although Bormetti et al. (2012); De Giuli et al. (2010); Quintana and Iglesias (2003) and Quintana et al. (2005a,b) exploited this feature for outlier detection.

2.1.2 Relationship with the Nested Dirichlet Process

We will discuss the relationship of the NPPMs with the Nested Dirichlet Process.

2.1.2.1 Nested Dirichlet Process (NDP)

We will describe the Nested Dirichlet Process (NDP) stating the two characterizations provided by Rodríguez et al. (2008) and discuss if they are equivalent or not.

Definition 2.4. *We say that G_j is a $NDP(\alpha, \beta, H)$ if*

$$
\begin{aligned}
G_j | Q & \overset{iid}{\sim} Q \\
Q & \sim DP(\alpha, DP(\beta, H)).
\end{aligned}
$$

Notice that the term $DP(\alpha, DP(\beta, H))$ has a profound meaning and is not just a matter of notation. In order to clarify this, we present the following discussion that appears in Rodríguez et al. (2008). Consider the probability space $(\mathbf{\Theta}, \mathcal{B}(\mathbb{R}^d), P)$ where $\mathbf{\Theta} \subset \mathbb{R}^d$, $\mathcal{B}$ corresponds to the Borel σ-algebra of subsets of $\mathbb{R}^d$ and P is the probability induced by the distribution function H. Consider the Dirichlet Process $DP(\beta, H)$ with $\beta > 0$. Let $\mathcal{X}$ be the space of probability measures over $(\mathbf{\Theta}, \mathcal{B})$, let $\mathcal{B}(\mathcal{X})$ the Borel sets induced by the sup norm, it is well known that $(\mathcal{X}, \|\cdot\|_\infty)$, is a complete separable metric space. Let ν be the probability induced in $\mathcal{B}(\mathcal{X})$ by $DP(\beta, H)$; then we have a probability space defined by the triplet $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$.

Recall the original definition of the DP introduced by Ferguson (1973) stated in Definition 1.3, the choice of $\mathbf{\Theta} \subset \mathbb{R}^n$ for the base space of the DP is merely a practical one, and the aforementioned results extend in general to any complete and separable metric space $\mathcal{X}$. In particular, because the space of probability distributions is complete and separable under the sup norm, we could have started by taking $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$ as our base space. Therefore, $DP(\alpha, \nu)$ is a random measure over the space of distributions on distributions. We can replace $\nu = DP(\beta, H)$ and we would obtain $DP(\alpha, DP(\beta, H)$

Definition 2.5.

$$G(\bullet) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*(\bullet)} \tag{2.4}$$

$$G_k^*(\bullet) \equiv \sum_{l=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*}(\bullet) \tag{2.5}$$

where $\theta_{lk}^* \sim H$, H *is a probability measure on* $(\mathbf{\Theta}, \mathcal{B})$,
$w_{lk}^* = u_{lk}^* \prod_{s=1}^{l-1}(1 - u_{sk}^*)$ *with* $u_{lk}^* \overset{i.i.d}{\sim} Beta(1, \beta)$,
$\pi_k^* = v_k^* \prod_{s=1}^{k-1}(1 - v_s^*)$ *with* $v_k^* \overset{i.i.d}{\sim} Beta(1, \alpha)$

Figures 2.1 and 2.2 are graphical representations of Definition 2.5.

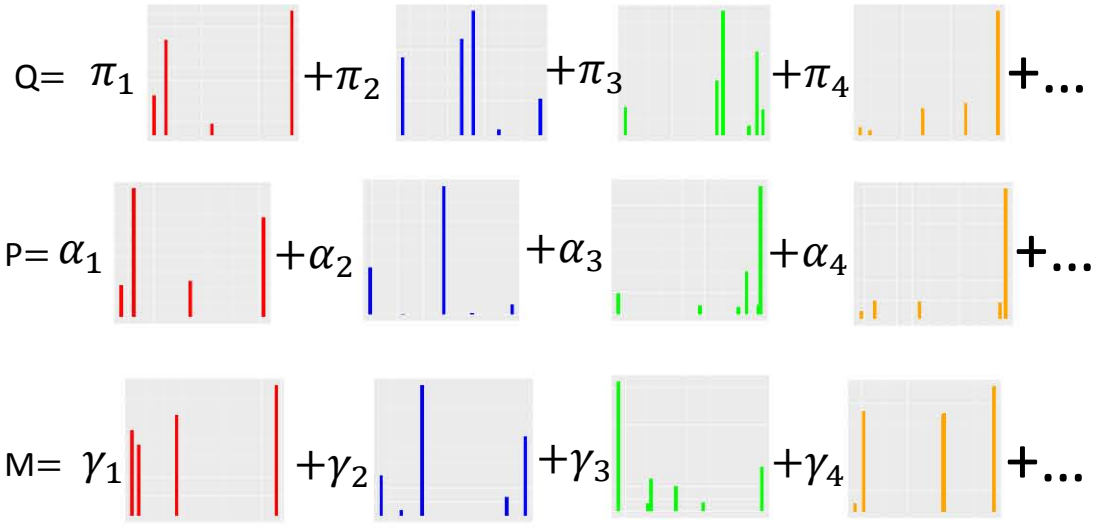


Figure 2.1: $P, Q, M \sim DP(\alpha, DP(\beta, H))$

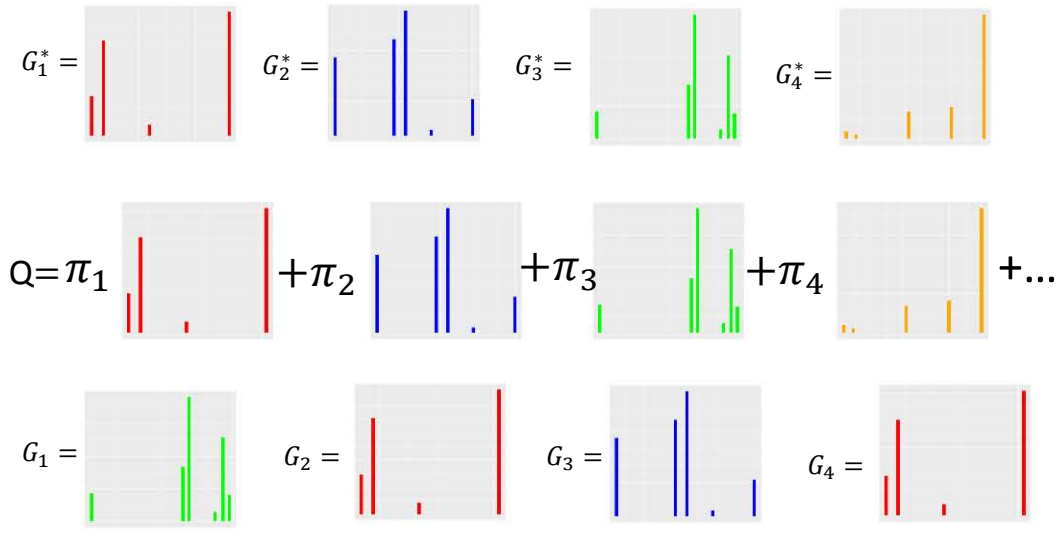


Figure 2.2: $G_i^* \sim DP(\beta, H))$, $Q \sim DP(\alpha, DP(\beta, H))$ and $G_j \sim Q$

These authors claim that the NDP can be characterized as a distribution on the space of distributions on distributions, but Müller and Nieto-Barajas (2008) argued

that this is not the case and it is only an element in the space of distributions on distributions. They also pointed out that the model is better described as random clustering of a set of random distribution. They justify it by noting that the argument in equation (2.4) is θ, not a random distribution. Moreover, if equation (2.4) is changed to

$$G(\bullet) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*}(\bullet) \tag{2.6}$$

then the argument $\bullet$ would be a measurable set of random measures and G_j would be defined on the space of distribution on distributions.

If we analyze Definition 2.4, we obtain a probability measure on the space of distributions on distributions, as stated by Rodríguez et al. (2008), whereas, using Definition 2.5 we obtain a random clustering of a set of random distributions therefore, definitions 2.4 and 2.5 can not be equivalent. In fact, if we change Definition 2.5 to the following definition, the definitions would be equivalent.

Definition 2.6.

$$G(\bullet) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*}(\bullet) \tag{2.7}$$

$$G_k^*(\bullet) \equiv \sum_{l=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*}(\bullet) \tag{2.8}$$

If we marginalize the Nested Dirichlet Process of Definition (2.4), we obtain a special case of the Nonparametric Product Partition model. This result is formalized in the following proposition.

Proposition 2.7. *Let $\mathcal{X} = \{F | F : \mathbb{R}^p \to [0,1]$ is a probability distribution on $\mathbb{R}^p\}$, and let $\|F\|_\infty = \sup_{\boldsymbol{x}} |F(\boldsymbol{x})|$ be the sup norm. It is a well known result that $(\mathcal{X}, \|\cdot\|)$ is a complete separable metric space. Let $\mathcal{B}(\mathcal{X})$ be the Borel sets induced by $\|\cdot\|$. The Dirichlet process $DP(\beta, F_0)$ induces a probability on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with $\beta \in \mathbb{R}^+$ and F_0 a distribution function on $\mathbb{R}^p$. Let ν be the probability measure over $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ induced by $DP(\beta, F_0)$, and consider the following model*

$$\begin{aligned} \boldsymbol{x}_1, ..., \boldsymbol{x}_n | F_1, ..., F_n &\overset{ind}{\sim} F_i(x_i) \\ F_1, ..., F_n | P &\overset{i.i.d.}{\sim} P \\ P | \alpha, \beta, F_0 &\sim DP(\alpha, \nu), \end{aligned}$$

which is equivalent to

$$\begin{aligned} \boldsymbol{x}_1, ..., \boldsymbol{x}_n | F_1, ..., F_n &\overset{ind}{\sim} F_i(x_i) \\ F_1, ..., F_n &\overset{i.i.d.}{\sim} P \\ P|\alpha, \beta, F_0 &\sim NDP(\alpha, \beta, F_0). \end{aligned}$$

If we integrate out P, we obtain the following NPPM

$$\begin{aligned} p\left(\boldsymbol{x}_1, ..., \boldsymbol{x}_n \mid \rho, F_Y^{S_j}\right) &\propto \prod_{i=1}^{n} F_i(\boldsymbol{x}_i) \\ F_X^{S_j} \mid \rho &\overset{ind}{\sim} DP\left(\beta, F_0\right) \\ P\left(\rho = \{S_1, ..., S_k\}\right) &\propto \prod_{i=1}^{k} c(S_i) \end{aligned}$$

with cohesions functions

$$c(S) = \alpha \times (|S| - 1)!.$$

Proof. Straightforward using Proposition 1.7 □

As in the parametric case, this relationship allows us to simulate from this class of NPPMs efficiently using algorithms proposed in Rodríguez et al. (2008) and Müller and Nieto-Barajas (2008).

We raise the following conjecture, which we have not been able to prove,

Conjecture 2.8. *If we marginalize the nested Dirichlet Process provided by Definition 2.5, we obtain a special case of the Nonparametric Product Partition model with cohesion functions of the form $c(S) = \alpha \times (|S| - 1)!$.*

2.2 Loss function

Let $S_j \in \rho$ be such that $i \in S_j$, let $Z_{S_j} \sim F_Z^{S_j}$, and define $\boldsymbol{\mu}_i = E(Z_{S_j})$. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n)$. Notice that $\boldsymbol{\mu}$ is a random vector with $\boldsymbol{\mu}_i \in \mathbb{R}^p$ and $\mu_i = \mu_k$ if $i, k \in S_j$

We define the loss function

$$l\left(\hat{\rho}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\hat{\rho}}\right) = \gamma \left\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{\rho}}\right\|^2 + (1 - \gamma)\left(|\hat{\rho}| - |\rho|\right) \tag{2.9}$$

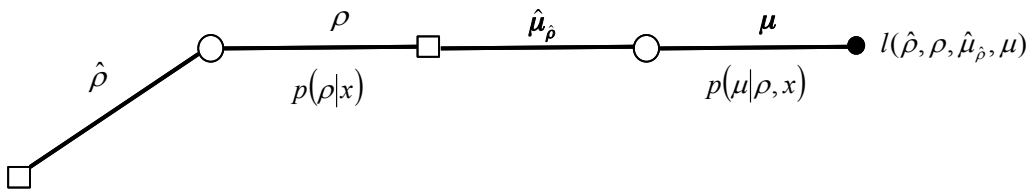
where the vector $\hat{\boldsymbol{\mu}}_{\hat{\rho}}$ is the estimate of the vector $\boldsymbol{\mu}$ associated with the estimated partition $\hat{\rho}$ and $|\hat{\rho}|$ denotes the cardinality of $\hat{\rho}$. Here, $0 \leq \gamma \leq 1$ is a cost-complexity

parameter. The choice of loss function implies a trade-off (controlled by the user-defined quantity γ) between the optimal estimator of $\boldsymbol{\mu}$ and model simplicity, by which we mean a model with a low number of clusters or strata. Notice that the term $(|\hat{\rho}| - |\rho|)$ penalizes when the estimated partition $\hat{\rho}$ contains more clusters than the true partition ρ. We follow the procedure discussed in Quintana and Iglesias (2003) to obtain the partition with the minimum expected loss.

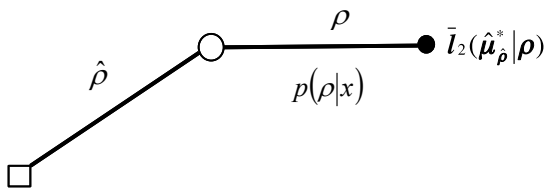
Theorem 2.9. *Let* $\hat{\boldsymbol{\mu}}_B(\boldsymbol{x}) = E(\boldsymbol{\mu}|\boldsymbol{x})$ *and* $\hat{\boldsymbol{\mu}}_\rho(\boldsymbol{x}) = E(\boldsymbol{\mu}|\rho, \boldsymbol{x})$*, then the expected loss minimization criterion leads to the choice* $\hat{\rho}^*$ *that minimizes*

$$SC(\hat{\rho}) = \gamma \|\hat{\boldsymbol{\mu}}_B(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}_\rho(\boldsymbol{x})\|^2 + (1-\gamma)|\hat{\rho}| \tag{2.10}$$

Proof. For a sequential decision problem, Bernardo and Smith (1994) state that one has to first solve the final nth stage by minimizing the appropriate loss function, then one has to solve the $(n-1)th$ stage by minimizing the expected loss function conditional on making the optimal choice at the nth stage; and so on, working backwards progressively, until the optimal first stage option has been obtained. To visualize the decision tree of our problem see Figure 2.3.



(a) Second stage



(b) First stage

Figure 2.3: Decision tree of Theorem 2.9

To solve the optimization of the second stage:

$$\begin{aligned}\bar{l}_2(\hat{\boldsymbol{\mu}}_{\hat{\rho}}|\rho) &= \int l(\hat{\rho},\rho,\hat{\boldsymbol{\mu}}_{\hat{\rho}},\boldsymbol{\mu})p(\boldsymbol{\mu}|\rho,\boldsymbol{x})d\boldsymbol{\mu} \\ &= \gamma\int\|\hat{\boldsymbol{\mu}}_{\hat{\rho}}-\boldsymbol{\mu}\|^2 p(\boldsymbol{\mu}|\rho,\boldsymbol{x})d\boldsymbol{\mu}+(1-\gamma)\left(|\hat{\rho}|-|\rho|\right).\end{aligned}$$

Let $\hat{\boldsymbol{\mu}}^*_{\hat{\rho}} = \text{argmin}_{\hat{\boldsymbol{\mu}}_{\hat{\rho}}}\{\bar{l}_2(\hat{\boldsymbol{\mu}}_{\hat{\rho}}|\rho)\}$, then the loss of choosing $\hat{\rho}$ when ρ is the real state of the world would be $\bar{l}_2(\hat{\boldsymbol{\mu}}^*_{\hat{\rho}}|\rho)$. Now we will solve the first stage (see Figure 2.3b). The expected loss is given by

$$\begin{aligned}\bar{l}(\hat{\rho}) &= \int \bar{l}_2(\hat{\boldsymbol{\mu}}^*_{\hat{\rho}}|\rho)p(\rho|\boldsymbol{x})d\rho \\ &= \gamma\int\left\{\int\|\hat{\boldsymbol{\mu}}^*_{\hat{\rho}}-\boldsymbol{\mu}\|^2 p(\boldsymbol{\mu}|\rho,\boldsymbol{x})d\boldsymbol{\mu}\right\}p(\rho|\boldsymbol{x})d\rho+(1-\gamma)\left(|\hat{\rho}|-\int|\rho|p(\rho|\boldsymbol{x})d\rho\right) \\ &= \gamma\int\|\hat{\boldsymbol{\mu}}^*_{\hat{\rho}}-\boldsymbol{\mu}\|^2\int p(\boldsymbol{\mu}|\rho,\boldsymbol{x})p(\rho|\boldsymbol{x})d\rho d\boldsymbol{\mu}+(1-\gamma)\left(|\hat{\rho}|-\int|\rho|p(\rho|\boldsymbol{x})d\rho\right) \\ &= \gamma\int\|\hat{\boldsymbol{\mu}}^*_{\hat{\rho}}-\boldsymbol{\mu}\|^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu}+(1-\gamma)\left(|\hat{\rho}|-\int|\rho|p(\rho|\boldsymbol{x})d\rho\right)\end{aligned}$$

Since $\hat{\boldsymbol{\mu}}^*_{\hat{\rho}}$ is the Bayes estimate under a quadratic loss function,

$$\hat{\boldsymbol{\mu}}^*_{\hat{\rho}} = E\left(\boldsymbol{\mu}|\rho,\boldsymbol{x}\right) = \hat{\boldsymbol{\mu}}_{\rho}\left(\boldsymbol{x}\right)$$

Finally,

$$\begin{aligned}\int\|\hat{\boldsymbol{\mu}}_{\rho}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}\|^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu} &= \int\sum_{i=1}^{n}(\hat{\boldsymbol{\mu}}_{\rho,i}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}_i)^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu} \\ &= \int\sum_{i=1}^{n}(\hat{\boldsymbol{\mu}}_{\rho,i}\left(\boldsymbol{x}\right)-\hat{\boldsymbol{\mu}}_{B,i}\left(\boldsymbol{x}\right)+\hat{\boldsymbol{\mu}}_{B,i}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}_i)^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu} \\ &= \int\sum_{i=1}^{n}(\hat{\boldsymbol{\mu}}_{\rho,i}\left(\boldsymbol{x}\right)-\hat{\boldsymbol{\mu}}_{B,i}\left(\boldsymbol{x}\right))^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu} \\ &\quad +2\int\sum_{i=1}^{n}(\hat{\boldsymbol{\mu}}_{\rho,i}\left(\boldsymbol{x}\right)-\hat{\boldsymbol{\mu}}_{B,i}\left(\boldsymbol{x}\right))(\hat{\boldsymbol{\mu}}_{B,i}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}_i)p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu} \\ &\quad +\int\sum_{i=1}^{n}(\hat{\boldsymbol{\mu}}_{B,i}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}_i)^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu} \\ &= \|\hat{\boldsymbol{\mu}}_{\rho}\left(\boldsymbol{x}\right)-\hat{\boldsymbol{\mu}}_{B}\left(\boldsymbol{x}\right)\|^2+0+\text{tr}(V(\boldsymbol{\mu}|\boldsymbol{x}))\end{aligned}$$

where $tr(A)$ denotes the trace of a given matrix A. Then

$$\gamma\int\|\hat{\boldsymbol{\mu}}_{\rho}(\boldsymbol{x})-\boldsymbol{\mu}\|^2 p(\boldsymbol{\mu}|\boldsymbol{x})d\boldsymbol{\mu}+(1-\gamma)(|\hat{\rho}|-|\rho|) = \gamma\|\hat{\boldsymbol{\mu}}_{\rho}(\boldsymbol{x})-\hat{\boldsymbol{\mu}}_{B}(\boldsymbol{x})\|^2+\gamma\text{tr}(V(\boldsymbol{\mu}|\boldsymbol{x}))$$

$$+(1-\gamma)|\hat{\rho}| - (1-\gamma)\int |\rho| p(\rho|\boldsymbol{x}) d\rho$$

The proof concludes by noting that the second and fourth term in the last expression does not depend on $\hat{\rho}$. □

The above proof shows that the optimal choice $\hat{\rho}^*$ will be the partition for which the resulting estimate $\hat{\boldsymbol{\mu}}_\rho(\boldsymbol{x})$ is closest to $\hat{\boldsymbol{\mu}}_B(\boldsymbol{x})$ with the smallest number of elements. We now introduce a loss function that depends on the distribution of the data and not only on the mean value. Recall that F_i is the marginal distribution of $\boldsymbol{X}_i$ given $\rho = \{S_1, \ldots, S_k\}$; if $i \in S_j$ then $F_i = F_X^{S_j}$, and, $F_i \sim DP\left(\alpha^{S_j}, G^{S_j}\right)$. We define the loss function

$$l(\hat{\rho}, \rho, \hat{\boldsymbol{F}}_{\hat{\rho}}, \boldsymbol{F}) = \sum_{i=1}^{n} \gamma_i \|F_i - \hat{F}_{\hat{\rho},i}\|_i^2 + (1-\gamma)\left(|\hat{\rho}| - |\rho|\right) \tag{2.11}$$

where $\boldsymbol{F} = (F_1, \ldots, F_n)$, $\hat{\boldsymbol{F}}_{\hat{\rho}}$ is the estimate of $\boldsymbol{F}$ associated with the estimated partition $\hat{\rho}$, i.e. $\hat{\boldsymbol{F}}_{\hat{\rho}} = \left(\hat{F}_{\hat{\rho},1}, \ldots, \hat{F}_{\hat{\rho},n}\right)$. We define the following class of norms

$$\|F\|_i^2 = \int W_i(\boldsymbol{y})(F(\boldsymbol{y}))^2 d\nu_i(\boldsymbol{y})$$

with $W_i(\boldsymbol{y})$ a nonnegative function which represents a weighting function, $\nu_i(\boldsymbol{y})$ a measure on $\mathbb{R}^p$ and $\gamma = 1 - \sum_{i=1}^n \gamma_i$ with $0 \leq \gamma_i$ and $\gamma \leq 1$. The term $(|\hat{\rho}| - |\rho|)$ promotes a smaller number of clusters.

Theorem 2.10. *Let $\hat{F}_{B,i} = E\left(F_i|\boldsymbol{x}\right)$ and $\hat{F}_{\rho,i} = E\left(F_i|\rho, \boldsymbol{x}\right)$, then the expected loss minimization criterion using the loss function defined by equation (2.11) leads to the choice $\hat{\rho}^*$ that minimizes*

$$SC\left(\hat{\rho}\right) = \sum_{i=1}^{n} \gamma_i \|\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\|_i^2 + (1-\gamma)|\hat{\rho}|$$

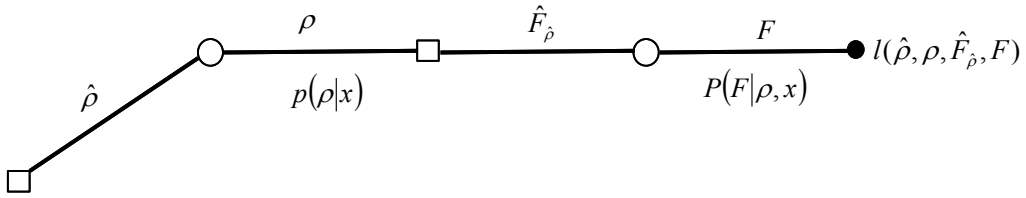
Proof. Let

$$\hat{\boldsymbol{F}}_\rho = \left(\hat{F}_{\rho,1}, \ldots, \hat{F}_{\rho,n}\right), \ \hat{\boldsymbol{F}}_B = \left(\hat{F}_{B,1}, \ldots, \hat{F}_{B,n}\right)$$

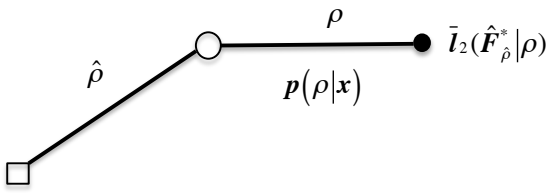
and

$$\|\hat{\boldsymbol{F}}_{\hat{\rho}} - \boldsymbol{F}\|^2 = \sum_{i=1}^{n} \gamma_i \|F_i - \hat{F}_{\rho,i}\|_i^2.$$

The corresponding decision tree is depicted in Figure 2.4.



(a) Second stage



(b) First stage

Figure 2.4: Decision tree for the decision problem of Theorem 2.10.

We begin by solving the optimization problem of the second stage:

$$\begin{aligned}
\bar{l}_2(\hat{\boldsymbol{F}}_{\hat{\rho}}|\rho) &= \boldsymbol{E}_{\boldsymbol{F}|\rho,\boldsymbol{x}}\left(l(\hat{\rho}, \rho, \hat{\boldsymbol{F}}_{\hat{\rho}}, \boldsymbol{F})\right) \\
&= \boldsymbol{E}_{\boldsymbol{F}|\rho,\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}_{\hat{\rho}} - \boldsymbol{F}\|^2 + (1-\gamma)\left(|\hat{\rho}| - |\rho|\right)\right) \\
&= \boldsymbol{E}_{\boldsymbol{F}|\rho,\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}_{\hat{\rho}} - \boldsymbol{F}\|^2\right) + (1-\gamma)\left(|\hat{\rho}| - |\rho|\right).
\end{aligned}$$

Let $\hat{\boldsymbol{F}}^*_{\hat{\rho}} = \operatorname{argmin}_{\hat{\boldsymbol{F}}_{\hat{\rho}}}\{\bar{l}_2(\hat{\boldsymbol{F}}_{\hat{\rho}}|\rho)\}$; then the loss of choosing $\hat{\rho}$ when ρ is the real state of the world would be $\bar{l}_2(\hat{\boldsymbol{F}}^*_{\hat{\rho}}|\rho)$. Since $\hat{\boldsymbol{F}}^*_{\hat{\rho}}$ is the Bayes estimate under a quadratic loss function,

$$\hat{\boldsymbol{F}}^*_{\hat{\rho}} = \boldsymbol{E}_{\boldsymbol{F}|\rho,\boldsymbol{x}}\left(\boldsymbol{F}\right) = \hat{\boldsymbol{F}}_{\rho}.$$

Now we will solve the second stage (see Figure 2.4b). The expected loss is given by

$$\begin{aligned}
\bar{l}(\hat{\rho}) &= \boldsymbol{E}_{\rho|\boldsymbol{x}}\left(\bar{l}_2(\hat{\boldsymbol{F}}^*_{\hat{\rho}}|\rho)\right) \\
&= \boldsymbol{E}_{\rho|\boldsymbol{x}}\left(\boldsymbol{E}_{\boldsymbol{F}|\rho,\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}^*_{\hat{\rho}} - \boldsymbol{F}\|^2\right) + (1-\gamma)\left(|\hat{\rho}| - |\rho|\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{E}_{\rho|\boldsymbol{x}}\left(\boldsymbol{E}_{\boldsymbol{F}|\rho,\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}^*_{\hat{\rho}} - \boldsymbol{F}\|^2\right)\right) + (1-\gamma)\left(|\hat{\rho}| - \boldsymbol{E}_{\rho|\boldsymbol{x}}(|\rho|)\right) \\
&= \boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}^*_{\hat{\rho}} - \boldsymbol{F}\|^2\right) + (1-\gamma)\left(|\hat{\rho}| - \boldsymbol{E}_{\rho|\boldsymbol{x}}(|\rho|)\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}^*_{\hat{\rho}} - \boldsymbol{F}\|^2\right) &= \boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\sum_{i=1}^{n}\gamma_i\|F_i - \hat{F}_{\hat{\rho},i}\|_i^2\right) \\
&= \boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\sum_{i=1}^{n}\gamma_i\int W_i(\boldsymbol{y})\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2 d\nu_i(\boldsymbol{y})\right) \\
&= \sum_{i=1}^{n}\gamma_i\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\int W_i(\boldsymbol{y})\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2 d\nu_i(\boldsymbol{y})\right).
\end{aligned}
$$

Applying Fubini's theorem,

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\int W_i(\boldsymbol{y})\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2 d\nu_i(\boldsymbol{y})\right) &= \int \boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(W_i(\boldsymbol{y})\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2\right) d\nu_i(\boldsymbol{y}) \\
&= \int W_i(\boldsymbol{y})\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2\right) d\nu_i(\boldsymbol{y}) \qquad (2.12)
\end{aligned}
$$

with

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2\right) &= \boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{B,i} + \hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)^2\right) \\
&= \boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{B,i}\big)^2\right) \\
&\quad +2\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{B,i}\big)\big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)\right) \\
&\quad +\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)^2\right). \qquad (2.13)
\end{aligned}
$$

Note that

$$
\begin{aligned}
\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{B,i}\big)\big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)\right) &= \big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)\,\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{B,i}\big)\right) \\
&= 0
\end{aligned}
$$

The first term in equation (2.13) does not depend on $\hat{\rho}$ and can be thought of as a constant K; hence

$$
\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2\right) = K + 0 + \big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)^2.
$$

Substituting into equation (2.12), we obtain

$$\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\int W_i(\boldsymbol{y})\big(F_i - \hat{F}_{\hat{\rho},i}\big)^2 d\nu_i(\boldsymbol{y})\right) = \int W_i(\boldsymbol{y})\big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)^2 d\nu_i(\boldsymbol{y}).$$

Therefore,

$$\begin{aligned}\boldsymbol{E}_{\boldsymbol{F}|\boldsymbol{x}}\left(\|\hat{\boldsymbol{F}}^*_{\hat{\rho}} - \boldsymbol{F}\|^2\right) &= \sum_{i=1}^{n}\gamma_i \int W_i(\boldsymbol{y})\big(\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\big)^2 d\nu_i(\boldsymbol{y}) + K_2 \\ &= \sum_{i=1}^{n}\gamma_i \|\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\|_i^2 + K_2\end{aligned}$$

where K_2 is another constant. Hence,

$$\bar{l}(\hat{\rho}) = \sum_{i=1}^{n}\gamma_i \|\hat{F}_{B,i} - \hat{F}_{\hat{\rho},i}\|_i^2 + (1-\gamma)|\hat{\rho}| - (1-\gamma)\boldsymbol{E}_{\rho|\boldsymbol{x}}(|\rho|) + K_2$$

The proof concludes by noting that the third and fourth terms in the last expression do not depend on $\hat{\rho}$. □

The above proof shows that the optimal choice $\hat{\rho}^*$ will be the partition for which the resulting estimate $\hat{\boldsymbol{F}}_\rho$ is closest to $\hat{\boldsymbol{F}}_B$.

Theorems 2.9 and 2.10 suggest a procedure based on distances to find the optimal $\hat{\rho}^*$. However, an exhaustive search on the space of all possible partitions is infeasible. Therefore we will adopt different heuristic algorithms depending on each application. Unfortunately, these strategies can not give us the optimal solution but they can lead to a reasonable one in a realistic amount of time.

2.3 Nonparametric product partition models for change-point analysis

2.3.1 Introduction

From the statistical point of view, a change point is a place or time point such that the observations follow one distribution up to that point and follow another distribution after that point. Multiple change-point problems can be defined similarly (Chen and Gupta, 2011). Change-point problems arise naturally in many disciplines such as economics, finance, medicine, psychology and geology, and statisticians have developed a number of methodologies to deal with this topic. The reader is referred, for example,

to Chen and Gupta (2011) for parametric classic statistical models, and to Csörgö and Horváth (1997) and Brodsky and Darkhovsky (2010) for nonparametric classical approaches and related results. For Bayesian parametric models see Ecley et al. (2011). Another recent Bayesian parametric approach is discussed by Killick et al. (2012) using a cost function to detect change points. Since our approach is based on product partition models (PPMs) for the underlying structure of the data, we start by briefly reviewing relevant related work.

We can see change-point analysis as a particular case of clustering where the observations are ordered, usually by time. Because of its inherent cluster properties, PPMs offer a convenient approach that can lead to good estimates in this setting. In fact, change-point analysis for predictions purposes is an early application of PPMs (Barry and Hartigan 1992, 1993). Later Loschi and Cruz (2005) and Quintana et al. (2005a) proposed criteria to identify change points in these models. The former authors suggest to use the marginal probability of being a change point at each location of the time series, while the latter authors use a weighted loss function to identify the shifts of behaviour. Many algorithms assume that each cluster has the same density (Chen and Gupta, 2011). This approach may be very restrictive, which is why many authors model the densities nonparametrically.

An early example of this approach from a Bayesian perspective can be found in Mira and Petrone (1996), who used mixture of Dirichlet processes to model the densities; this ensures flexibility of the distribution for each cluster although their approach is computationally very demanding when dealing with several change points. More recently, Yau et al. (2011) proposed another approach which uses hidden Markov models for the cluster structure and mixture of Dirichlet processes to model the density at each state. In this model, we need to establish the number of states in advance, which can be very limiting for some applications. Furthermore, Yau et al. (2011) use the marginal probability to detect change points. This criterion has several limitations which are discussed in Yau and Holmes (2013). A recent Bayesian parametric approach was introduced by Killick et al. (2012) using a cost function to detect change points.

The main goal of this section is to introduce a flexible model for change-point detection that can discover fluctuations in the distribution in sequentially observed data. Let $x_1, ..., x_n$ be a data sequence and consider the index set $I = \{1, ..., n\}$. Consider a random partition $\rho = \{i_0, i_1, ..., i_b\}$ of the set $I \cup \{0\}$, with ordered points

$0 = i_0 < i_1 < i_b = n$, and a random variable B which denotes the number of blocks in ρ. Consider that each partition divides the data sequence into b contiguous subsequences, which will be denoted here by $\boldsymbol{x}_{i_{(r-1)}i_r} = \left(x_{i_{(r-1)}+1}, ..., x_r\right)^T$, for $r = 1, ..., b$.
Let c_{ij} be the prior cohesion associated with block $ij = \{i+1, ..., j\}$, for $i, j \in I \cup \{0\}$, and $j > 1$, that represents the degree of similarity among the observations in $\boldsymbol{x}_{i,j}$.
Let $F_1, ..., F_n$ be a sequence of marginal distributions of $x_1, ..., x_n$ respectively. Given a partition ρ, we have that $F_i = F_{i_{(r-1)}i_r}$, for every $i_{(r-1)} < i \leq i_r$, and $F_{i_0 i_1}, ..., F_{i_{(b-1)}i_b}$ are independent, with $F_{ij} \sim DP(\alpha_{ij}, G_{ij})$.

We say that the random quantity $(x_1, ..., x_n; \rho)$ follows a nonparametric product partition model (NPPM), denoted by $(x_1, ..., x_n; \rho) \sim NPPM$, if

1. the prior distribution of ρ is

$$p(\rho = \{i_0, i_1, ..., i_b\}) \propto \prod_{j=1}^{b} c_{i_{(j-1)}i_j} \tag{2.14}$$

2. conditional on $\rho = \{i_0, i_1, ..., i_b\}$, we have the following hierarchical model

$$\begin{aligned} F_i|\alpha_i, G_i \quad &| \quad \overset{ind}{\sim} DP(\alpha_i, G_i) \\ x_1, ..., x_n \quad &| \quad F_1, ..., F_n, \rho \sim \prod_{i=1}^{n} F_i(x_i) \end{aligned} \tag{2.15}$$

For a graphical representation for the NPPMs for change-point analysis, see Figure 2.5

Remark 2.11. *The hyperparameters α_i and G_i are fixed.*

We now calculate the posterior distribution of the partition for the NPPMs applied to change point analysis.

Proposition 2.12. *The corresponding posterior distribution of ρ is again of the form of equation (2.14), with cohesions given by $c(S)p_S(\boldsymbol{x}_S) = c_{ij}p_{ij}(\boldsymbol{x}_{ij})$ where $p_{ij}(\boldsymbol{x}_{ij})$ is the predictive distribution defined in proposition 2.3 with $S = \{i+1, ..., j\}$*

Proof. Straightforward using Proposition 2.3. □

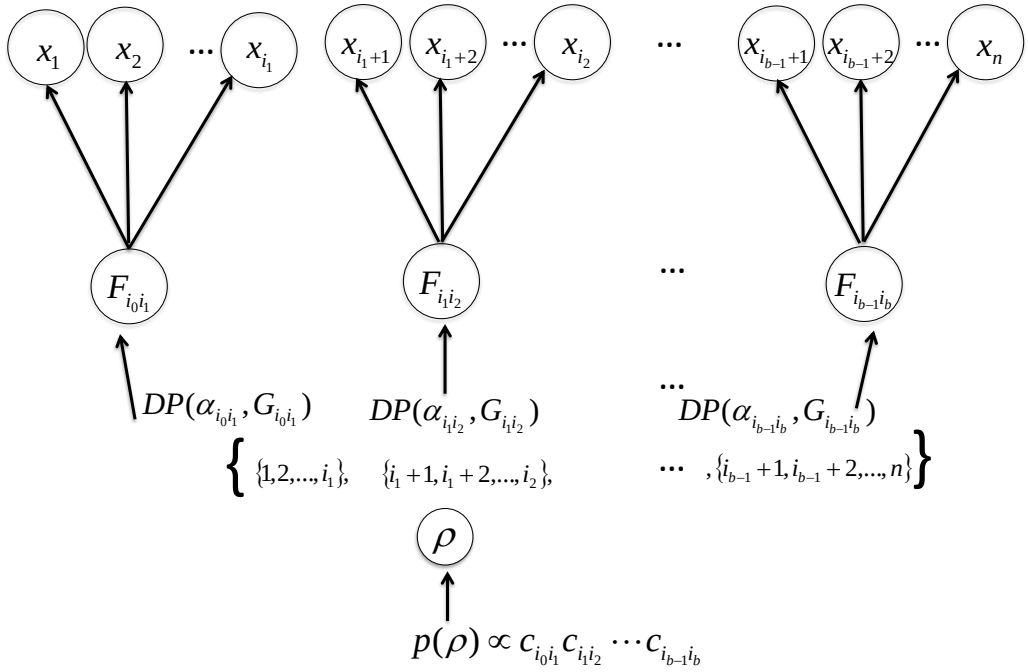


Figure 2.5: Graphical representation of NPPMs for change-point analysis.

2.3.2 Estimation of the weights in the loss function

The choice of loss function implies a trade-off (controlled by the user-defined quantity γ) between the optimal estimator of $\boldsymbol{F}$ and model simplicity. Let,

$$l(\hat{\rho}, \hat{\boldsymbol{F}}_{\hat{\rho}}, \boldsymbol{F}) = \sum_{i=1}^{n} \gamma_i \|F_i - \hat{F}_{\hat{\rho},i}\|_i^2 + (1-\gamma)|\hat{\rho}| \tag{2.16}$$

where $\boldsymbol{F} = (F_1, ..., F_n)$, $\hat{\boldsymbol{F}}_{\hat{\rho}}$ is the estimate of $\boldsymbol{F}$ associated to the estimated partition $\hat{\rho}$, $\hat{\boldsymbol{F}}_{\hat{\rho}} = \left(\hat{F}_{\hat{\rho},1}, ..., \hat{F}_{\hat{\rho},n}\right)$, $\|G\|_i^2 = \int |G(\boldsymbol{x})|^2 dW_i(\boldsymbol{x})$, with $W_i(\boldsymbol{x})$ a probability distribution function on $\mathbb{R}^p$ and $\gamma = \sum_{i=1}^{n} \gamma_i$ with $0 \leq \gamma_i$ and $\gamma \leq 1$.

We use Theorem 2.10 and Algorithm 1.1 to estimate the change points.

Notice that, for a fixed value of γ, we obtain the change points by minimizing equation (2.11) which we will denote by

$$\rho_\gamma = \left\{i_0^\gamma, i_1^\gamma, ..., i_b^\gamma\right\}.$$

Denote the sum of squared errors associated to γ as

$$SSE_\gamma = \sum_{i=1}^{n} (x_i - C_i^\gamma)^2,$$

where C_i^γ is the mean of observations $\{x_{i_{r-1}^\gamma+1}, ..., x_{i_r^\gamma}\}$ for $i_{r-1}^\gamma < i \leq i_r^\gamma$. Clearly, for a fixed value γ_0 of the parameter γ, there exists an interval $I = (\gamma_a, \gamma_b)$ such that for all $\gamma \in I$ we obtain $\rho_\gamma = \rho_{\gamma_0}$; therefore, it is enough to consider the number of change points versus the SSE_γ for the analysis.

2.3.3 Exact computational procedures

Barry and Hartigan (1993) provide an exact computational procedure to estimate the gold standard $\hat{\boldsymbol{F}}_B$ used in Theorem 2.10. In this section, we will describe it briefly for the nonparametric case.

Define

$$\lambda_{ij} = \sum \prod_{k=1}^{b} c_{i_{k-1} i_k},$$

where the summation is over all sets of integers $i = i_0 < i_1 < ... < i_b = j$. The quantity λ_{ij} is the sum of products of cohesions over all possible partitions of the set $\{i+1, i+2, ..., j\}$. Let the relevance r_{ij} be the probability that the block ij is included in the partition ρ. Then

$$r_{ij} = \frac{\lambda_{0i} c_{ij} \lambda_{jn}}{\lambda_{0n}},$$

The quantities λ_{0i} and λ_{jn} may be calculated in $O(n^2)$ steps using the recursive formulas

$$\begin{aligned}
\lambda_{01} &= c_{01}, \\
\lambda_{0i+1} &= c_{0i+1} + \sum_{k=1}^{i} \lambda_{0k} c_{ki+1}, \\
\lambda_{n-1n} &= c_{n-1n}, \\
\text{and} & \\
\lambda_{jn} &= c_{jn} + \sum_{k=j+1}^{n-1} c_{jk} \lambda_{kn}
\end{aligned}$$

In our case, we want to calculate $\hat{\boldsymbol{\mu}}_B$ and $\hat{\boldsymbol{F}}_B$ given the observations. The posterior relevances $r_{ij}(\boldsymbol{x})$ are computed from the posterior cohesions by recursive formulas like

those just listed. Now

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{B,k} &= E(\boldsymbol{\mu}_k|\boldsymbol{x}) \\ &= \sum_{i<k\leq j} E_{ij}(\boldsymbol{\mu}_k|x_{ij})r_{ij}(\boldsymbol{x}),\end{aligned}$$

where $E_{ij}(\boldsymbol{\mu}_k|x_{ij})$ denotes the posterior expectation of $\boldsymbol{\mu}_k$ when the block ij lies in the partition. Similarly for $\hat{\boldsymbol{F}}_B$, i.e.

$$\begin{aligned}\hat{\boldsymbol{F}}_{B,k} &= E(F_k|\boldsymbol{x}) \\ &= \sum_{i<k\leq j} E_{ij}(F_k|\boldsymbol{x_{ij}})r_{ij}(\boldsymbol{x}),\end{aligned}$$

where $E_{ij}(F_k|x_{ij})$ denotes the posterior expectation of F_k when the block ij lies in the partition. In the model defined by equations (2.14) and (2.15), the posterior cohesion functions are given by

$$p\left(\rho = \{i_0, i_1, ..., i_b\}\,|\boldsymbol{x}\right) \propto \prod_{j=1}^{b} c_{i_{(j-1)}i_j} p(\boldsymbol{x}_{i_{(j-1)}i_j}) \tag{2.17}$$

with

$$p(\boldsymbol{x}_{ij}) = \frac{\alpha_{ij}^{K_{ij}} \prod_{k=1}^{K_{ij}} (|\boldsymbol{x}_k^*| - 1)!}{\alpha(1+\alpha)...(N-1+\alpha)} \prod_{k=1}^{K_{ij}} G_{ij}(\boldsymbol{x}_k^*),$$

where $\boldsymbol{x}_k^*$ for $k = 1, ..., K_{ij}$ are the distinct values in block ij and $|\boldsymbol{x}_k^*|$ is the number of times that the value $\boldsymbol{x}_k^*$ is repeated in block ij. The number of possible blocks in n data points is $\binom{n+1}{2}$. For small data sets, this procedure is feasible ($n < 500$) but for greater values of n, it is computationally expensive. We tackle this issue using a Gibbs sampling scheme described in the following section.

2.3.4 Gibbs sampling for change-point analysis

Let U_i be an auxiliary random quantity that reflects whether or not a change point occurs at time i (Barry and Hartigan, 1993); i.e.

$$U_i = \begin{cases} 1 & \text{if} \quad F_i = F_{i+1}, \\ 0 & \text{if} \quad F_i \neq F_{i+1}, \end{cases}$$

for $i = 1, ..., n-1$. Each partition $\left(U_1^s, ..., U_{n-1}^s\right)$, $s \geq 1$, is generated by using Gibbs sampling. Starting from an initial value $\left(U_1^0, ..., U_{n-1}^0\right)$, the r-th element at step s, U_r^s, is generated from the conditional distribution:

$$U_r \mid U_1^s, ..., U_{r-1}^s, U_{r+1}^{s-1}, ..., U_{n-1}^{s-1}; \boldsymbol{x}_{0n}$$

for $r = 1, ..., n-1$. To avoid unnecessary calculations, it is enough to consider the following ratio:

$$R_r = \frac{P\left(U_r = 1 \mid V_r^s; \boldsymbol{x}_{0n}\right)}{P\left(U_r = 0 \mid V_r^s; \boldsymbol{x}_{0n}\right)}$$

for $r = 1, ..., n-1$, in which $V_r^s = \left\{U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1}\right\}$

$$P\left(U_r = 1 \mid V_r^s; \boldsymbol{x}_{0n}\right) = \frac{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 1, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid \boldsymbol{x}_{0n}\right)}{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid \boldsymbol{x}_{0n}\right)}$$

and

$$P\left(U_r = 0 \mid V_r^s; \boldsymbol{x}_{0n}\right) = \frac{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 0, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid \boldsymbol{x}_{0n}\right)}{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid \boldsymbol{x}_{0n}\right)}.$$

Then

$$R_r = \frac{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 0, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid \boldsymbol{x}_{0n}\right)}{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid \boldsymbol{x}_{0n}\right)}$$

.

Let

$$x = \begin{cases} \max\{i, s.t : 0 < i < r, U_i^s = 0\} & \text{if} \quad U_i^s = 0, \text{ for some } i \in \{1, ..., r-1\} \\ 0 & \text{otherwise} \end{cases}$$

and

$$y = \begin{cases} \min\{i, s.t : r < i < n, U_i^s = 0\} & \text{if} \quad U_i^s = 0, \text{ for some } i \in \{r+1, ..., n-1\} \\ n & \text{otherwise.} \end{cases}$$

Hence

$$R_r = \frac{c_{xy} \int \prod_{i=x+1}^{y} F\left(x_i\right) p\left(F\right) dF}{c_{xr} c_{ry} \int \prod_{i=x+1}^{r} F\left(x_i\right) p\left(F\right) dF \cdot \int \prod_{i=r+1}^{y} F\left(x_i\right) p\left(F\right) dF}.$$

Consequently, the criterion for choosing the values U_i^s, $i = 1, ..., n-1$, becomes:

$$U_r^s = \begin{cases} 1, & \text{if} \quad R_r \geq \frac{1-u}{u} \\ 0, & \text{otherwise} \end{cases}$$

for $r = 1, ..., n-1$, in which $u \sim Unif\left(0, 1\right)$.

Fuentes-García et al. (2010) and Martínez and Mena (2014) provide a clever Gibbs sampler using $p(k, n_1, ..., n_k)$ where k is the number of blocks and n_i is the number of observations within block i.

2.3.5 Multiple change-point analysis with missing values

Missing data arise in many applications such as aCGH analysis that we will describe in Section 2.4.3.2. To overcome this difficulty, we will start by assuming that the data are missing at random, which is a reasonable assumption in many situations. Our model allows us to handle this problem naturally since a common practice to estimate a missing value in change-point analysis is to average values near the point where a missing value appears. In our setting, we have random partitions which allow us to average the distributions functions of all possible partitions of the data. It can be easily shown that

$$p(\rho|X_1,...,X_n) \propto p(\rho_c|\boldsymbol{X}_c),$$

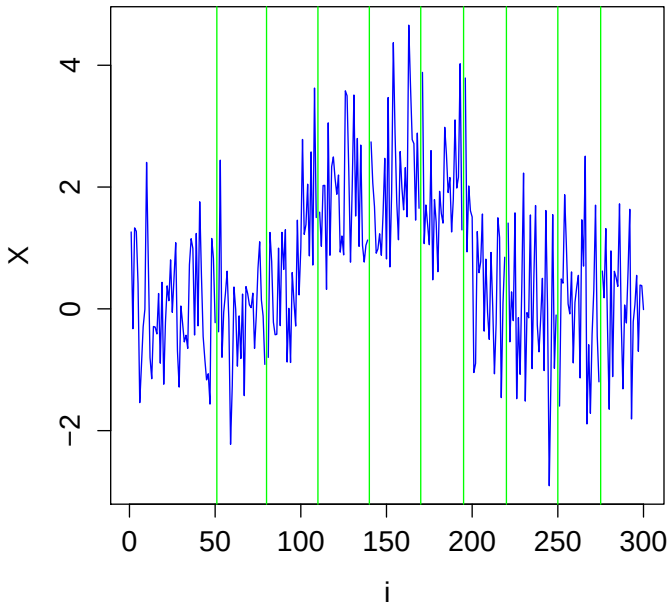
where $\boldsymbol{X}_c$ denotes the data and ρ_c the corresponding partition without missing values. Therefore, the quantities of interest of our analysis, F_i, $F_{B,i}$ can be calculated without extra computations. For instance, if we have 101 data points with a missing value at $i = 80$, consider the partition $\rho = [1,50][51,80][81,101]$, then $X_c = \{X_1,...,X_{79},X_{81},...,X_{101}\}$ and $\rho_c = [1,50][51,79][81,101]$. Notice that we will be able to estimate $E[F_k|X_c,\rho_c] = E[F_{80}|X_1,...,X_{100},\rho]$ with $51 \leq k \leq 79$.

We will provide two examples where we apply this idea.

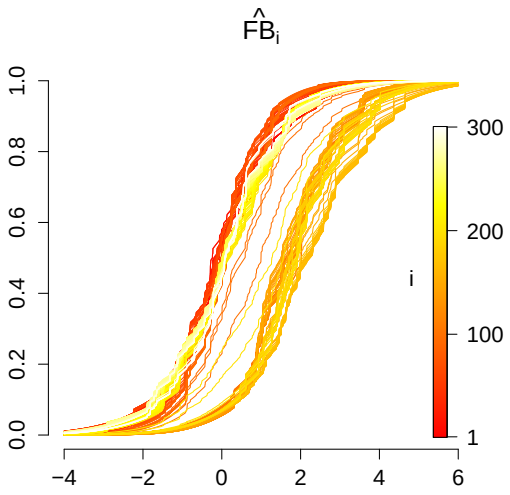
Example 2.13. *We simulated a data set as follows,*

$$X_i \overset{ind}{\sim} N(\mu_i,1) \begin{cases} \mu_i = 0, & \text{if } 1 \leq i \leq 100 \\ \mu_i = 2, & \text{if } 101 \leq i \leq 200 \\ \mu_i = 0, & \text{if } 201 \leq i \leq 300, \end{cases}$$

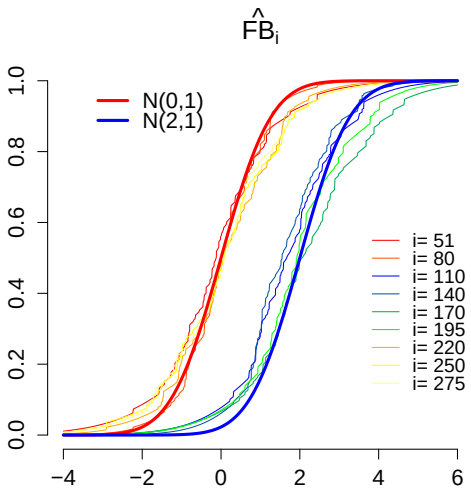
with missing values at $i \in \{51,80,110,140,170,195,220,250,275\}$. *Figure 2.6a shows the simulated data with missing values. The estimated distributions at each point are displayed in Figure 2.6b, where we can distinguish two clusters which correspond to the two change points present in the data. The estimated distributions for missing values together with the true distributions are presented in Figure 2.6c, where we can see that the estimated distributions are close to the true ones.*



(a) X_i. Green lines indicates missing values.



(b) $\hat{F}_{B,i}$ for $1 \leq i \leq 300$



(c) $\hat{F}_{B,i}$ for missing values.

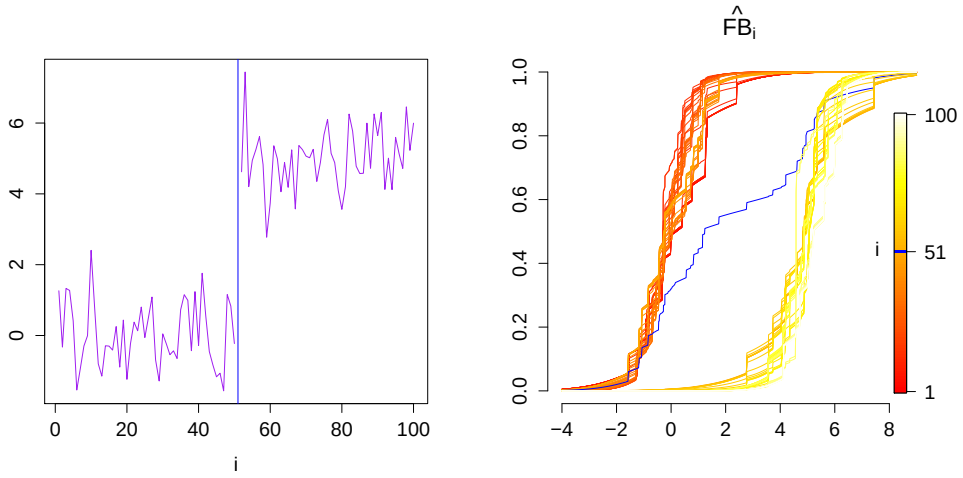
Figure 2.6: Graphics for Example 2.13.

What happens when the missing value appears at a change point? The following example will clarify this question.

Example 2.14. *We simulated a data set as follows,*

$$X_i \overset{ind}{\sim} N(\mu_i, 1) \begin{cases} \mu_i = 0, & \text{if } 1 \leq i \leq 50 \\ \mu_i = 2, & \text{if } 51 \leq i \leq 100 \end{cases}$$

with missing values at $i = 51$*. Figure 2.7a shows simulated data with a missing value at* $i = 51$*. The estimated distributions at each point are displayed in Figure 2.7b, where we can also distinguish two clusters which corresponds to the change point presented in the data. Nevertheless, the estimated distribution (in blue) for* $i = 51$ *is a mixture of two distributions: the first one of block* $\{1, ..., 50\}$ *and the second one of block* $\{52, ..., 100\}$*.*



(a) X_i (b) $\hat{F}_{B,i}$ for $1 \leq i \leq 100$.

Figure 2.7: Graphics for Example 2.14.

2.4 Simulation experiments and applications

In the examples of this section we use the following cohesion function

$$c_{ij} = c_{NP} \times p \times (1-p)^{j-i-1},$$

where $c_{NP} > 0$ and $p \in (0, 1)$ are constants. We note that p can be interpreted as the probability of change at an arbitrary point, so that $p \times (1-p)^{j-i-1}$ is the probability

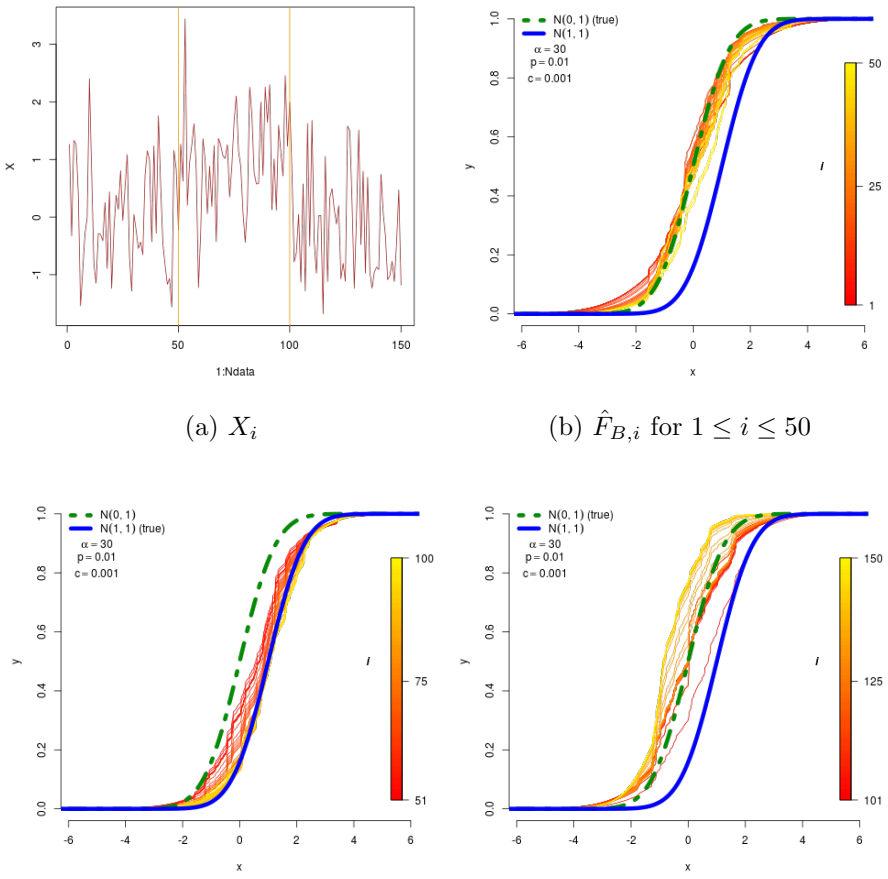
of the event that a change point occurs at j. The parameter c_{NP} allows us to promote fewer change-points when $c_{NP} < 1$. In what follows, we set $c_{NP} = 0.001$ and $p = 0.01$. For the parameters of the Dirichlet process $DP(\alpha_{ij}, G_{ij})$ we use a Normal distribution, $G_{ij} = N\left(\mu_{ij}, \sigma_{ij}^2\right)$. We employ an empirical Bayes approach to estimate μ_{ij} and σ_{ij} as follows: $\hat{\mu}_{ij} = \text{median}(\boldsymbol{x}_{ij})$ and $\hat{\sigma}_{ij} = \text{IQR}(\boldsymbol{x}_{ij})/1.349$ where IQR stands for interquartile range. We use these robust estimators so as to allow for observations from heavy-tailed distributions. For the disperpersion parameters α_{ij}, we use the following rule of thumb: if the number of observations is less than 50, we set $\alpha_{ij} = 1$; for larger samples, we use $\alpha_{ij} = 30$.

2.4.1 Simulation study

Example 2.15. *Here we give an example of the distribution functions estimated $\hat{F}_{B,i}$. We simulated a data set as follows,*

$$X_i \overset{ind}{\sim} N(\mu_i, 1) \begin{cases} \mu_i = 0, & \text{if } 1 \leq i \leq 50 \\ \mu_i = 1, & \text{if } 51 \leq i \leq 100 \\ \mu_i = 0, & \text{if } 101 \leq i \leq 150. \end{cases}$$

Figure 2.8 shows the simulated data together with the estimated distributions at each point.



(a) X_i

(b) $\hat{F}_{B,i}$ for $1 \leq i \leq 50$

(c) $\hat{F}_{B,i}$ for $51 \leq i \leq 100$

(d) $\hat{F}_{B,i}$ for $101 \leq i \leq 150$

Figure 2.8: Graphics for Example 2.15.

In the next examples, we will also present simulation results from the nonparametric product partition models using different function weights W_i and measure ν_i in the loss function of Theorem 2.10 as follows. The first one is the NPPMs with $W_i(x) = 1$ and $\nu_i(x)$ corresponding to the Lebesgue measure; for the second one, we set $W_i(x) = 1$ and $\nu_i(x) = \hat{F}_{B,i}(x)$, which we name NPPMsB; finally, inspired by the Anderson-Darling test, we use $W_i(x) = \frac{1}{\hat{F}_{B,i}(x)\left(1-\hat{F}_{B,i}(x)\right)}$ with $\nu_i(x) = \hat{F}_{B,i}(x)$, and refer to it as NPPMsBB. We study the behavior of the probability of change point of the NPPMs as well and this will be identified by NPPMsP.

We compare the performance with that of parametric and nonparametric procedures that have recently appeared in the literature, such as the Pruned Exact Linear Time

(PELT) procedure (Killick et al., 2012), which is based on a penalization function, and ECP (E-divisive change-point analysis) introduced by Matteson and James (2014), which is based on a divergence measure that can determine whether two independent random vectors are identically distributed. We also compare our proposal with the parametric product partition model introduced by Barry and Hartigan (1992, 1993) (bcp) which detects changes in mean with known variance, and the normal-gamma product partition model studied by Loschi et al. (2003) which detects changes in mean and variance in normally distributed data. For this model we analyze the marginal probability of change point (NG P) and the loss function (NG) criterion. For each method, we evaluate their effectiveness in detecting change in mean, variance, tail and skewness for 20, 150 and 300 observations using the Rand index introduced by Rand (1971) which we will describe briefly. Suppose that we have n data points with two ways of clustering the data: U and V. Let A be the set of pairs of data points that are together in U and V. Let B the set of pairs of data points that are not clustered together under U and V. Then the Rand index is defined as follows

$$RandI = \frac{|A| + |B|}{\binom{n}{2}}.$$

In words, the Rand index is the percentage of pairs of data in which the partitions A and B cluster in the same way; that is, if both observations are clustered together or not.

Example 2.16. ***Small data sets.*** *Each simulation applies the methods described earlier to a set of* 1000 *independent sequences of* 20 *observations with one change point at* $i = 11$ *with distributions* G_1 *and* G_2 *respectively for each block. To assess the performance for changes in mean, we set* $G_1 = N(\mu_1, 1)$ *and* $G_2 = N(\mu_2, 1)$*; for changes in variance we define* $G_1 = N(0, \sigma_1^2)$ *and* $G_2 = N(0, \sigma_2^2)$*; in the case of changes in tail we specify* $G_1 = N(0, \sigma^2)$ *and* $G_2 = t_\nu$*. Finally, to study changes in skewness, we define* $G_1 = N(\mu, 1)$ *and* $G_2 = sN(0, 1, \alpha)$ *where* sN *denotes the skew normal distribution and* $\mu = Mode(G_2)$*. Table 2.1 shows the mean and standard deviation of the Rand index for each method and each case considered. For the purpose of facilitating the interpretation of results, the table is colored with darker tones denoting better performance. We can see that the methods based on loss functions (NPPMs, NPPMsB, NPPMsBB, NG, ECP and PELT) perform better than methods that rely on the marginal probability of change point (NPPMsP, bcp and NG P). The NPPMs with loss function have similar efficiency and perform better than the other parametric and nonparametric approaches.*

Change in mean ($N = 20$)									
(μ_1, μ_2)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG P**	**ECP**	**PELT**
$(0, 0)$	$0.82_{0.25}$	$0.97_{0.10}$	$0.98_{0.09}$	$0.97_{0.10}$	$0.97_{0.10}$	$0.99_{0.05}$	$0.93_{0.16}$	$0.95_{0.13}$	$0.99_{0.03}$
$(0, 1)$	$0.63_{0.20}$	$0.82_{0.16}$	$0.82_{0.16}$	$0.82_{0.16}$	$0.62_{0.17}$	$0.76_{0.18}$	$0.65_{0.19}$	$0.80_{0.17}$	$0.74_{0.18}$
$(0, 2)$	$0.80_{0.21}$	$0.92_{0.11}$	$0.90_{0.14}$	$0.93_{0.09}$	$0.80_{0.19}$	$0.91_{0.13}$	$0.80_{0.20}$	$0.95_{0.10}$	$0.90_{0.16}$
$(0, 4)$	$0.96_{0.07}$	$0.98_{0.03}$	$0.98_{0.03}$	$0.99_{0.03}$	$0.98_{0.06}$	$0.99_{0.03}$	$0.96_{0.06}$	$0.999_{0.004}$	$0.996_{0.02}$
Change in variance ($N = 20$)									
(σ_1^2, σ_2^2)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG P**	**ECP**	**PELT**
$(1, 2.25)$	$0.58_{0.17}$	$0.81_{0.14}$	$0.82_{0.13}$	$0.82_{0.13}$	$0.62_{0.14}$	$0.70_{0.16}$	$0.66_{0.13}$	$0.71_{0.16}$	$0.69_{0.13}$
$(1, 5)$	$0.65_{0.20}$	$0.82_{0.17}$	$0.86_{0.12}$	$0.86_{0.12}$	$0.650_{0.12}$	$0.77_{0.15}$	$0.70_{0.09}$	$0.73_{0.16}$	$0.75_{0.17}$
$(1, 10)$	$0.73_{0.22}$	$0.88_{0.12}$	$0.88_{0.10}$	$88_{0.12}$	$0.67_{0.13}$	$0.82_{0.16}$	$0.72_{0.12}$	$0.79_{0.17}$	$0.84_{0.15}$
$(1, 64)$	$0.90_{0.15}$	$0.93_{0.08}$	$0.91_{0.13}$	0.95_{08}	$0.74_{0.07}$	$0.90_{0.14}$	$0.83_{0.20}$	$0.92_{0.13}$	$0.96_{0.08}$
Change in tail ($N = 20$)									
(σ, ν)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG p**	**ECP**	**PELT**
$(1, 2.1)$	$0.6_{0.17}$	$0.75_{0.16}$	$0.8_{0.13}$	$0.74_{0.17}$	$0.59_{0.14}$	$0.67_{0.14}$	$0.65_{0.09}$	$0.69_{0.16}$	$0.67_{0.13}$
$(1, 3)$	$0.58_{0.16}$	$0.78_{0.15}$	$0.8_{0.13}$	$0.8_{0.14}$	$0.6_{0.13}$	$0.67_{0.14}$	$0.65_{0.09}$	$0.7_{0.15}$	$0.67_{0.13}$
$(1, 4)$	$0.58_{0.16}$	$0.81_{0.12}$	$0.8_{0.13}$	$0.8_{0.13}$	$0.59_{0.12}$	$0.67_{0.15}$	$0.65_{0.09}$	$0.7_{0.15}$	$0.66_{0.13}$
$(1, 8)$	$0.58_{0.15}$	$0.81_{0.12}$	$0.81_{0.13}$	$0.8_{0.14}$	$0.58_{0.12}$	$0.64_{0.15}$	$0.65_{0.1}$	$0.68_{0.15}$	$0.67_{0.1}$
Change in skewness ($N = 20$)									
(μ, α)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG p**	**ECP**	**PELT**
$(0.5, 1)$	$0.57_{0.16}$	$0.81_{0.13}$	$0.81_{0.13}$	$0.8_{0.14}$	$0.57_{0.12}$	$0.62_{0.15}$	$0.65_{0.13}$	$0.69_{0.15}$	$0.67_{0.13}$
$(0.45, 3)$	$0.6_{0.18}$	$0.84_{0.12}$	$0.84_{0.11}$	$0.82_{0.14}$	$0.59_{0.13}$	$0.66_{0.16}$	$0.65_{0.14}$	$0.68_{0.15}$	$0.67_{0.12}$
$(0.35, 5)$	$0.62_{0.19}$	$0.84_{0.12}$	$0.85_{0.11}$	$0.85_{0.12}$	$0.61_{0.13}$	$0.67_{0.16}$	$0.66_{0.14}$	$0.7_{0.16}$	$0.67_{0.12}$
$(0.25, 10)$	$0.65_{0.2}$	$0.85_{0.12}$	$0.85_{0.11}$	$0.85_{0.12}$	$0.63_{0.14}$	$0.68_{0.15}$	$0.67_{0.14}$	$0.72_{0.17}$	$0.67_{0.12}$

Color key	$0 - 0.6$	$0.6 - 0.7$	$0.7 - 0.8$	$0.8 - 0.85$	$0.85 - 0.9$	$0.9 - 0.95$	$0.95 - 1$	

Table 2.1: Rand index mean and standard deviation for $n = 20$.

Example 2.17. ***150 observations.*** *In this simulation experiment we include the ECP2 which is the ECP method without specifying the number of change points. Each simulation applies the methods described earlier to a set of* 1000 *independent sequences of* 150 *observations with two change points at* $i = 51$ *and* $i = 101$ *with distributions* G_1, G_2, G_1 *for successive blocks. To assess the performance for changes in mean, we set* $G_1 = N(\mu_1, 1)$ *and* $G_2 = N(\mu_2, 1)$, *for changes in variance we define* $G_1 = N(0, \sigma_1^2)$ *and* $G_2 = N(0, \sigma_2^2)$, *in the case of changes in tail, we specify* $G_1 = N(0, 1)$ *and* $G_2 = t_\nu$. *Finally, to study changes in skewness, we define* $G_1 = N(\mu, \sigma^2)$ *and* $G_2 = sN(0, 1, \alpha)$ *with* $\mu = Mode(G_2)$ *and* $\sigma^2 = Var(G_1)$. *Table 2.2 presents the mean and the variance of the Rand index. As in the case of 20 observations, we can see that the methods using a loss function criterion perform better than the methods using the marginal probability of change point. For changes in mean, PELT, ECP, ECP2 perform best although NPPMBs perform adequately as well. For changes in variance, our proposal performs slightly better than the other parametric and nonparametric procedures for small changes, although the other methodologies are more effective when dealing with big changes in variance. As expected, the parametric approaches assuming normality can not deal well with changes in tail or skewness. For such cases, the nonparametric approaches are most effective, NPPMs and NPPMsB being the methods that perform best.*

Change in mean ($n = 150$)										
(μ_1, μ_2)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG P**	**ECP**	**ECP2**	**PELT**
$(0, 0)$	$0.87_{0.21}$	$0.99_{0.06}$	$1_{0.03}$	$1_{0.05}$	$0.99_{0.05}$	1_{0}	$1_{0.04}$	$0.98_{0.1}$	$0.98_{0.09}$	$0.98_{0.08}$
$(0, 1)$	$0.5_{0.21}$	$0.88_{0.07}$	$0.91_{0.08}$	$0.89_{0.14}$	$0.44_{0.2}$	$0.88_{0.18}$	$0.61_{0.23}$	$0.94_{0.06}$	$0.91_{0.12}$	$0.93_{0.11}$
$(0, 2)$	$0.76_{0.21}$	$0.95_{0.12}$	$0.98_{0.02}$	$0.96_{0.12}$	$0.83_{0.2}$	$0.98_{0.02}$	$0.73_{0.02}$	$0.99_{0.01}$	$0.99_{0.02}$	$0.99_{0.02}$
$(0, 4)$	$0.97_{0.05}$	$0.98_{0.08}$	$1_{0.01}$	$1_{0.02}$	$1_{0.02}$	$1_{0.01}$	$0.99_{0.03}$	1_{0}	$1_{0.01}$	$1_{0.01}$
Change in variance ($n = 150$)										
(σ_1^2, σ_2^2)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG P**	**ECP**	**ECP2**	**PELT**
$(1, 3)$	$0.63_{0.2}$	$0.86_{0.1}$	$0.85_{0.06}$	$0.81_{0.15}$	$0.5_{0.23}$	$0.79_{0.2}$	$0.74_{0.17}$	$0.77_{0.11}$	$0.47_{0.22}$	$0.83_{0.2}$
$(1, 5)$	$0.68_{0.19}$	$0.91_{0.07}$	$0.82_{0.19}$	$0.76_{0.24}$	$0.7_{0.22}$	$0.87_{0.18}$	$0.83_{0.1}$	$0.87_{0.1}$	$0.69_{0.26}$	$0.95_{0.06}$
$(1, 7)$	$0.79_{0.09}$	$0.89_{0.17}$	$0.88_{0.12}$	$0.86_{0.1}$	$0.8_{0.16}$	$0.94_{0.09}$	$0.86_{0.07}$	$0.92_{0.08}$	$0.81_{0.23}$	$0.96_{0.03}$
$(1, 10)$	$0.78_{0.17}$	$0.93_{0.03}$	$0.9_{0.14}$	$0.89_{0.13}$	$0.79_{0.09}$	$0.96_{0.04}$	$0.87_{0.04}$	$0.95_{0.06}$	$0.92_{0.13}$	$0.98_{0.02}$
Change in tail ($n = 150$)										
ν	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG p**	**ECP**	**ECP2**	**PELT**
2.1	$0.61_{0.21}$	$0.82_{0.11}$	$0.81_{0.09}$	$0.77_{0.15}$	$0.74_{0.18}$	$0.76_{0.18}$	$0.68_{0.2}$	$0.74_{0.1}$	$0.38_{0.14}$	$0.78_{0.17}$
3	$0.51_{0.21}$	$0.77_{0.15}$	$0.79_{0.09}$	$0.76_{0.12}$	$0.61_{0.23}$	$0.69_{0.2}$	$0.54_{0.22}$	$0.71_{0.11}$	$0.35_{0.1}$	$0.72_{0.19}$
4	$0.48_{0.19}$	$0.78_{0.1}$	$0.79_{0.07}$	$0.77_{0.1}$	$0.51_{0.22}$	$0.65_{0.2}$	$0.46_{0.2}$	$0.72_{0.1}$	$0.35_{0.1}$	$0.71_{0.18}$
10	$0.44_{0.17}$	$0.79_{0.06}$	$0.78_{0.07}$	$0.77_{0.08}$	$0.37_{0.12}$	$0.69_{0.15}$	$0.35_{0.09}$	$0.7_{0.11}$	$0.35_{0.09}$	$0.68_{0.17}$
Change in skewness ($n = 150$)										
(μ, σ, α)	**NPPMsP**	**NPPMs**	**NPPMsB**	**NPPMsBB**	**bcp**	**NG**	**NG p**	**ECP**	**ECP2**	**PELT**
$(0.5, 0.7, 2)$	$0.45_{0.18}$	$0.78_{0.09}$	$0.8_{0.08}$	$0.77_{0.13}$	$0.34_{0.07}$	$0.68_{0.16}$	$0.34_{0.05}$	$0.72_{0.13}$	$0.39_{0.14}$	$0.7_{0.17}$
$(0.45, 0.65, 3)$	$0.46_{0.18}$	$0.81_{0.07}$	$0.83_{0.07}$	$0.77_{0.17}$	$0.35_{0.1}$	$0.61_{0.22}$	$0.34_{0.06}$	$0.74_{0.12}$	$0.39_{0.15}$	$0.76_{0.11}$
$(0.35, 0.62, 5)$	$0.53_{0.22}$	$0.86_{0.08}$	$0.89_{0.05}$	$0.77_{0.24}$	$0.47_{0.21}$	$0.86_{0.08}$	$0.4_{0.16}$	$0.86_{0.1}$	$0.76_{0.16}$	$0.81_{0.07}$
$(0.25, 0.6, 10)$	$0.51_{0.21}$	$0.86_{0.08}$	$0.89_{0.05}$	$0.78_{0.23}$	$0.45_{0.19}$	$0.85_{0.12}$	$0.38_{0.14}$	$0.88_{0.11}$	$0.7_{0.26}$	$0.82_{0.07}$

Color key	**0 – 0.6**	**0.6 – 0.7**	**0.7 – 0.8**	**0.8 – 0.85**	**0.85 – 0.9**	**0.9 – 0.95**	**0.95 – 1**	

Table 2.2: Rand index mean and standard deviation for $n = 150$.

Example 2.18. ***300 observations.*** *We use the same settings of the experiment of 150 observations with change points at* $i = 101$ *and* $i = 201$*. Comparing Table 2.2 and Table 2.3, we can observe better performance when the number of observations increases. We notice the same behavior presented in the* 150*-observation experiment although the nonparametric product partition models with loss function using weights seems to better detect the change points than the ones which do not.*

Change in mean ($n = 300$)										
(μ_1, μ_2)	NPPMsP	NPPMs	NPPMsB	NPPMsBB	bcp	NG	NG P	ECP	ECP2	PELT
$(0, 0)$	$0.76_{0.25}$	$0.98_{0.09}$	$1_{0.02}$	$0.99_{0.05}$	$1_{0.03}$	1_{0}	$1_{0.04}$	$0.98_{0.08}$	$0.98_{0.08}$	$0.98_{0.08}$
$(0, 1)$	$0.5_{0.2}$	$0.86_{0.16}$	$0.94_{0.06}$	$0.93_{0.1}$	$0.4_{0.16}$	$0.94_{0.13}$	$0.67_{0.2}$	$0.97_{0.03}$	$0.97_{0.04}$	$0.97_{0.03}$
$(0, 2)$	$0.76_{0.22}$	$0.91_{0.15}$	$0.99_{0.01}$	$0.97_{0.12}$	$0.86_{0.18}$	$0.99_{0.01}$	$0.7_{0.01}$	$1_{0.01}$	$0.99_{0.01}$	$0.99_{0.02}$
$(0, 4)$	$0.98_{0.04}$	$0.73_{0.09}$	$1_{0.01}$	$1_{0.01}$	$1_{0.03}$	1_{0}	$1_{0.02}$	1_{0}	$1_{0.01}$	$0.99_{0.01}$
Change in variance ($n = 300$)										
(σ_1^2, σ_2^2)	NPPMsP	NPPMs	NPPMsB	NPPMsBB	bcp	NG	NG P	ECP	ECP2	PELT
$(1, 3)$	$0.73_{0.14}$	$0.88_{0.04}$	$0.88_{0.1}$	$0.89_{0.08}$	$0.52_{0.24}$	$0.88_{0.15}$	$0.81_{0.1}$	$0.79_{0.1}$	$0.52_{0.23}$	$0.94_{0.05}$
$(1, 5)$	$0.55_{0.23}$	$0.87_{0.2}$	$0.93_{0.11}$	$0.88_{0.19}$	$0.78_{0.19}$	$0.95_{0.09}$	$0.85_{0.06}$	$0.94_{0.07}$	$0.93_{0.12}$	$0.97_{0.03}$
$(1, 7)$	$0.63_{0.24}$	$0.93_{0.09}$	$0.93_{0.15}$	$0.9_{0.13}$	$0.85_{0.11}$	$0.97_{0.06}$	$0.86_{0.04}$	$0.96_{0.05}$	$0.96_{0.08}$	$0.98_{0.02}$
$(1, 10)$	$0.77_{0.19}$	$0.9_{0.19}$	$0.92_{0.18}$	$0.87_{0.21}$	$0.78_{0.07}$	$0.98_{0.03}$	$0.86_{0.03}$	$0.98_{0.02}$	$0.98_{0.02}$	$0.99_{0.02}$
Change in tail ($n = 300$)										
ν	NPPMsP	NPPMs	NPPMsB	NPPMsBB	bcp	NG	NG p	ECP	ECP2	PELT
2.1	$0.57_{0.22}$	$0.86_{0.09}$	$0.86_{0.09}$	$0.82_{0.13}$	$0.81_{0.11}$	$0.84_{0.16}$	$0.77_{0.15}$	$0.73_{0.1}$	$0.39_{0.15}$	$0.87_{0.07}$
3	$0.45_{0.19}$	$0.85_{0.07}$	$0.85_{0.06}$	$0.84_{0.08}$	$0.71_{0.2}$	$0.79_{0.16}$	$0.66_{0.21}$	$0.71_{0.11}$	$0.36_{0.1}$	$0.8_{0.15}$
4	$0.45_{0.18}$	$0.78_{0.14}$	$0.82_{0.08}$	$0.81_{0.08}$	$0.56_{0.23}$	$0.73_{0.18}$	$0.53_{0.22}$	$0.72_{0.1}$	$0.37_{0.12}$	$0.75_{0.17}$
10	$0.42_{0.16}$	$0.75_{0.13}$	$0.8_{0.06}$	$0.78_{0.07}$	$0.37_{0.12}$	$0.69_{0.14}$	$0.37_{0.12}$	$0.68_{0.13}$	$0.35_{0.09}$	$0.69_{0.17}$
Change in skewness ($n = 300$)										
(μ, σ, α)	NPPMsP	NPPMs	NPPMsB	NPPMsBB	bcp	NG	NG p	ECP	ECP2	PELT
$(0.5, 0.7, 2)$	$0.43_{0.16}$	$0.77_{0.09}$	$0.82_{0.08}$	$0.81_{0.09}$	$0.34_{0.05}$	$0.75_{0.11}$	$0.34_{0.05}$	$0.72_{0.14}$	$0.39_{0.16}$	$0.77_{0.09}$
$(0.45, 0.65, 3)$	$0.45_{0.18}$	$0.81_{0.08}$	$0.86_{0.08}$	$0.85_{0.09}$	$0.36_{0.11}$	$0.74_{0.19}$	$0.35_{0.1}$	$0.75_{0.14}$	$0.45_{0.2}$	$0.78_{0.05}$
$(0.35, 0.62, 5)$	$0.56_{0.2}$	$0.79_{0.14}$	$0.9_{0.06}$	$0.88_{0.06}$	$0.39_{0.15}$	$0.88_{0.08}$	$0.36_{0.11}$	$0.88_{0.11}$	$0.73_{0.26}$	$0.79_{0.2}$
$(0.25, 0.6, 10)$	$0.49_{0.2}$	$0.84_{0.14}$	$0.93_{0.04}$	$0.93_{0.06}$	$0.41_{0.17}$	$0.92_{0.05}$	$0.38_{0.14}$	$0.95_{0.06}$	$0.94_{0.1}$	$0.87_{0.13}$

Color key	0 – 0.6	0.6 – 0.7	0.7 – 0.8	0.8 – 0.85	0.85 – 0.9	0.9 – 0.95	0.95 – 1	

Table 2.3: Rand index mean and standard deviation for $n = 300$.

2.4.2 Sensitivity Analysis

In this section we include a sensitivity analysis of our model. We carried out 100 simulations for each of the following combinations of the parameters.

$\alpha = 1, 30$ $c_{NP0} = 0.001, 0.002$ $p_{NP0} = 0.01, 0.05$ using the following cohesion functions:

$$c_{ij} = p_{NP0}(1 - p_{NP0})^{j-i-1} c_{NP0}$$

For each simulation, we generated a data set as follows:

$$X_i \overset{ind}{\sim} N(\mu_i, 1) \begin{cases} \mu_i = 0, & \text{if } 1 \leq i \leq 50 \\ \mu_i = 4, & \text{if } 51 \leq i \leq 100 \\ \mu_i = 0, & \text{if } 101 \leq i \leq 150. \end{cases}$$

Each simulation applies the methods described earlier to a set of 100 independent sequences of 150 observations with two change points at $i = 51$ and $i = 101$ with distributions $N(\mu_i = 0, 1)$, $N(\mu_i = 4, 1)$, $N(\mu_i = 0, 1)$ for successive blocks.

In Tables 2.4 and 2.5 we present the histograms of the frequency the change points estimated for each method proposed, with different parameters α, cohesion functions and probabilities of change.

The number of change points detected by the model NNPPMsP is quite unstable when $\alpha = 1$; this may be explained by the flexibility and variability of the Dirichlet Process for small values of α. Using loss functions to estimate the change points is more stable for a wider choice of parameters.

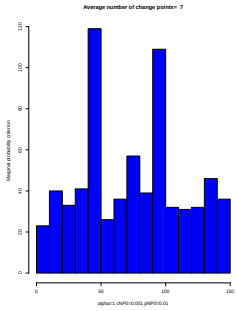
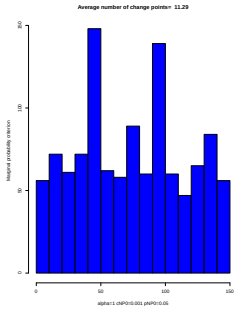
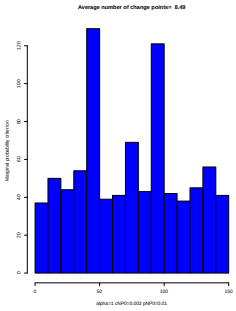
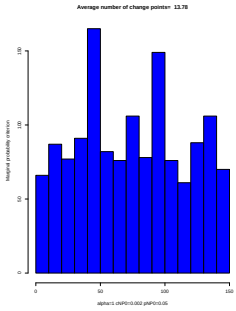
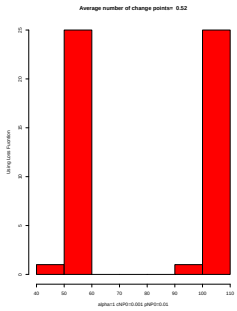
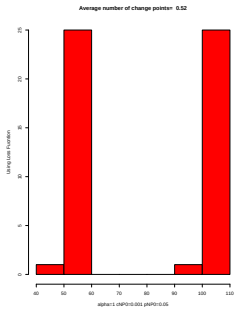
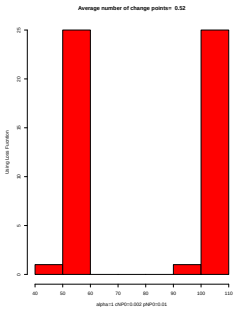
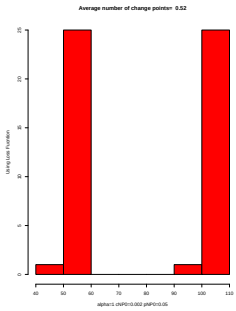
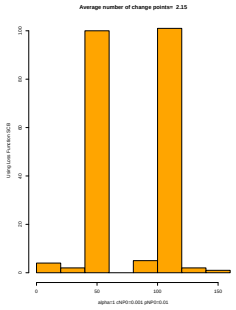
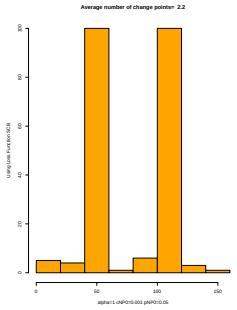
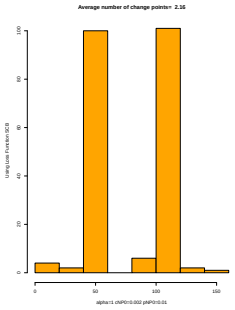
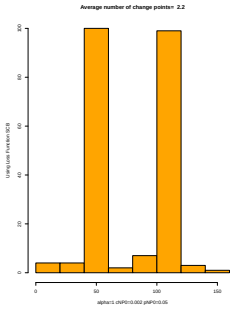
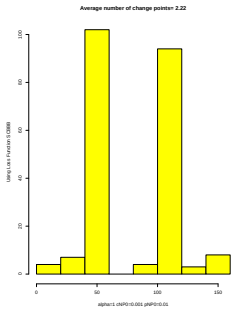
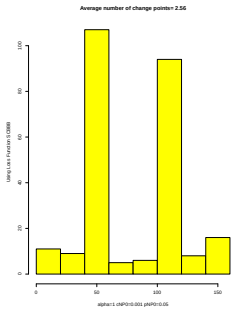
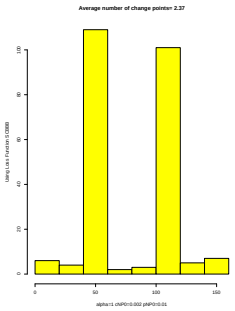
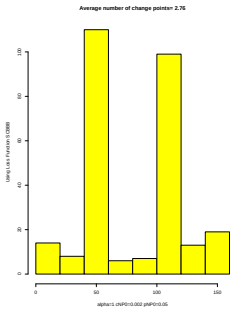
	$\alpha = 1$ $c_{NP0} = 0.001$ $p_{NP0} = 0.01$	$\alpha = 1$ $c_{NP0} = 0.001$ $p_{NP0} = 0.05$	$\alpha = 1$ $c_{NP0} = 0.002$ $p_{NP0} = 0.01$	$\alpha = 1$ $c_{NP0} = 0.002$ $p_{NP0} = 0.05$
NPPMsP				
NPPMs				
NPPMsB				
NPPMsBB				

Table 2.4: Histogram of the change points detected in the simulations

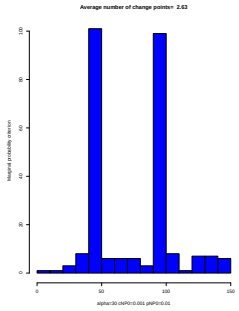
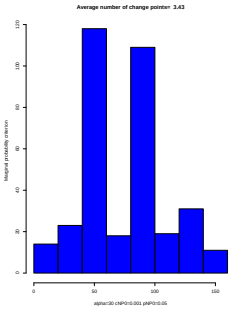
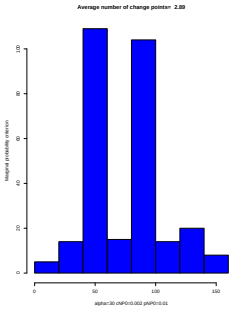
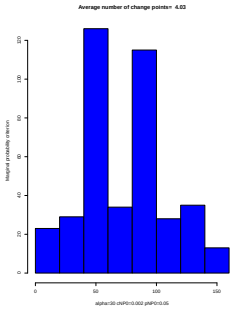
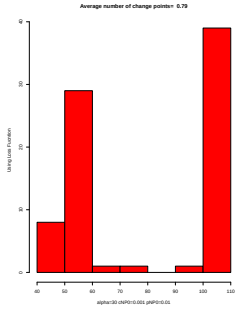
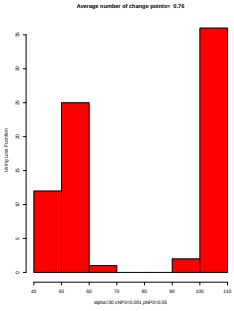
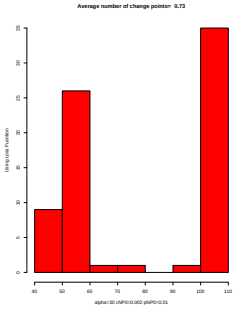
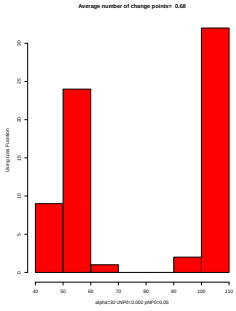
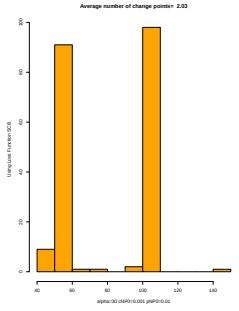
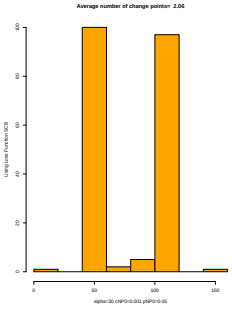
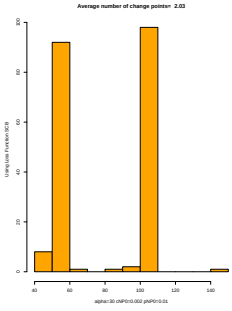
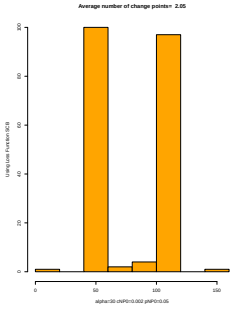
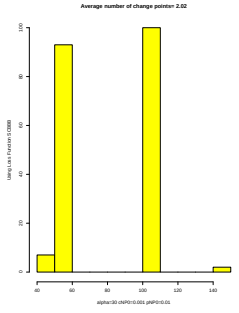
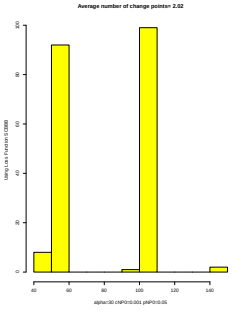
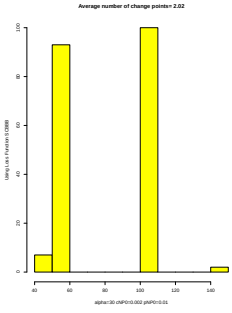
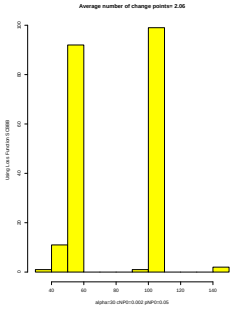
	$\alpha = 30$ $c_{NP0} = 0.001$ $p_{NP0} = 0.01$	$\alpha = 30$ $c_{NP0} = 0.001$ $p_{NP0} = 0.05$	$\alpha = 30$ $c_{NP0} = 0.002$ $p_{NP0} = 0.01$	$\alpha = 30$ $c_{NP0} = 0.002$ $p_{NP0} = 0.05$
NPPMsP				
NPPMs				
NPPMsB				
NPPMsBB				

Table 2.5: Histogram of the change points detected in the simulations

2.4.3 Applications to real data sets

2.4.3.1 Dow Jones industrial average

We will study the weekly closing values of the Dow-Jones industrial average over the period from July 1st, to August 2nd, 1974 (the week before former President Nixon resigned). The data can be downloaded from the R package *strucchange* (Zeileis et al., 2002). We transform the data into a series of rates return,

$$R_t = \frac{P_{t+1}}{P_t} - 1$$

where P_t are the index values at week t, $t = 1, 2, ..., 161$. For the analysis, we will assume as (Hsu, 1979), that the values are exchangeable.

Figure 2.9 presents the number of change points detected by the NPPM, NPPMB and NPPMBB versus SSE_γ for the Dow Jones data set described before. We can see that NPPMB and NPPMBB have almost the same values, while NPPM presents different behavior. We maximize the difference of SSE when we change more zero change points to one change point for models NPPMB and NPPMBB. In the case of NPPM, we maximize this difference when we change from 2 to 3 change points. Therefore, for models NPPMB and NPPMBB, we choose γ in such a way that we obtain one change point. In the case of NPPM, we select this value to obtain three change points. Table 2.6 shows change points detected by different methods and their respective SSE. Although PELT, ECP, NPPMB and NPPMBB detect only one point, the last two have minimum SSE. Since NPPM detects three, naturally, it is the method with minimum SSE. Figure 2.10 shows the change points detected by the methods studied. In Figure 2.11, we can observe the estimated distribution for each data point and distinguish mainly two groups of distributions, the first one in color blue and the second one in green. Although both groups appear to have the same mean, the second one has larger variance. The distribution corresponding to the change point found by NPPMB and NPPMBB is indicated in red.

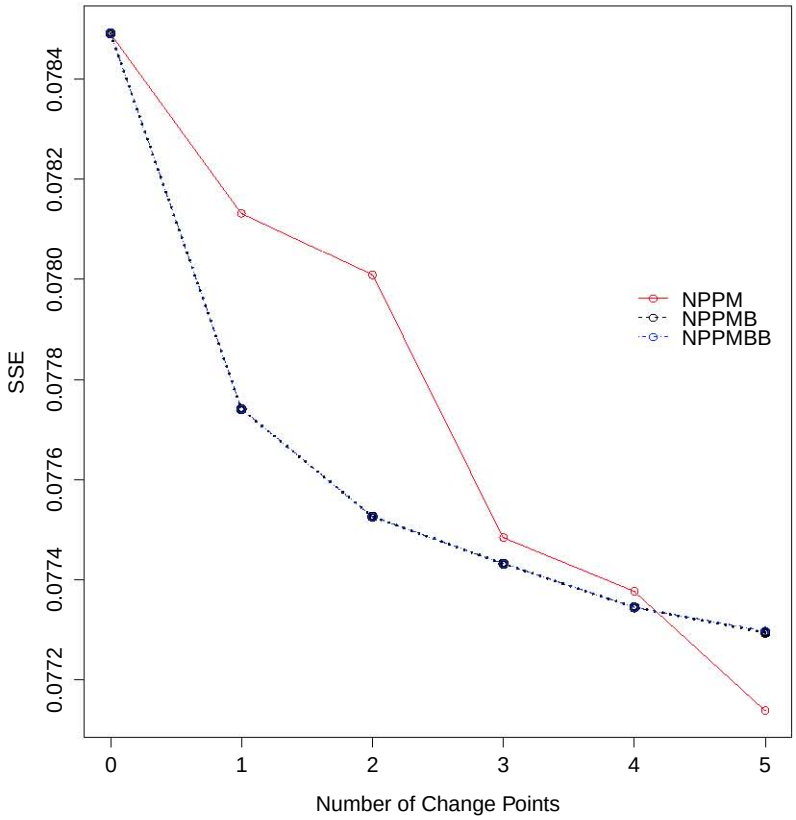


Figure 2.9: Number of Change Points vs SSE for NPPM, NPPMB and NPPMBB.

Method	Number of Change Points	Change Points	SSE
NPPM	3	$\{24, 71, 91\}$	0.07748426
NPPMB	1	84	0.07774066
NPPMBB	1	84	0.07774066
PELT	1	90	0.07784357
ECP	1	90	0.07784357

Table 2.6: Dow Jones data and change points detected by different methods.

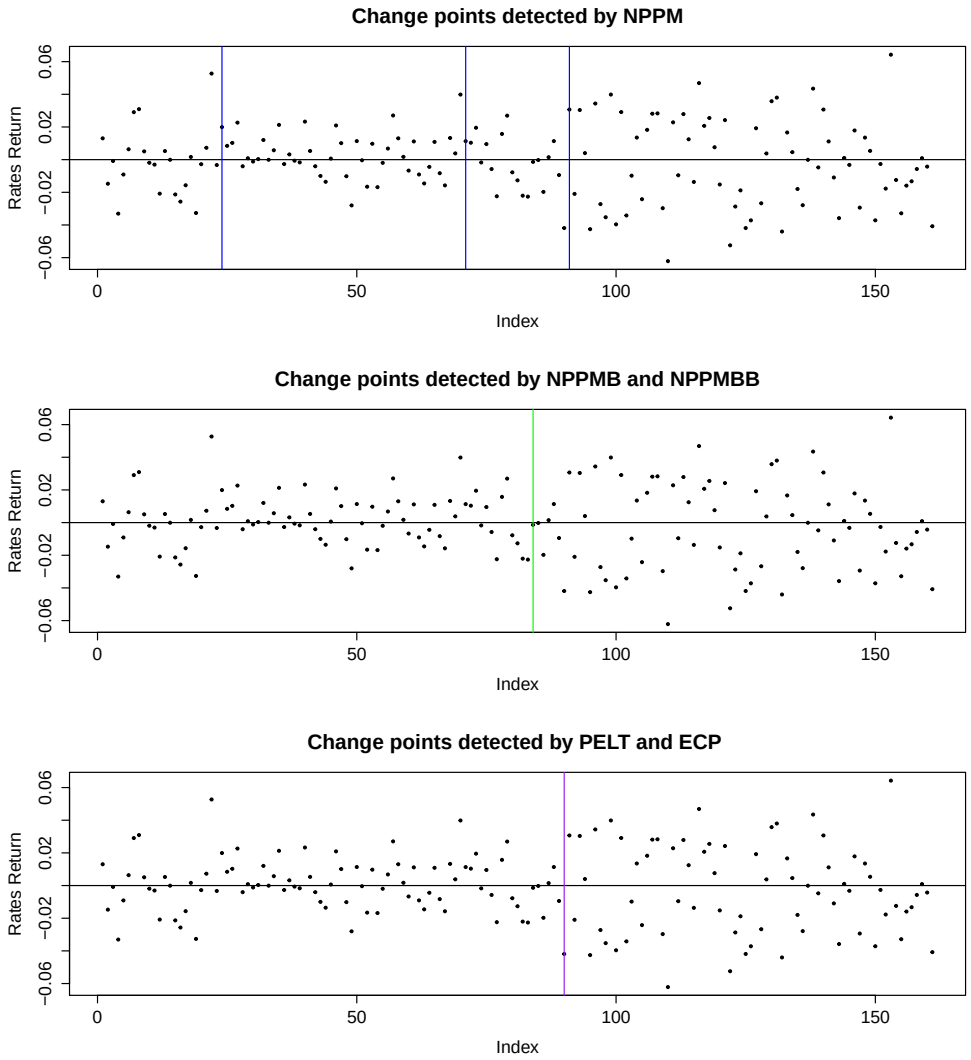


Figure 2.10: Change points detected by NPPM, NPPMB, NPPMBB, PELT and ECP.

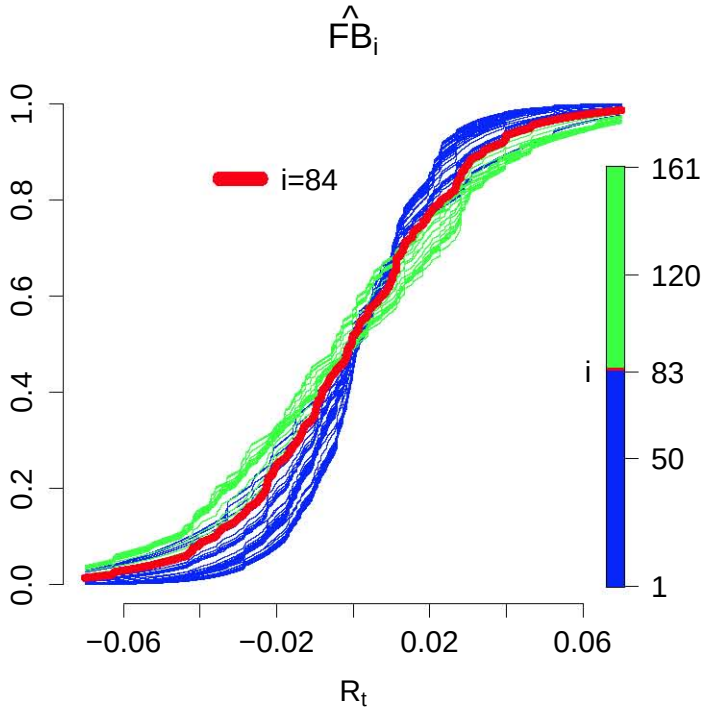


Figure 2.11: Estimated distributions ($\hat{F}_{B,i}$) for $i = 1, ..., 161$.

2.4.3.2 Human genome

Biological and medical research reveals that some forms of cancer are caused by somatic or inherited mutations in oncogenes and tumor suppressor genes; cancer development and genetic disorders often result in chromosomal DNA copy number changes or copy number variations (CNVs). Consequently, identification of these loci where the DNA copy number changes or CNVs have taken place will (at least partially) facilitate the development of medical diagnostic tools and treatment regimes for cancer and other genetic diseases (Chen and Gupta, 2011).

Copy number variation can be discovered by cytogenetic techniques such as array comparative genomic hybridization (aCGH), which consists of the steps described in Figure 2.12.

Missing values are common in this type of data and occur for diverse reasons, including insufficient resolution, image corruption, or simply due to dust or scratches

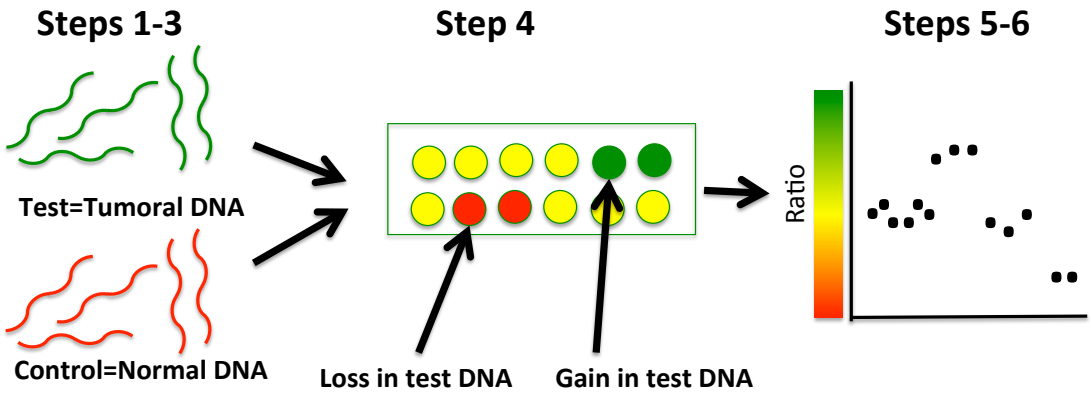


Figure 2.12: Diagram of the microarray-based comparative genomic hybridization process. Steps 1-3: Test and control DNA are labeled with fluorescent dyes, combined equal amounts of DNA and applied to the microarray. Step 4: Test and control DNA compete to attach, or hybridize, to the microarray. Steps 5-6: The microarray scanner measures the fluorescent signals and computer software calculates the Log-Ratios of the fluorescence intensities of the test and reference samples along the chromosome.

on the slide. Missing data may also occur systematically as a result of the robotic methods used to create them (Troyanskaya et al., 2001).

Snijders et al. (2001) performed aCGH experiments on 15 fibroblast cell lines and obtained normalized averages of the Log-Ratios. This data has been studied extensively in the literature and can be downloaded from several sources, such as the GLAD package of the R software (Hupe, 2011). Next, we present the analysis of the genome gm01524 of the Snijders database using NPPMs and compare it with ECP and PELT. Because these methodologies do not consider missing data, we imputed the absent values as the average of their neighboring observations. To apply our methodology, we assume that missing values are at random which is consistent with Figure 2.13 where we do not recognize any pattern at the points where a missing value is present.

Figure 2.14 shows the distribution functions estimated at each position of the genome. We can see that all the chromosome positions share the same distribution function with different mean, as assumed by several algorithms for change-points detection for aCGH data.

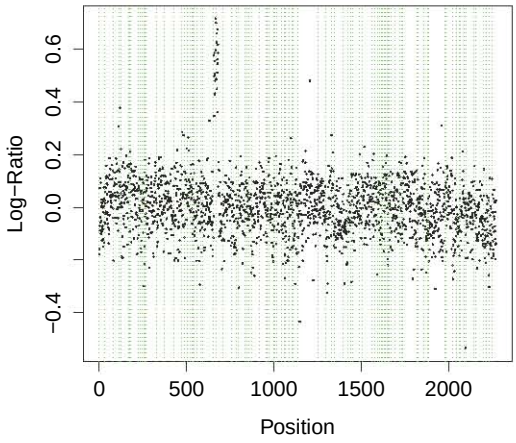


Figure 2.13: Genome gm01524 of Snijders data (2271 observations including 112 missing values). Green lines indicates missing values.

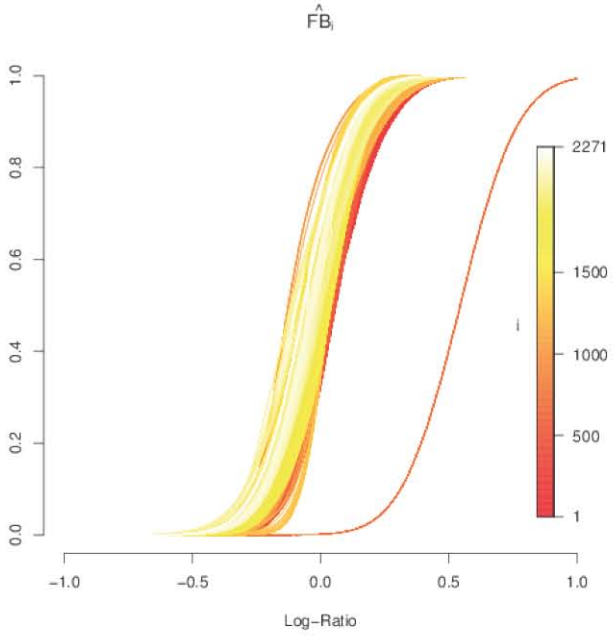


Figure 2.14: Estimated distributions for each data point. Snijders data.

We will specify the number of change points using the graphic of the sum of square errors displayed in Figure 2.15. The number of change points depends on the "elbow" point at which the remaining SSE are relatively small and all about the same size. This point is not very evident in Figure 2.15, but we can still say that the eighteenth point is our "elbow" point.

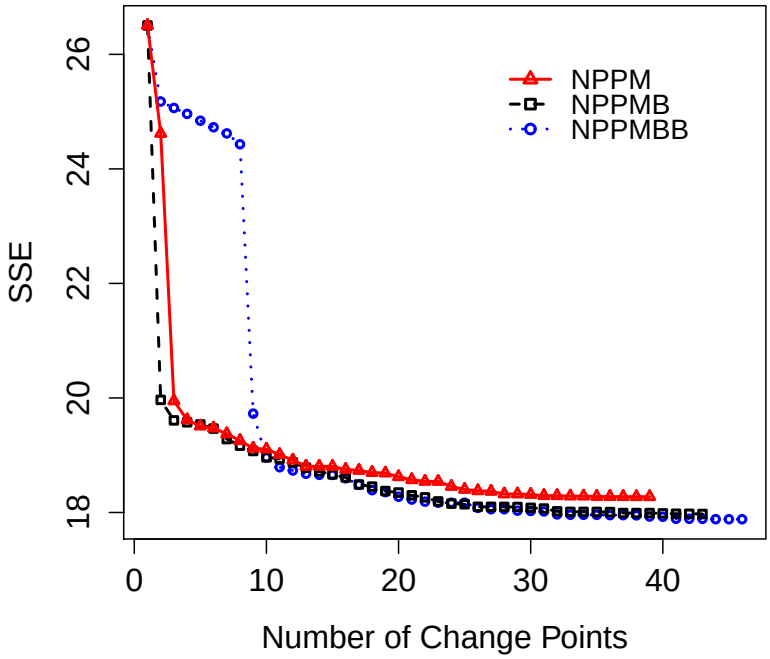


Figure 2.15: Number of change points vs SSE for NPPM, NPPMB and NPPMBB in Snijders et al. data.

Table 2.7 shows the number of changes points obtained by ECP and PELT. We compare the SSE obtained by these methods with the SSE obtained by NPPMs for the same number of change points.

Number of Change Points	PELT	ECP	NPPM	NPPMB	NPPMBB
18	-	17.10636	18.6968	18.45198	18.39047
23	18.41243	-	18.54185	18.19333	18.17079

Table 2.7: SSE and number of change points detected by different methods for the Snijders data.

Unlike NPPMs, ECP and PELT do not consider the variability of missing values, which can explain the differences in Table 2.7. Figure 2.16 shows change points detected by different methods.

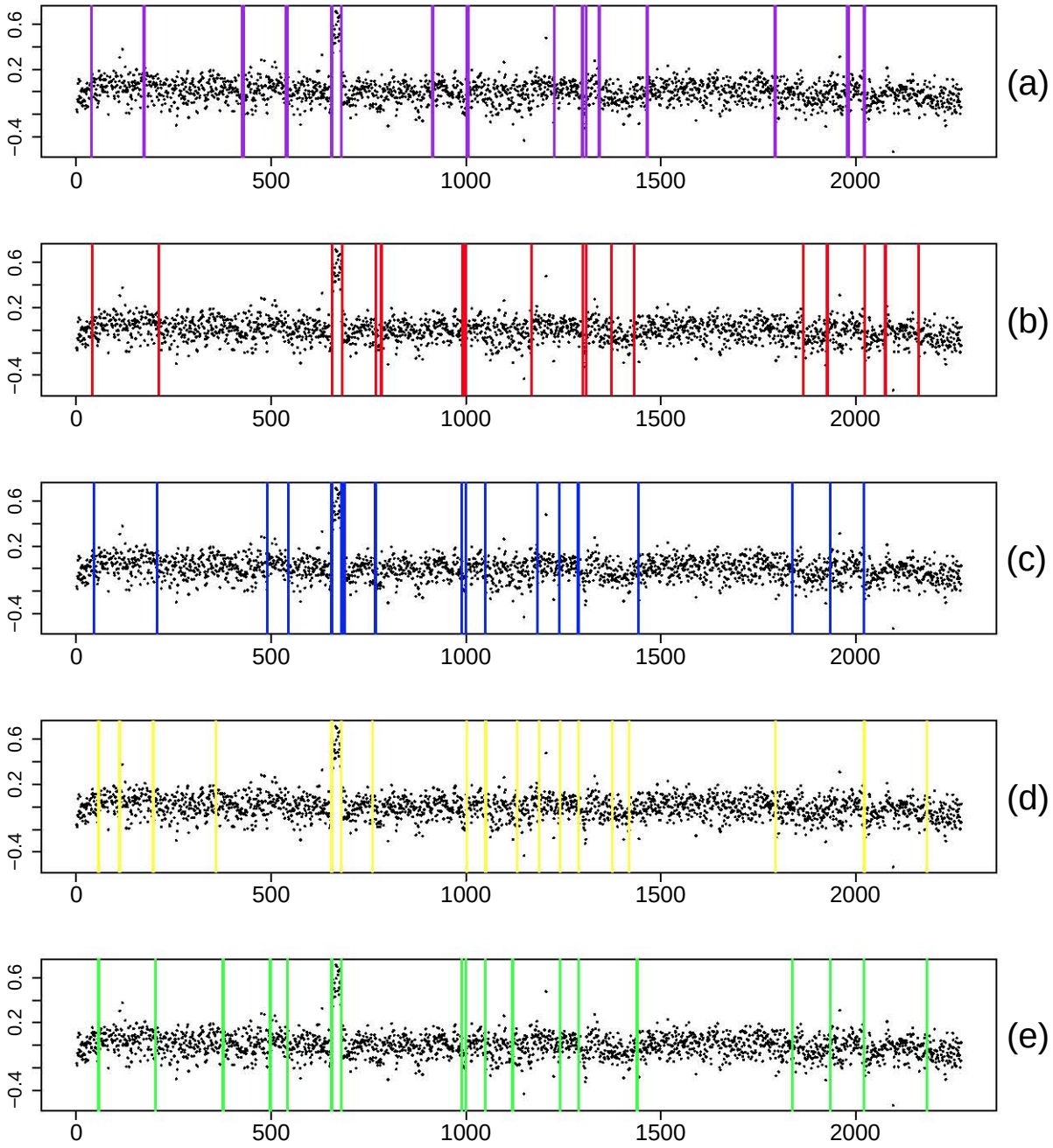


Figure 2.16: Snijders data and change points detected by different methods. (a) PELT (b) ECP (c) NPPM (d) NPPMB (e) NPPMBB.

2.5 Discussion

In this chapter, we used the Dirichlet process to extend the PPMs to the NPPMs which does not impose any particular parametric form on the distribution function. We also provided a methodology which uses loss functions to detect change points that can also be applied to other models such as the nonparametric hidden Markov models of Yau et al. (2011). For the loss function proposed in equation (2.11), we suggested a procedure to determine the value of the parameter γ, which appears in similar loss functions in different settings such as, for example, variable selection (Hahn and Carvalho, 2015).

In many applications such as the analysis of aCGH data, missing data occurs frequently. To tackle this difficulty, we took advantage of the random partition structure of our model in order to estimate the distribution function of each missing value without using multiple imputation (which is computationally less efficient). Then, we were able to detect the change points using the loss functions described earlier. This procedure can also be used in the parametric product partition models for change-point analysis.

Also, we established a relationship between NDP and NPPMs in the same way that the PPMs are related to the Dirichlet process: By integrating out the Dirichlet process in the NDP, we obtained a particular case of our model.

Finally, we have shown through simulations that methods based on loss functions may perform better that the ones using the marginal probability. This is specially true in the case of the NPPMs, in which this criterion detects change points very poorly. Moreover, we have shown for different data sizes that our proposal performs better than parametric and nonparametric models recently discussed in the literature.

Mena and Ruggiero (2016) discuss change-point analysis in a nonparametric setting; they ask: If the model is too flexible, does it make sense to define a change point? We argue that our loss function allows us to define it because we measure the shift using a distance between distributions.

Chapter 3

A New Approach to Bayesian Post-Stratification

Those who have no compassion have no wisdom. Knowledge, yes; cleverness, maybe; wisdom, no. A clever mind is not a heart. Knowledge doesn't really care. Wisdom does.

Benjamin Hoff

In the context of survey sampling, post-stratification is a reweighting of the sample using an auxiliary variable which is highly correlated with the target variable. The basic idea is that, if we know the population is composed of distinct groups (strata) that differ with regard to the quantity we are interested in estimating and we know the sizes of these strata in our population, then we can obtain a more accurate estimate of the quantity of interest by correcting for any imbalance in the representation of the strata in the sample. This correction is obtained by using a weighted average (using the known weights from the population) of the averages within strata as our estimate of the population mean. The aim of this chapter is to provide a Bayesian framework to model the prior knowledge concerning the structure of the population using random partition models. We will then be able to combine many polls to obtain better estimates through a Bayesian learning process.

3.1 Finite Population Sampling

In this section we will give a brief overview of finite population sampling.

3.1.1 Methods in finite population sampling

In finite population sampling there exist two main approaches to estimating the quantities of interest: design-based and model-based inference. Little (2004) discusses the main differences between them. More precisely, he divides these procedures as follows:

- Design-based inference.
- Model-based inference.
  - Superpopulation modelling.
  - Bayesian modelling.

In the following subsections we will describe each one briefly. Little proposes a third view, in which the researcher takes into account the sample design and makes weak parametric assumptions that can produce reliable and efficient inferences in survey settings.

Design-based inference

Pure design-based inference is the most common in traditional sampling theory. Well-known sampling texts such as Cochran (1977) use this type of approach. Here, the population of interest is considered as a finite collection of elements. Design-based inference assumes that the population is fixed. Each sample is viewed as a realization of a random process, so a different sample may have chosen a different set of units. The probabilistic nature of the sample is the only source of randomness that plays a part when making inference about the population.

Design-based inference has several advantages:

- It takes into account the survey design.
- Accurate inferences in large samples without making strong assumptions such as a distribution of the data.
- Easy computational implementation.
- Popularity among practical statisticians.

Model-based inference

Unlike design-based inference, pure model-based inference does not regard the population of interest as fixed. Here we assume that an infinite superpopulation or superpopulation model, which includes a random component, is responsible for creating the elements in the finite population. One way to think of the superpopulation model is as the process used to create the elements in the population. The specific form of the proposed infinite superpopulation model is often borrowed from classical methods such as regression.

Superpopulation modelling

In the superpopulation approach we assume that the population is a random sample from a model. We assume a probability distribution $p(\mathrm{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters that we want to study in order to characterize the population. A relevant example studied in Little (2004) is model-based inference for the mean from a stratified random sample. We want to estimate the mean of the population Y and we have a stratification based on the discrete variable X. We will consider the basic normal post-stratification model. Let y_i denote the value of Y for unit i in the sample. Let x_i be the value of X for unit i. We will assume that x_i given $z_i = h$ is normal with mean μ_h and variance σ_h^2. Suppose we do not have any prior knowledge about μ_h and σ_h^2. A simple Bayesian model that reflects this is

$$\begin{aligned} p(y_i|x_i = h, \mu_h, \sigma_h^2) &\overset{independent}{=} N(\mu_h, \sigma_h^2) \\ p(\mu_h, \log \sigma_h) &\propto const \end{aligned}$$

In this model the structure is fixed. Suppose that we know σ_h^2; then standard calculations yield

$$E(\hat{\overline{y}}|x_1, \ldots, x_n, \sigma_h^2) = \overline{y}_{st} = \sum_{h=1}^{H} P_h \overline{y}_h$$

$$Var(\hat{\overline{y}}|x_1, \ldots, x_n, \sigma_h^2) = v_{st} = \sum_{h=1}^{H} P_h^2 \sigma_h^2 (\frac{1}{n_h} - \frac{1}{N_h})$$

where P_h is the probability of sampling from stratum h, n_h is the number of observations in the sample that belong to stratum h and N_h is the number of individuals in the

population that belong to stratum h. We name this model the basic normal post-stratification model (BNPM). When we replace the $\{\sigma_h^2\}$ by the estimates $\{s_h^2\}$, we obtain the same estimates as in design-based modelling. This procedure is often justified when we have large samples.

Note that integrating over the posterior distribution of $\{\sigma_h^2\}$ rather than simply plugging in estimates yields a useful small-sample correction not readily available from design-based and superpopulation approaches (see Little 2004).

Bayesian modelling

Ericson (1969) proposed a Bayesian approach using Pólya urns to predict the non observed values of the population given the data (see section 1.3). Suppose we have a population of size N and we observe a random sample of size n. In a first urn we put the n observed values $x_1, \ldots, x_n$. In a second urn we put $N - n$ balls without any values. These balls represent the unknown values of the rest of the population $X_{n+1}, \ldots, X_N$. In the next step, we extract a ball from each urn, then we assign the value of the ball of the first urn to the ball of the second one. Next we put the two balls in the first urn. Now we will have $n+1$ balls in the first urn, and $N-n-1$ balls in the second one. We continue this process until the second urn is empty. We thus obtain a simulation of $P(x_{n+1}, \ldots, x_N | x_1, \ldots, x_n)$. Suppose we have an infinite population and we have a random sample of size n. Suppose also that we believe that the distribution of the infinite population is G.

Blackwell and MacQueen (1973) extended the Pólya urn model as follows: In the first urn we put the n observed values $x_1, \ldots, x_n$, in the second urn we put an infinite number of balls without any values; as before, these balls represent the unknown values of the rest of the population. In the next step we select a ball from the first urn with probability $\dfrac{n}{n+\alpha}$ and with probability $\dfrac{\alpha}{n+\alpha}$ a new value from G; then we extract a ball from the second urn and assign the value obtained to the ball. Then we put this ball in the first urn. Now we have $n+1$ balls in the first urn. The parameter α is a dispersion parameter. With this procedure, we obtain the Dirichlet process $DP(\alpha, G)$ defined by Ferguson (1973) and discussed in section 1.3.

The relationship between finite population sampling and Bayesian non-parametric statistics arises naturally from these Pólya urn schemes. Despite this link, there are only a few papers that relate these two fields of statistics. An example can be found

in Binder (1982) who uses Bayesian non-parametric models for estimating population percentiles and develops a procedure for interval estimates in stratified sampling.

Lo (1986) studies the Dirichlet multinomial model for finite population sampling. He obtains the following result: *If the population size tends to infinity* (*the sample size is fixed*), *sampling without replacement from a Dirichlet multinomial process is equivalent to the iid sampling from a Dirichlet process.* In general, we feel that Bayesian inference for finite population sampling is a challenging area of study that needs more attention from the Bayesian community.

3.1.2 Stratification

Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling. If the stratification is adequate, it helps us obtain more accurate estimates because we include correct information in the inference. Under specific circumstances, stratification yields better estimates than the ones obtained with simple random sampling, but this is not always the case.

Optimal stratification

What is the best characteristic for the construction of the strata? How should the boundaries between the strata be determined? How many strata should there be? In Cochran (1977) the author gives some advices concerning these questions. Given the number of strata, the equations for determining the best stratum boundaries have been studied by Dalenius and Hodges (1959) under proportional and optimal allocation (the size of each stratum is proportional to the standard deviation of the distribution of the variable; hence larger samples are taken in the strata with the greatest variability to generate the least possible sampling variance). An interesting problem arises when we are dealing with multivariate variables. How can we choose the optimal stratification when we have continuous and discrete random variables together?. One possibility is to use Bayesian trees (see Denison et al. 2002) which induce a partition over both discrete and continuous covariates.

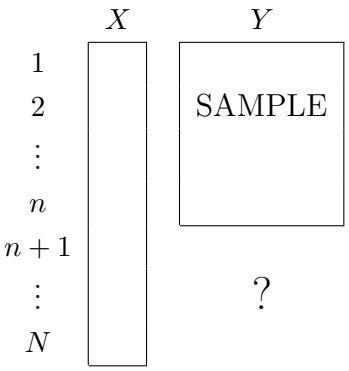


Figure 3.1: Graphical representation of post-stratification

3.1.3 Post-Stratification, weighting and calibration

Post-Stratification, weighting and calibration are related methods to use auxiliary information available on the whole population. For example, if our sample consists of 60% women, when we know from a census that in the population women are in fact 52%, this may introduce bias into the estimate of interest because we will give greater weight to those people we oversampled. In this section we will describe some of the main methods to use auxiliary variables in order to obtain better estimates.

3.1.3.1 Post-Stratification

When we are conducting a survey it is very important to obtain a representative sample of the population. But sometimes we oversample some kinds of observations and undersample others. As mentioned above, this may introduce bias into the estimate of the quantity of interest. We can correct these biases statistically with a post-stratification survey weight. In order to calculate a post-stratification weight, we need an auxiliary variable X to which we can compare our sample data. Then we can estimate a parameter of the target variable Y. Figure 3.1 represents this situation.

Selection of post-strata: Considering all post-strata is not always practical because of problems of empty strata and lack of sufficient population information. Little (1993) develops an algorithm to collapse post-strata. He uses the reduction of the variance of the estimated mean as a criterion to form the post-strata. Let us consider the BNPM. Collapsing two post-strata h_1 and h_2 is interpreted as combining the population

proportions P_{h_1} and P_{h_2} and modifying the model accordingly. The posterior mean under this collapsed model is

$$\overline{y}_{ps}^{(h_1h_2)} = \sum_{h \neq h_1,h_2} P_h \overline{y}_h + (P_{h_1} + P_{h_2})\overline{y}_{(h_1h_2)},$$

where $\overline{y}_{h_1h_2}$ is the mean of the collapsed post-strata h_1 and h_2. The posterior variance of the collapsed post-strata is denoted by $v_{ps}^{(h_1h_2)}$ Hence the difference in posterior variance can be written as: $\triangle v_{h_1,h_2} = v_{ps} - v_{ps}^{(h_1h_2)}$ Little (1993) observes that if Y and $X \times \mathbb{I}_{\{h_1,h_2\}}$ are independent, then the variance of the estimate can be reduced because the distributions of $Y|X = h_1$ and $Y|X = h_2$ are the same. Then, when we collapse, we reduce the number of parameters to be estimated (μ_{h_1} and μ_{h_2} to $\mu_{h_1h_2}$) and increase the number of observations in the estimation of this parameter. But when $Y|X = h_1$ and $Y|X = h_2$ are associated, we are modelling a mixture of two normals with only one normal distribution and then the replacement of the parameters μ_{h_1} and μ_{h_2} with $\mu_{h_1h_2}$ can affect the global estimation of $\overline{y}$. Little (1993) proposes the following algorithm:

Algorithm 3.1 Collapsing post-strata

1. *Order the post-strata so that neighbors are a priori relatively homogeneous. If they are based on an ordered variable* (*such as age*), *then this step is not needed.*
2. *Collapse the post-stratum pair* $(i, i+1)$ *that maximizes* $E(\triangle v_{i,i+1})$
3. *Proceed sequentially until a reasonable number of post-strata remain or* $E(\triangle v_{i,i+1})$ *becomes noticeably negative.*

Note that in this algorithm the prior knowledge about the population structure is used and determines the final post-stratification.

Remark 3.1. *The assignment of sample size to post-strata is irrelevant if there are enough observations to make inference.*

3.1.3.2 Weighting

Weighting the post-stratification: After selection of strata, we need to weight the observations in each strata. In many applications, survey weights can be the inverse of the selection probabilities but this is not always the case. We can also use regression as an alternative for weighting. Bethlehem and Keller (1987) use regression to provide a method for weighting and to relate the target variables of the survey to auxiliary variables. We now briefly describe this idea. Let $1, ..., N$ be the labels of the elements

of the finite population. We will construct the following matrices: $Y = [y_{ij}]_{N\times q}$ and $X = [x_{ij}]_{N\times p}$ where $Y = (y_1, \ldots, y_N)^T$ are the target variables and $X = (x_1, \ldots, x_n)^T$ are the auxiliary variables. The objective of the sample is to estimate the q-vector of means: $\overline{y} = Y_T \frac{e_N}{N}$ where e_N is the N-vector consisting of ones ($e_N = [1]_{N\times 1}$). Similarly the p-vector of population means for the p auxiliary variables is denoted by $\overline{x} = X^T \frac{e_N}{N}$. A sample from the finite population x can be represented as an $N \times N$-diagonal matrix $D = [d_{ij}]_{N\times N}$. $d_{ii} = 1$ if the i-element of the population is in the sample and $d_{ii} = 0$ otherwise. Then $E[D] = W$ where $W = [w_{ij}]_{N\times N}$ is the $N \times N$ diagonal matrix where w_{ii} are the probabilities of inclusion of the i-element of the population. If the auxiliary variables are correlated with the target variables, an estimator can be constructed as follows:

$$Y = XB + E$$

where E is an $N \times q$-matrix of residuals. Applying the ordinary least squares method results in $B = (X^T X)^{-1} X^T Y$. An estimator for B, based in the sample data, is defined as $\hat{B} = (X^T W^{-1} DX)^{-1} XW^{-1} DY$. The regression estimator is then given by $\hat{\overline{y}}_R = \hat{B}^T Y^T \frac{e_N}{N} = \hat{B}^T \overline{x}$. In the case of Simple Random Sampling (SRS), let y_s be the n-vector of sampled values of the target variable and X_s be the $n \times p$-matrix of auxiliary variables corresponding to sampled elements. The regression estimator then reduces to $\hat{\overline{y}}_R = \hat{\beta}^T \overline{x}_s$, where $\overline{x}_s$ is the p-vector of sample means of the auxiliary variables and $\hat{\beta} = (X_S^T X_S)^{-1} X_s^T y_s$.

We can also include qualitative auxiliary variables. Suppose our auxiliary variable has H categories. For each category there is a dummy variable which takes the value 1 if the particular element belongs to that stratum, otherwise it takes the value 0. The matrix X has dimension $N \times H$ and each row contains exactly one 1. The columns of X sum up to the sub-population totals $N_1, N_2, \ldots, N_H$ where $N_1 + \ldots + N_L = N$. We can retain the notation used in the case of SRS. The columns of X_s will sum up to the random sample totals $n_1, n_2, \ldots, n_H$ in the strata, where $n_1 + n_2 + \ldots + n_H = n$. The vector of population means of the auxiliary variables is equal to $\overline{x} = \frac{(N_1, N_2, .., N_H)^T}{N}$ and the corresponding vector of sample means is equal to $\overline{x}_s = \frac{(n_1, \ldots, n_H)}{n}$. Due to the special structure of the matrix X the matrix $X_S^T X_S$ is a diagonal matrix with diagonal elements equal to $n_1, \ldots, n_H$. We obtain

$$\hat{\beta} = (\overline{y}_s^{(1)}, \ \ldots, \ \overline{y}_s^{(H)})$$

where $\overline{y}_s^h$ is the sample mean of the target variable in stratum h. Finally, we obtain the following regression estimator

$$\hat{\overline{y}}_R = \sum_{h=1}^{H} \frac{N_h \hat{\overline{y}}_s^h}{N} = \hat{\overline{y}}_{ps},$$

where the subscript ps denotes the traditional post-stratification estimator. If there are no observations in one or more strata, then some of the diagonal elements of $X_S^T X_S$ are zero, in which case $X_S^T X_S$ is singular. We can make $X_S^T X_S$ non-singular by collapsing strata. Suppose that we have the auxiliary variables SEX, AGE and REGION. We can consider the following linear model SEX $\times$AGE + REGION where $\times$ denotes the crossing of auxiliary variables. This kind of models are known as incomplete multi-way stratification. Note that in this case we do not consider all the possible interactions between the levels of the categorical variables.

3.1.3.3 Calibration

The method of calibration for estimation of population totals is described by Deville and Särndal (1992). An implicit objective of the method is to use auxiliary information to obtain estimators that are approximately unbiased with smaller variance. As before, consider a finite population labelled by $1, ..., N$. Let y_i be the value of the target variable of the i-unit. Let x_i be the value of the auxiliary variable of the i-unit. Consider s, a random sample from the population. Let $w_i = P(i \in s)$ the probability that the i-unit belongs to the sample. The population total of x, $t_x = \sum_i^N x_i$ is assumed known by the researcher. Deville and Särndal (1992) consider a calibration estimator $\hat{T}_C = \sum_{i \in s} c_i x_i$ where the calibration weights c_i are chosen to minimize a given distance ϕ from the basic weights $c_i^{'} = \frac{1}{w_i}$ subject to $\sum_{i \in s} c_i x_i = t_x$. When a weighted sum of squares is used to measure the distance between the two sets of weights, the estimator obtained through calibration corresponds to a generalized regression estimator.

3.1.4 Nonparametric methods for calibration in finite population sampling

3.1.4.1 Introduction

In linear weighting, (Bethlehem and Keller, 1987) and calibration methods (Deville and Särndal, 1992) it is implicitly assumed that a linear relationship holds between the target and the auxiliary variables. In many cases this assumption is not warranted and can yield poor estimates. In order to obtain more efficient estimators for the population means and the population distributions, in recent years frequentist statisticians began to employ nonparametric models that are more flexible. In the Bayesian approach there are even fewer papers, even though nonparametric Bayesian statistics is a growing area of research. The aim of this section is to provide an overview of the latest nonparametric methods used for the calibration for finite population sampling using auxiliary information, both from the classical and the Bayesian approaches. We also discuss some draft ideas about the use of Bayesian nonparametric models in this important matter for finite population sampling.

3.1.4.2 Frequentist approach

Consider a population $U = \{1, 2, \ldots, N\}$ of N units from which a random sample s of size n is selected according to a specific sampling design with probabilities of inclusion π_i for the i-th element of the population. Let y_i be the value of the target variable Y and x_i the auxiliary variable X for the i-th element.

Deville and Särndal (1992); Bethlehem and Keller (1987) describe the most common methodologies used to include auxiliary variables X in order to make inference about the mean of the target variable Y. Usually, a parametric model is used to represent the relationship $m(\cdot)$ between the auxiliary data and the target variable.

A frequentist approach was suggested by Kuo (1988) for the distribution function; she adopts a nonparametric model-based approach, which is more flexible when modelling the relation between X and Y. Breidt and Opsomer (2000) use the local polynomial regression estimator for the unknown regression function $m(\cdot)$. A more recent model using local polynomial regression estimators can be founded in Rueda and Sánchez-Borrego (2009). These latter authors assume that the population can be described

by

$$y_i = m(x_i) + \epsilon_i, \tag{3.1}$$

where ϵ_i are independent and identically distributed with $E[\epsilon_i] = 0$, and constant variance σ^2. After the sample has been observed, we can estimate $\overline{y}$ with

$$\overline{y} = f\overline{y}_s + (1-f)\overline{y}_{s^c} \tag{3.2}$$

where $\overline{y}_s = \frac{1}{n}\sum_{i\in s} y_i$, $\overline{y}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} y_i$ and f is a real number that helps us to weight the mean of the sample and the mean of the non sampled elements. For example, we can set $f = \sum_{i\in s}\pi_i$. The first term of equation (3.2) is known, and estimating $\overline{y}$ is equivalent to predicting the mean $\overline{y}_{s^c}$ in the non sample data. If x is known for all the population, then a natural way to do the predicting is to use a regression model that treats the unknown values y_i $i \in s^c$ as predicted values of $\hat{y}_i = m(x_i)$ $i \in s^c$. Clearly, when we know the values x_i for the complete population U, an estimator of $\overline{y}$ is

$$\hat{\overline{y}} = f\overline{y}_s + (1-f)\frac{1}{N-n}\sum_{i\in s^c}\hat{y}_i.$$

In practice, we only have the values of the x_i's in the sample and the empirical distribution of X for the complete population U. Chambers et al. (1993) use a fixed bandwidth kernel smoothing to get estimates of the function $m(\cdot)$ evaluated in x_i $i \in s^c$. The kernel smoother is simply a weighted average of all data points.

$$\hat{m}(x) = \frac{\sum_{i\in s} w(x, x_i, h) y_i}{\sum_{i\in s} w(x, x_i, h)}, \tag{3.3}$$

where the weights are specified using a kernel function K and a bandwidth h

$$w(x, x_i, h) = K\left(\frac{x - x_i}{h}\right).$$

The kernel function K is normally a non-negative-valued function, symmetric about zero. The interpretation of the bandwidth depends on the kernel function used, but it is generally true that a larger bandwidth leads to smoother estimates by giving progressively more weight to observations further away from x. As the bandwidth gets smaller, only observations close to x get significant weight and therefore the kernel smooth estimate is largely determined by local observations.

Montanari and Ranalli (2005) develop a different approach using a neural network model for the calibration estimator. In many applications we require more than the mean of the population. We may need, for example, to estimate the quantiles or the density distribution of the target variable Y. (See Rueda et al. (2010) for an extended review of the frequentist methods). Using the same notation as before, the finite population distribution function of the study variable Y, is given by

$$F_Y(t) = \sum_{i \in U} \frac{\triangle(t - y_i)}{N},$$

with

$$\triangle(t - y_i) = \begin{cases} 1 & \text{if } t \geq y_i \\ 0 & \text{otherwise.} \end{cases}$$

A design-based estimator of the distribution function of Y is the Horvitz-Thompson estimator (HT), defined by

$$\hat{F}_{YH}(t) = \sum_{i \in s} d_i \frac{\triangle(t - y_i)}{N},$$

with $d_i = \dfrac{1}{\pi_i}$, which we call the basic design weights. The HT is unbiased but, in general, is not a distribution function since $\lim_{t \to +\infty} \hat{F}_{YH}(t) \neq 1$; also, it does not use the auxiliary information provided by the variable x. We will now describe the model proposed by Rueda et al. (2010). They assume the model described by equation (3.1) and use the same estimator of $m(\cdot)$ defined by equation (3.3). To incorporate the auxiliary information, they propose the calibration estimator based on m

$$\hat{F}_{Ymc}(t) = \sum_{i \in s} \omega_i \frac{\triangle(t - y_i)}{N},$$

where the calibrated weights ω_i are modified from d_i by minimizing the chi-square distance measure

$$\Phi_s = \sum_{i \in s} \frac{(\omega_i - d_i)^2}{d_i q_i},$$

with q_i known positive constants unrelated to d_i, subject to the calibration equations

$$\frac{1}{N} \sum_{i \in s} \omega_i \triangle(t_j - \hat{m}(x_i)) = F_{\hat{m}}(t_j),$$

with t_j for $j = 1, 2 \ldots, P$ arbitrarily chosen points such that $t_1 < t_2 < \ldots < t_P$ and where $F_{\hat{m}}(t_j)$ denotes the finite distribution function or empirical distribution of $\hat{m}(x_i)$ $i \in U$ evaluated at the point t_j. The proposed estimator $\hat{F}_{Ymc}(t)$ is a genuine distribution function. Then the construction of the estimates of the finite population α-quantile of Y denoted by $Q_Y(\alpha)$ is straightforward:

$$\hat{Q}_{Ymc}(\alpha) = \inf\{t : \hat{F}_{Ymc}(t) \geq \alpha\}.$$

In this approach we do not obtain a confidence interval for the α-quantiles estimates.

3.1.4.3 Bayesian approach

Zheng and Little (2003) develop a Bayesian nonparametric method to estimate the total of a finite population with probability-proportional-to-size sampling (PPS), in which the values of π_i are proportional to the values x_i of the size variable X and are usually known for the whole population before s is drawn. Zheng and Little consider the following model

$$y_i = f(\pi_i, \beta) + \epsilon_i$$

with $\epsilon_i \overset{iid}{\sim} N(0, \pi_i^{2k}\sigma^2)$. Here f is a function of π_i that is continuous up to the $(p-1)th$ derivate with unknown parameters β. The exponent k models error heteroscedasticity, and for simplicity it is assumed to be known. The function f is estimated by splines, which are piecewise polynomial functions that are smooth to a certain degree. Zheng and Little (2004) use splines to make inference for the mean from two-stage sample designs. Another approach has been explored by Ciampi et al. (2007) who use Bayesian regression tree models for calibration in the context of micro-array analysis. In this case we have a regression model for each terminal node of the tree, which provides flexibility in the relationship between Y and X since the linear dependence is local. Nelson and Meeden (1998) propose an extension of the Pólya posterior that can be used to estimate population quantiles with auxiliary variables.

3.2 A Bayesian approach to post-stratification

3.2.1 Toy example

In this example we will describe a new approach to post-stratification using random partitions models. Suppose we want to estimate the proportion of people that will vote for the PRI in the Mexico City. We also have a simple random sample of size 100. We have census information which divided the population into three zones: X^1, X^2 and X^3. To obtain better estimates the researchers want to make an adequate post-stratification using this information. They have five ways to make the post-stratification:

$$[\{X^1, X^2, X^3\}][\{X^1, X^2\}, \{X^3\}][\{X^1, X^3\}, \{X^2\}][\{X^1\}, \{X^2, X^3\}][\{X^1\}, \{X^2\}, \{X^3\}]$$

Suppose that, in past elections, the researchers have observed that people from X^1 and people from X^2 have roughly the same behavior. Suppose also the number of people in X^1 is small, so it would be convenient to merge these zones. The researchers want to find the best post-stratification using their knowledge. Let $S_0 = \{1, 2, 3\}$ and $X = \{X^1, X^2, X^3\}$ Let $\rho = \{S_1, \ldots, S_k\}$ be a partition of S_0 into k subsets. We define $X_S = \{X_i, i \in S\}$.

Remark 3.2. *Note that we are inducing a partition over the range of the covariate ZONE.*

Let us exemplify this idea using the following data.

i	y_i	x_i
1	0	1
2	1	1
3	1	2
4	1	2
5	1	2
6	1	2
7	0	2
8	0	3
9	0	3
10	1	3

Table 3.1: Data for the toy example

Suppose that $\rho = \{\{S_1\}\{S_2\}\}$ with $S_1 = \{1, 3\}$ and $S_2 = \{2\}$, then $\{i : x_i \in X_{S_1}\} = \{1, 2, 8, 9, 10\}$ and $\{i : x_i \in X_{S_2}\} = \{3, 4, 5, 7\}$

In a PPM, the partition is over the observations. Müller and Quintana (2010); Müller et al. (2011), use the covariates only to increase the probability that two data points with the same covariates will be in the same cluster (see equation (1.5)). They do not obtain a partition over the range of covariates. Clearly, using the partition over the range of the covariates we obtain a partition over the population. Jordan et al. (2007) presented a partition over the observed covariates of the sample for prediction purposes. Hegarty and Barry (2008) also use a partition over the range of covariates to estimate the relative risk of disease in each area. Holmes et al. (2005) proposed a Bayesian partition model that constructs arbitrarily complex regression and classification surfaces by splitting the covariate space but none of these works study the properties of those models (such as the Bayesian learning process concerning the random partition ρ).

Let $y = (y_1, \ldots, y_n)$ be the sample of size $n = 100$, where

$$y_i = \begin{cases} 1 & \text{if the } i\text{-th individual votes for the PRI,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $x = (x_1, \ldots, x_n)$ where x_i is the zone of the observation y_i. We define $\boldsymbol{y}_S^X = \{y_i : x_i \in X_S\}$. For example, if $S = \{1, 2\}$ then $X_S = \{X^1, X^2\}$ and $\boldsymbol{y}_S^X$ is the set of observations that belong to X^1 or X^2. We will assume that the probability distribution that is associated with $\boldsymbol{y}$ is parameterized by a vector of the form $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. If $\rho = \{S_1, \ldots, S_k\}$, then

$$\boldsymbol{\theta} = \sum_{i=1}^{k} \left(\theta^{S_i}\delta_1(S_i), \ldots, \theta^{S_i}\delta_n(S_i)\right),$$

where θ^S is a common value for the θ_i's, $i \in S$. We denote this particular form of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_\rho$ and note that $\boldsymbol{\theta}$ can also be represented as $(\theta^{S_1}, \ldots, \theta^{S_{|\rho|}}, \rho)$.

Then we have the following hierarchical model:

$$\begin{aligned} \boldsymbol{y}_{S_j}^X | (\theta^{S_1}, \ \ldots, \ \theta^{S_{|\rho|}}, \ \rho) \quad &\sim \quad \prod_{\{i : x_i \in X_{S_j}\}} Bernoulli(y_i | \theta^{S_j}) \\ \theta^{S_j} | \rho \quad &\overset{ind}{\sim} \quad Beta(a^{S_j}, b^{S_j}) \text{ with } S_j \in \rho \\ \rho \quad &\sim \quad RP(S_0), \end{aligned}$$

where θ^{S_j} is a parameter associated to the set or stratum X_{S_j}.

3.2.1.1 Posterior inference

In this section we will calculate the posterior probability $p(\rho = \{S_1, \ldots, S_k\}|y_1, \ldots, y_n)$
Suppose that we have fixed the size of the sample n.

$$\begin{aligned} p(\rho|y_1,\ldots,y_n) &\propto p(y_1,\ldots,y_n|\rho)p(\rho) \\ &\propto p(\rho)p(\boldsymbol{y}^X_{S_1}|\rho) \times \ \ldots \ \times p(\boldsymbol{y}^X_{S_{|\rho|}}|\rho) \\ &\propto p(\rho) \times \int p(\boldsymbol{y}^X_{S_1}|\rho,\theta^{S_1})p(\theta^{S_1}|\rho)d\theta^{S_1} \times \\ &\quad \ldots \times \int p(\boldsymbol{y}^X_{S_{|\rho|}}|\rho,\ \theta_{S_{|\rho|}})p(\theta^{S_{|\rho|}}|\rho)d\theta^{S_{|\rho|}} \end{aligned}$$

We now calculate $\int p(\boldsymbol{y}^X_{S_j}|\rho,\theta^{S_j})p(\theta^{S_j}|\rho)d\theta^{S_j}$.

For any $t \in \mathbb{R}$ we have

$$\begin{aligned} Bernoulli(\boldsymbol{y}^X_{S_j}|t) \times Beta(t|a,b) &= t^{\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i}(1-t)^{n_{S_j}-\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}t^{a-1}(1-t)^{b-1} \\ &= t^{\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+a-1}(1-t)^{n_{S_j}-\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+b-1} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}, \end{aligned}$$

where $n_{S_j} = |S_j|$.

Then we have

$$\begin{aligned} \int_0^1 Bernoulli(\boldsymbol{y}^X_{S_j}|t) \times Beta(t|a,b)dt &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma^1(b)} \times \\ &\quad \int_0^1 t^{\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+a-1}(1-t)^{n_{S_j}-\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+b-1}dt \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &\quad \times \frac{\Gamma(\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+a)\Gamma(n_{S_j}-\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+b)}{\Gamma(a+b+n_{S_j})}. \end{aligned}$$

Finally,

$$\begin{aligned} p(\rho=\{S_1,\ldots,S_k\}|y_1,\ldots,y_n) \propto \prod_{j=1}^{k} \Bigg\{ &\frac{\Gamma(a_{S_j}+b_{S_j})}{\Gamma(a_{S_j})\Gamma(b_{S_j})} \\ &\times \frac{\Gamma(\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+a_{S_j})\Gamma(n_{S_j}-\sum_{y_i\in\boldsymbol{y}^X_{S_j}} y_i+b_{S_j})}{\Gamma(a_{S_j}+b_{S_j}+n_{S_j})} \Bigg\} \\ &\times p(\rho=\{S_1,\ldots,S_k\}). \end{aligned}$$

Remark 3.3.

$$\lim_{n_{S_j}\to 0} \frac{\Gamma(\sum_{y_i\in \boldsymbol{y}^X_{S_j}} y_i + a_{S_j})\Gamma(n_{S_j} - \sum_{y_i\in \boldsymbol{y}^X_{S_j}} y_i + b_{S_j})}{\Gamma(a_{S_j} + b_{S_j} + n_{S_j})} = \Gamma(a_{S_j})\Gamma(b_{S_j})\Gamma(a_{S_j} + b_{S_j})$$

Hence, when there are missing observations in some S_j,

$$p(\boldsymbol{y}^X_{S_1}, \ldots, \boldsymbol{y}^X_{S_{|\rho|}}|\rho) \propto \prod_{\{j|\boldsymbol{y}^X_{S_j}\neq\varnothing\}} p(\boldsymbol{y}^X_{S_j}|\rho)$$

3.2.2 General model

In this section we will introduce a more general model. Let (X, Y) be a random vector. Here X is a finite discrete random variable with $X \in \{X^1, \ldots, X^M\}$. Let $S_0 = \{1, \ldots, M\}, \rho = \{S_1, \ldots, S_k\}$ be a partition of S_0 into k subsets. We define $\boldsymbol{X}_S = \{X^i | i \in S\}$ Let $(x_1, y_1) \ldots (x_n, y_n)$ be a simple random sample of (X, Y). We define $\boldsymbol{y}^X_S = \{y_i | x_i \in X_S\}$. We have the following model

$$\begin{aligned} \boldsymbol{y}^X_{S_j}|(\theta^{S_1},\ \ldots,\ \theta^{S_{|\rho|}},\ \rho) &\sim \prod_{\{i:x_i\in \boldsymbol{X}_{S_j}\}} p(y_i|\theta^{S_j}) \\ \theta^{S_j}|\rho &\overset{ind}{\sim} p(\theta^{S_j}|\boldsymbol{\psi}^{S_j}) \text{ with} S_j \in \rho \\ \rho &\sim RP(S_0), \end{aligned} \tag{3.4}$$

where θ^{S_j} is a parameter associated to the set or stratum $\boldsymbol{X}_{S_j}$ and $\boldsymbol{\psi}^{S_j}$ the respective vector of hyperparameters.

3.2.2.1 Posterior inference

The posterior distribution of the partition is given by

$$\begin{aligned} p(\rho|y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|\rho)p(\rho) \\ &\propto p(\rho)p(\boldsymbol{y}^X_{S_1}|\rho) \times \ldots \times p(\boldsymbol{y}^X_{S_{|\rho|}}|\rho) \\ &\propto p(\rho) \times \int p(\boldsymbol{y}^X_{S_1}|\rho, \theta^{S_1})p(\theta^{S_1}|\rho)d\theta^{S_1} \times \ldots \\ &\quad \times \int p(\boldsymbol{y}^X_{S_{|\rho|}}|\rho,\ \theta_{S_{|\rho|}})p(\theta^{S_{|\rho|}}|\rho)d\theta^{S_{|\rho|}} \end{aligned}$$

For each $j = 1, \ldots, |\rho|$ we have

$$p(\boldsymbol{y}^X_{S_j}|\rho, \theta^{S_j}) = \prod_{\{y_i\in \boldsymbol{y}^X_{S_j}\}} p(y_i|\rho, \theta^{S_j}).$$

When there are missing observations for any S_j,

$$p(\boldsymbol{y}^X_{S_1}, \ldots, \boldsymbol{y}^X_{S_{|\rho|}}|\rho) \propto \prod_{\{j|\boldsymbol{y}^X_{S_j} \neq \varnothing\}} p(\boldsymbol{y}^X_{S_j}|\rho)$$

3.2.3 Bayesian Learning process

3.2.3.1 Introduction

The model defined in the previous section provides an interesting learning process concerning the structure of the population. We will devote this section to study this process. Suppose that each year we make the same study about preferences on the same product. How can we use previous polls to make better estimates? If it seems reasonable to assume that the structure of the population remains the same through time, then we could learn about the post-stratification of the population. This idea can be very useful when there is post-strata with only a few observations. Sometimes it is technically more convenient to describe a partition by a set of cluster membership indicators, $s_i = j$ if $i \in S_j, i = 1, \ldots, n$. Recall that we follow the convention that clusters are labelled in order of appearance, to avoid any inconsistency.

Example 3.4. *Let $X = \{X^1, X^2, X^3, X^4\}$. We consider the vector (s_1, s_2, s_3, s_4) to represent the partition ρ. For example, the vector (1, 1, 2, 3) represents the partition $\{\{1, 2\}, \{3\}, \{4\}\}$, in terms of covariates: $\{\{X^1, X^2\}, \{X^3\}, \{X^4\}\}$*

We simulated data points from the partition (1, 1, 2, 2) and parameters (0.7, 0.7, 0.3, 0.3), then we simulated data from the same partition with parameters (0.6, 0.6, 0.4, 0.4). In Table 3.2, we show the probability of each partition with and without the learning process. The true partition has probability 0.719 with learning process and 0.2026 without it. This property allows us to use previous polls to make better estimates.

	iteration 1	iteration 2		Without learning process
Partition	(0.7,0.7,0.3,0.3)	(0.6,0.6,0.4,0.4)		(0.6,0.6,0.4,0.4)
1,2,3,4	0.08122034	0.01708853		0.0285
1,1,2,3	**0.20027720**	0.08200028		0.0555
1,2,1,3	0.00125464	0.00048465		0.0523
1,2,3,1	0.02916935	0.01866985		0.0867
1,2,2,3	0.00001025	0.00000031		0.0041
1,2,3,2	0.00132556	0.00016207		0.0166
1,2,3,3	**0.19515760**	**0.15001130**		**0.1041**
1,1,1,2	0.00002205	0.00000308		0.0189
1,1,2,1	0.00524628	0.00272366		0.0703
1,2,1,1	0.00502333	0.00900247		0.2427
1,2,2,2	0.00001885	0.00000184		0.0132
1,1,2,2	**0.48122950**	**0.71983760**		**0.2026**
1,2,1,2	0.00002048	0.00000460		0.0304
1,2,2,1	0.00000368	0.00000034		0.0125
1,1,1,1	0.00002089	0.00000948		0.0615

Table 3.2: Simulation of the learning process

Example 3.5. *We simulated data from partition (1, 1, 2, 2) using parameters chosen at random at each iteration. In Table 3.3 we show the parameters used at each iteration.*

	θ_1	θ_2
1	0.725	0.947
2	0.658	0.356
3	0.583	0.423
4	0.985	0.572
5	0.799	0.967
6	0.767	0.591
7	0.86	0.655
8	0.84	0.784
9	0.636	0.066
10	0.992	0.346
11	0.898	0.847
12	0.622	0.405
13	0.387	0.426
14	0.387	0.337
15	0.099	0.091
16	0.251	0.288
17	0.292	0.111
18	0.95	0.422
19	0.727	0.867
20	0	0.663

Table 3.3: Values of θ_1 and θ_2

In Figure 3.2 we can see the probability of the partition (1, 1, 2, 2).

In the first and third iteration we can see that the parameters θ^{s_1} and θ^{s_2} are close. The probability is near 0.2. In the 22th and 25th iterations, the parameters θ^{s_1} and θ^{s_2} are close too, but because of the learning process, the probability of the true partition is near 1. The figure suggests consistency.

To assess the performance of the model, in this toy example we construct a **weighted distance** between the true partition ρ and the partition $\hat{\rho}$. This distance is based on the loss function of Binder (1978), which is given by

$$\sum_{\hat{\rho}} W(\rho, \hat{\rho}) \left(\delta_{\rho}(\{\hat{\rho}\}) - P(\hat{\rho})\right)^2$$

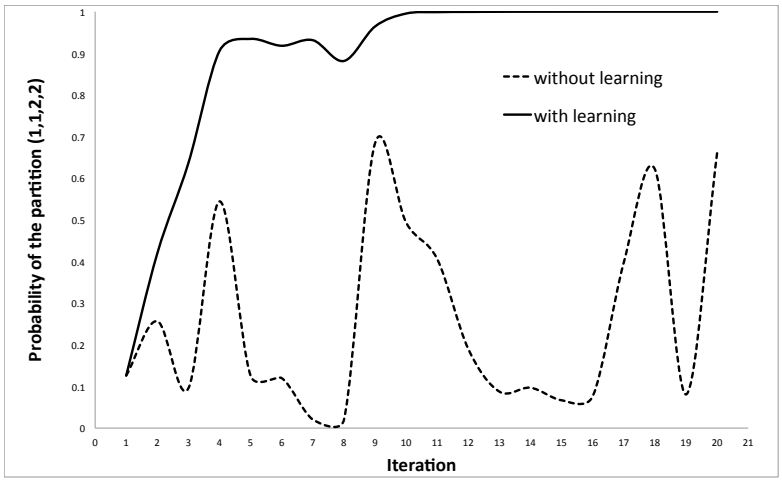


Figure 3.2: Probability of the partition (1,1,2,2) with and without the learning process.

with

$$W(\rho, \hat{\rho}) = \sum_{i<j} \{U(i,j) + V(i,j)\},$$

where $U(i,j) = 1$ if and only if i and j are in the same cluster under partition ρ but in different clusters under $\hat{\rho}$; similarly, $V(i,j) = 1$ if and only if i and j are together in $\hat{\rho}$ but separated in ρ. Figure 3.3 shows the results obtained for different number of simulations for $P(X_i = 1) = 0.25 \quad P(X_i = 2) = 0.25 \quad P(X_i = 3) = 0.1 \quad P(X_i = 4) = 0.4$ $Y_i|\{X_i \in \{1,2\}\} \sim Bernoulli(0.8) \;\; Y_i|\{X_i \in \{3,4\}\} \sim Bernoulli(0.2)$

.

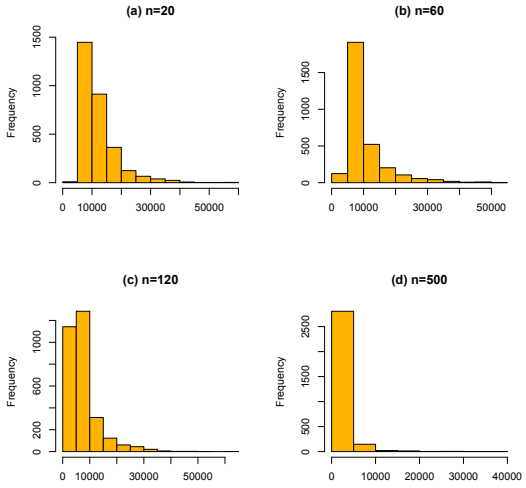


Figure 3.3: Using weighted distance to assess the performance of Example 3.3.

3.2.3.2 Product partition parameters correlated in time

A common practice in finite population sampling is to use previous polls to estimate the variance of the population and calculate the number of observation given a desired precision. For this purpose, we define a new model in which the parameters associated for each stratum can be dependent of previous polls. The model is given by

$$\begin{aligned} \boldsymbol{y}_{S_j}^X|(\theta_{s_1}^t,\ \ldots,\ \theta_{s_{|\rho|}}^t,\ \rho) &\sim \prod_{\{i,x_i\in X_j^*\}} Bernoulli\ (y_i|\theta_{s_j}^t) \\ \theta_{s_j}^t|\rho &\overset{ind}{\sim} Beta(a_{S_j}^t, b_{s_j}^t) \text{with} S_j \in \rho \\ a_{s_j}^t, b_{s_j}^t &\sim p(\phi_{S_j}) \\ \rho &\sim RP(S_0) \end{aligned} \tag{3.5}$$

where $\theta_{s_j}^t$ is the parameter associated to stratum X_j^* at time t. Let $E(\phi) = \phi_0$; if $var(\phi) \to 0$ then, $p(\phi) \to \delta_{\phi_0} \Rightarrow$

$$p(\theta^1, \ldots, \theta^T) = \int \prod_{i=1}^{T} p(\theta^i|\phi)p(\phi)d\phi \to \prod_{i=1}^{T} p(\theta^i|\phi_0). \tag{3.6}$$

When $var(\phi) \to \infty$, $\theta^1, \ldots, \theta^T$ are highly correlated.

For a graphical representation of this model see Figure 3.4.

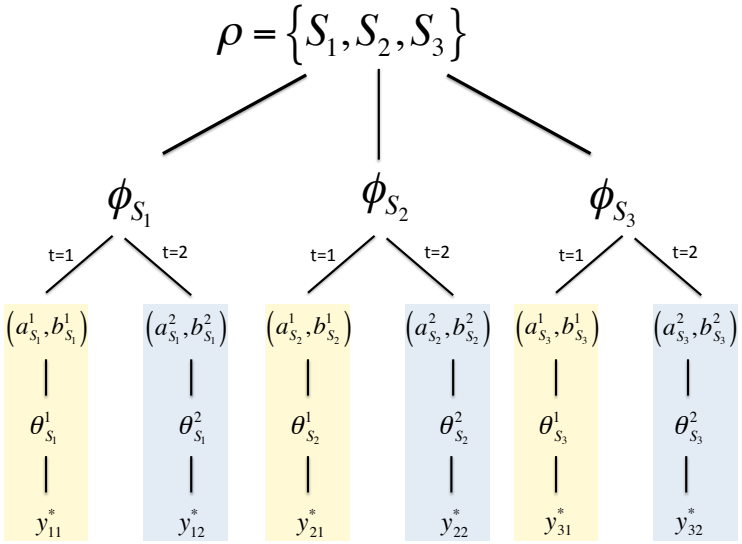


Figure 3.4: Graphical representation of PPM parameters correlated in time.

3.2.4 Loss function

We define the loss function

$$l\left(\rho, \hat{\rho}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\hat{\rho}}\right) = \gamma \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\hat{\rho}} \right\|^2 + (1-\gamma)\,|\hat{\rho}.| \tag{3.7}$$

In Appendix A, we state and provide a more detailed proof clarifying all the elements of the sequential decision problem of the theorem of Quintana and Iglesias (2003). This will allow us to find the partition which minimizes the expected loss.

3.3 Discussion

In this chapter, we proposed a Bayesian post-stratification approach where we model the post-stratification as a random partition in order to include our prior knowledge about the structure of the population. We defined a hierarchical Bayesian model tailored for such purpose. We also provided a toy example to illustrate this kind of models. Note that, if we have a set of five discrete covariates, each covariate with 3 levels, then we will have $3^5 = 243$ basic strata. Using the Bell number in equation (1.1), we have $B_{243} > 10^{275}$. This is the number of possible ways that we can stratify our population

using the five covariates. Because of this, we were not able to carry out the simulations for more realistic problems.

Chapter 4

Variable Selection

Human existence is based upon two pillars: Compassion and knowledge. Compassion without knowledge is ineffective; knowledge without compassion is inhuman

Weisskopf Victor (Physicist)

Variable selection is an important issue in regression analysis. In Bayesian statistics, Bayes Factors have played a major role when dealing with this problem; in fact, despite the importance of decision theory within the Bayesian framework, there are few proposals that use a utility function to address this problem. Inspired by product partition models, we propose a probability measure over all possible models and use a loss function suitable for prediction purposes. We point out some of the differences from other recent theoretic-decision approaches such as that of Hahn and Carvalho (2015), who use a similar loss function but different probability and decision settings, or Barbieri and Berger (2002) who studied conditions to minimize the squared error loss in orthogonal designs or nested models.

4.1 Introduction

In the framework of regression analysis, we usually want to explain the data or predict future observations in the simplest way, so redundant predictors must be removed. Suppose that we have n observations with k covariates and a response variable $\boldsymbol{Y}$.

Set $p = k + 1$, $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{X} = [\boldsymbol{x_{ij}}]$ is a $n \times p$ matrix with $i = 1, ..., n$, $j = 0, ..., k$ and $x_{i0} = 1$ for all i. Let $S_0 = \{1, ..., k\}$ and $S \subseteq S_0$. We define the

model $M_S : \boldsymbol{Y} = \boldsymbol{X}_S\boldsymbol{\beta}_S + \boldsymbol{\varepsilon}_S$ with $\boldsymbol{X}_S = [x_{ij}]_{j\in\{0\}\cup S}$ a $n \times n_S$ matrix $n_S = |S| + 1$, $\boldsymbol{\varepsilon}_S \sim N_n\left(0, \sigma_S^2\boldsymbol{I}_n\right)$ where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix and $|S|$ is the cardinality of the set S. $\boldsymbol{\beta}_S = \left(\beta_0^S, ..., \beta_{|S|}^S\right)$. In other words,

$$M_S : \boldsymbol{Y} \sim N_n\left(\boldsymbol{y}|\boldsymbol{X}_S\boldsymbol{\beta}_S, \sigma_S^2\boldsymbol{I}_n\right)$$

Note that we have $T = 2^k$ models.

Example 4.1. *Suppose that we have 3 observations and 5 covariates, then*

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ 1 & x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \end{pmatrix}.$$

If $S = \{2, 3, 5\}$ *then*

$$\boldsymbol{X}_S = \begin{pmatrix} 1 & x_{12} & x_{13} & x_{15} \\ 1 & x_{22} & x_{23} & x_{25} \\ 1 & x_{32} & x_{33} & x_{35} \end{pmatrix}$$

with $\boldsymbol{\beta}_S = \left(\beta_0^S, \beta_2^S, \beta_3^S, \beta_5^S\right)$

4.2 A new theoretic decision approach for variable selection

In this section we set up the theoretic decision framework for our problem. In variable selection, we typically first choose a model among the T possibilities, then we make predictions using the chosen model. This decision process can be modeled as a sequential decision problem. In the first stage, the space of actions is $\mathcal{A}_1 = \{\text{select } M_{\hat{S}_i}, 1 \leq i \leq T\}$ and the space of uncertain events is $\mathcal{E}_1 = \{M_{S_i}, 1 \leq i \leq T\}$ (see Figure 4.1). We are ready to define the loss function associated with choosing $M_{\hat{S}}$ when the *true model* is M_S. Let

$$l(M_{\hat{S}}, M_S) = A\psi(X_{\hat{S}}) + B(|\hat{S}| - |S|), \tag{4.1}$$

where $\psi\left(\boldsymbol{X}_{\hat{S}}\right)$ is a measure of the collinearity of the covariates, $|S|$ is the number of covariates of the model M_S and $A, B \geq 0$. The constants A and B help us standardize the measures $\psi(X_{\hat{S}})$ and $|\hat{S}| - |S|$, making them commensurable. Recall that our main purpose is prediction with the minimum number of covariates, which should ideally not be correlated. The term $|\hat{S}| - |S|$ promotes fewer covariates. In fact, when the estimated

model $M_{\hat{S}}$ has less covariates than the *true model* M_S, the loss function rewards it; otherwise, it penalizes it. For instance, suppose X_1, X_2, X_3 and X_4 are independent variables such as $\boldsymbol{Y} = X_1 + X_2 + X_3 + X_4 + \boldsymbol{\epsilon}$, $X_5 = X_1 + X_2$, $X_6 = X_3 + X_4$ and $X_7 = X_1 + X_2 + X_3 + X_4$. Defining $S = \{1, 2, 3, 4\}$, $\hat{S}_1 = \{5, 6\}$ and $\hat{S}_2 = \{7\}$ and $A = 0$, we obtain

$$l(M_{\hat{S}_1}, M_S) = -2 > -3 = l(M_{\hat{S}_2}, M_S),$$

which reflects our preference to predict with the fewest number of covariates.

The setting for the second stage of our decision problem is straightforward: Given that we have chosen $M_{\hat{S}}$, the space of actions is $\mathcal{A}_2 = \{\text{select } \hat{\boldsymbol{y}}_{M_{\hat{S}}} \in \mathbb{R}^n | M_{\hat{S}}\}$. The space of uncertain events is $\mathcal{E}_2 = \{\boldsymbol{y}_{M_S} \in \mathbb{R}^n | \textit{true model} \text{ is } M_S\}$. A natural loss function of choosing $\hat{\boldsymbol{y}}_{M_{\hat{S}}}$ when the true value is $\boldsymbol{y}_{M_S}$ could be

$$l'(\hat{\boldsymbol{y}}_{M_{\hat{S}}}, \boldsymbol{y}_{M_S}) = C\|\hat{\boldsymbol{y}}_{M_{\hat{S}}} - \boldsymbol{y}_{M_S}\|^2 \tag{4.2}$$

where $\boldsymbol{y}_{M_S}$ is the *true value* of $\boldsymbol{y}$ predicted by the *true model* M_S and $C \geq 0$.

Finally, we introduce the loss function composed by l and l', which can be seen as the entire loss of the sequential decision problem previously described. For the sequential decision problem depicted in Figure 4.1, the loss function is given by

$$\begin{aligned} L(M_{\hat{S}}, M_S; \hat{\boldsymbol{y}}_{M_{\hat{S}}}, \boldsymbol{y}_{M_S}) &= l'(\hat{\boldsymbol{y}}_{M_{\hat{S}}}, \boldsymbol{y}_{M_S}) + l(M_{\hat{S}}, M_S) \\ &= C\|\boldsymbol{y}_{M_S} - \hat{\boldsymbol{y}}_{M_{\hat{S}}}\|^2 + A\psi(\boldsymbol{X}_{\hat{S}}) \\ &\quad + B(|\hat{S}| - |S|) \end{aligned} \tag{4.3}$$

with $A, B, C \geq 0$

Note that C, A, B describe the trade-off between precision, collinearity and model complexity. Note also that the terms in equation (4.3), $A\psi(\boldsymbol{X}_{\hat{S}}) + B(|\hat{S}| - |S|)$, can be seen as the loss incurred when for choosing the *wrong model*. Since our main goal is prediction with minimum number of covariates and that the covariates should be uncorrelated, these penalizations seem reasonable. Given that we have chosen model $\hat{M}_S$, the term $\|\boldsymbol{y}_{M_S} - \hat{\boldsymbol{y}}_{M_{\hat{S}}}\|^2$ penalizes for the difference of the *true value* $\boldsymbol{y}_{M_S}$ and the estimated value $\hat{\boldsymbol{y}}_{\hat{M}_S}$ using model $\hat{M}_S$.

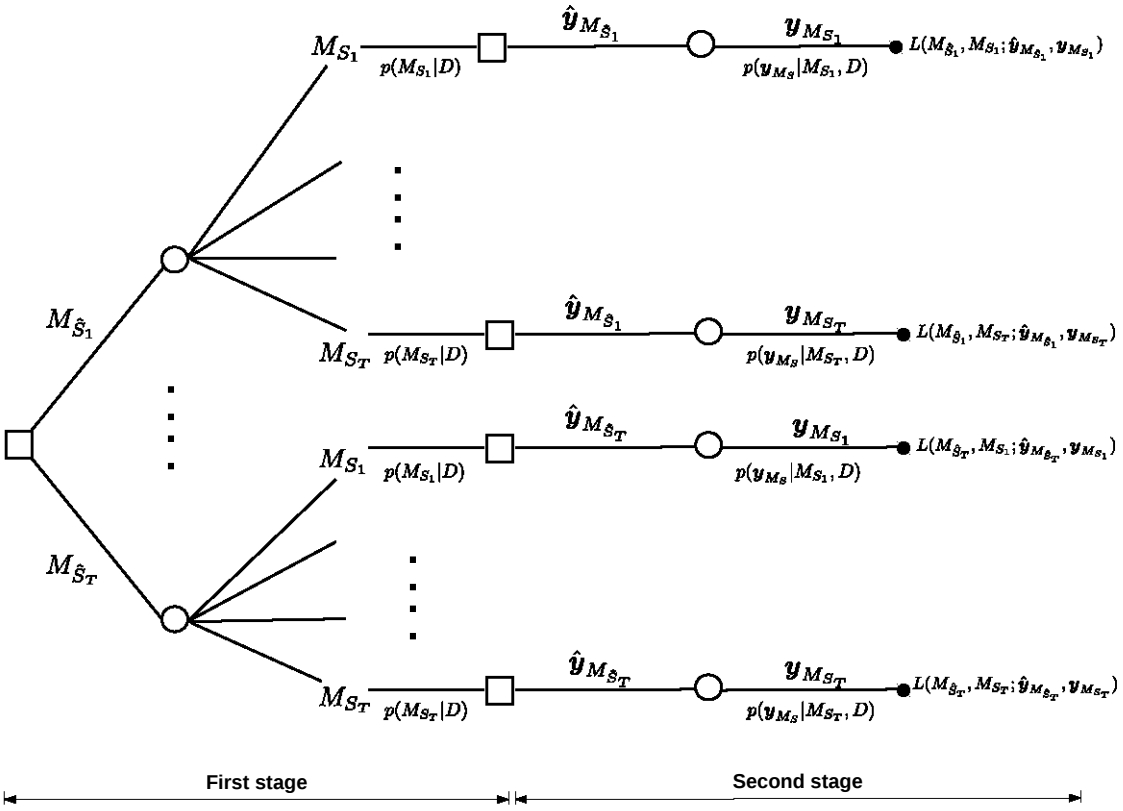


Figure 4.1: Sequential decision problem for variable selection

Concerning the probability measure on the space of uncertain events, we define the following hierarchical model:

$$\begin{aligned} p(M_S) &\propto \prod_{i \in S} c_i \\ \boldsymbol{Y}|M_S, \boldsymbol{\beta}_S, \sigma_S^2 &\sim N_n\left(\boldsymbol{y}|\boldsymbol{X}_S\boldsymbol{\beta}_S, \sigma_S^2\boldsymbol{I}_n\right) \\ \boldsymbol{\beta}_S, \sigma_S^2 &\sim \boldsymbol{\pi}(\boldsymbol{\beta}_S, \sigma_S^2), \end{aligned}$$

where $c_i \geq 0$ quantifies the prior belief that the i-th covariate should be included in the model and $\boldsymbol{\pi}(\boldsymbol{\beta}_S, \sigma_S^2)$ is the prior probability function of the parameters of the model M_S. We now calculate $p\left(M_S|D\right)$ where D is the observed data. It is given by

$$p\left(M_S|D\right) \propto p\left(D|M_S\right) p(M_S)$$

$$\propto \prod_{i \in S} c_i \int N_n \left(\boldsymbol{y} | \boldsymbol{X}_S \boldsymbol{\beta}_S, \sigma_S^2 \boldsymbol{I}_n \right) \boldsymbol{\pi}(\boldsymbol{\beta}_S, \sigma_S^2) d\boldsymbol{\beta}_S d\sigma_S^2$$

We are ready for the following proposition, which will allow us to find the model $M_{\hat{S}}$ that minimizes the loss function defined previously. We will prove this result using the ideas exposed in Quintana and Iglesias (2003).

Proposition 4.2. *Let $\hat{\boldsymbol{y}}_B = E(\boldsymbol{y}|D)$ and $\overline{\boldsymbol{y}}_{M_{\hat{S}}} = E\left(\boldsymbol{y}|M_{\hat{S}}, D\right)$. Then the expected loss minimization using the loss function of equation (4.3) leads to the choice of $M_{\hat{S}}^*$ that minimizes*

$$SC(M_{\hat{S}}) \quad = \quad \gamma_1 \|\hat{\boldsymbol{y}}_B - \overline{\boldsymbol{y}}_{M_{\hat{S}}}\|^2 + \gamma_2 \psi(\boldsymbol{X}_{\hat{S}}) + (1 - \gamma_1 - \gamma_2)|\hat{S}| \tag{4.4}$$

Proof. We begin by solving the optimization problem at the second stage (see Figure 4.1):

$$\begin{aligned}
\bar{l}_2 \left(\hat{\boldsymbol{y}}_{M_{\hat{S}}} | M_S \right) &= \int L(M_{\hat{S}}, M_S; \hat{\boldsymbol{y}}_{M_{\hat{S}}}, \boldsymbol{y}_{M_S}) p(\boldsymbol{y}_{M_S} | M_S, D) d\boldsymbol{y}_{M_S} \\
&= \int \left[\gamma_1 \|\boldsymbol{y}_{M_S} - \hat{\boldsymbol{y}}_{M_{\hat{S}}}\|^2 + \gamma_2 \psi(\boldsymbol{X}_{\hat{S}}) + (1 - \gamma_1 - \gamma_2)(|\hat{S}| - |S|) \right] p(\boldsymbol{y}_{M_S} | M_S, D) d\boldsymbol{y}_{M_S} \\
&= \gamma_1 \int \|\boldsymbol{y}_{M_S} - \hat{\boldsymbol{y}}_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}_{M_S} | M_S, D) d\boldsymbol{y}_{M_S} + \gamma_2 \psi(\boldsymbol{X}_{\hat{S}}) + (1 - \gamma_1 - \gamma_2)(|\hat{S}| - |S|)
\end{aligned}$$

Now let $\hat{\boldsymbol{y}}^*_{M_{\hat{S}}} = \arg\min_{\hat{\boldsymbol{y}}_{M_{\hat{S}}}} \left\{ \bar{l}_2 \left(\hat{\boldsymbol{y}}_{M_{\hat{S}}} | M_S \right) \right\}$; then the loss of choosing the model $M_{\hat{S}}$ when M_S is the real state of the world would be $\bar{l}_2 \left(\hat{\boldsymbol{y}}^*_{M_{\hat{S}}} | M_S \right)$. Since $\hat{\boldsymbol{y}}^*_{M_{\hat{S}}}$ is the Bayes estimate under quadratic loss, we obtain:

$$\hat{\boldsymbol{y}}^*_{M_{\hat{S}}} = E\left(\boldsymbol{y}_{M_{\hat{S}}} | M_{\hat{S}}, D \right) = E\left(\boldsymbol{y} | M_{\hat{S}}, D \right) = \overline{\boldsymbol{y}}_{M_{\hat{S}}}$$

We will now solve the first stage (see Figure 4.2).

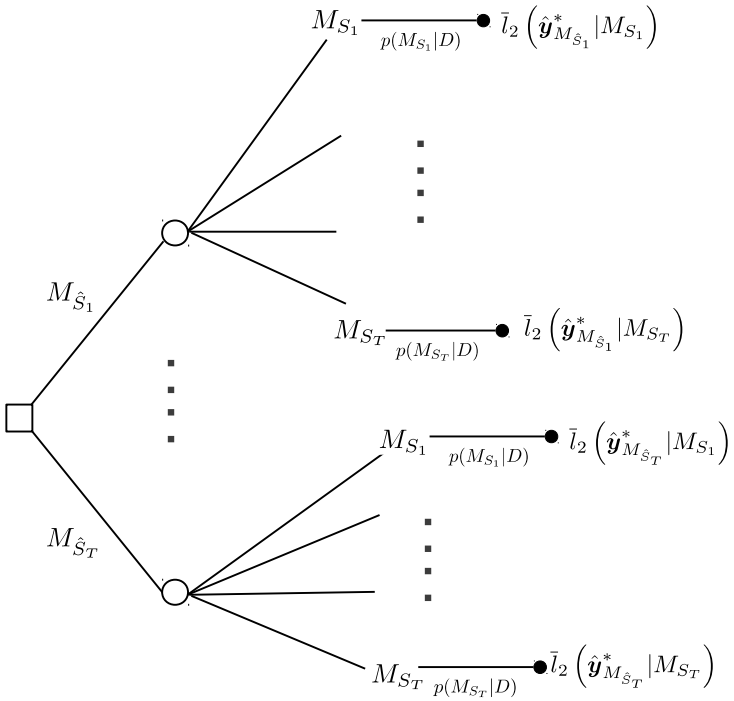


Figure 4.2: First stage of the decision problem

The expected loss is given by

$$\begin{aligned}
\bar{l}\left(M_{\hat{S}}\right) &= \sum_{S\subset S_0} \bar{l}_2\left(\hat{\boldsymbol{y}}^*_{M_{\hat{S}}}|M_S\right) p\left(M_S|D\right) \\
&= \sum_{S\subset S_0} \left\{\gamma_1 \int \|\boldsymbol{y}_{M_S} - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}_{M_S}|M_S, D) d\boldsymbol{y}_{M_S} + (1-\gamma_1-\gamma_2)|S|\right\} p\left(M_S|D\right) \\
&\quad +\gamma_2\psi(\boldsymbol{X}_{\hat{S}}) + (1-\gamma_1-\gamma_2)|\hat{S}| \\
&= \gamma_1 \int \|\boldsymbol{y}_{M_S} - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 \left\{\sum_{S\subset S_0} p\left(M_S|D\right) p(\boldsymbol{y}_{M_S}|M_S, D)\right\} d\boldsymbol{y}_{M_S} + (1-\gamma_1-\gamma_2)E(|S|) \\
&\quad +\gamma_2\psi(\boldsymbol{X}_{\hat{S}}) + (1-\gamma_1-\gamma_2)|\hat{S}| \\
&= \gamma_1 \int \|\boldsymbol{y} - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}|D) d\boldsymbol{y} + (1-\gamma_1-\gamma_2)E(|S|) + \gamma_2\psi(\boldsymbol{X}_{\hat{S}}) + (1-\gamma_1-\gamma_2)|\hat{S}|.
\end{aligned}$$

We now calculate $\int \|\boldsymbol{y} - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}|D) d\boldsymbol{y}$

$$\begin{aligned}\int \|\boldsymbol{y} - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}|D)d\boldsymbol{y} &= \int \|\boldsymbol{y} - \hat{\boldsymbol{y}}_B + \hat{\boldsymbol{y}}_B - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}|D)d\boldsymbol{y} \\ &= \int \|\boldsymbol{y} - \hat{\boldsymbol{y}}_B\|^2 p(\boldsymbol{y}|D)d\boldsymbol{y} \\ &\quad +2\int (\boldsymbol{y} - \hat{\boldsymbol{y}}_B) \cdot \left(\hat{\boldsymbol{y}}_B - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\right) p(\boldsymbol{y}|D)d\boldsymbol{y} \\ &\quad + \int \|\hat{\boldsymbol{y}}_B - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2 p(\boldsymbol{y}|D)d\boldsymbol{y} \\ &= V(\boldsymbol{y}|\boldsymbol{D}) + 0 + \|\hat{\boldsymbol{y}}_B - \hat{\boldsymbol{y}}^*_{M_{\hat{S}}}\|^2.\end{aligned}$$

Note that the first term in the last expression and $E(|S|)$ do not depend on the model $M_{\hat{S}}$ and can be regarded as constants.

Hence, finding the optimum of $\left(M^*_{\hat{S}}, \hat{y}^*_{M_S}\right)$ which minimizes the expected loss is equivalent to minimizing

$$SC(M_{\hat{S}}) = \gamma_1 \|\hat{\boldsymbol{y}}_B - \overline{\boldsymbol{y}}_{M_{\hat{S}}}\|^2 + \gamma_2 \psi(\boldsymbol{X}_{\hat{S}}) + (1 - \gamma_1 - \gamma_2)|\hat{S}|.$$

□

This solution can be interpreted as the best model with the loss function described by equation (4.3), whose predictions are the closest to the average prediction based on all the models. Moreover, Proposition 4.2 suggests a procedure based on distances to find the optimal $M^*_{\hat{S}}$. Unfortunately, an exhaustive search on the space of all possible partitions is unfeasible. However, we can adopt different heuristic algorithms depending on each application. These strategies may not give us the optimal solution but can lead us to a reasonable suboptimal one in a realistic amount of time.

4.2.1 Consistency

Consistency is a desirable property. When the size of n increases, our predictions should be near the predictions under the true model.

Proposition 4.3. *If $\boldsymbol{\pi}(\boldsymbol{\beta}_S, \sigma^2_S)$ is proper and $M^*_{\hat{S}}$ minimizes*

$$SC(M_{\hat{S}}) = \|\hat{\boldsymbol{y}}_B - \overline{\boldsymbol{y}}_{M_{\hat{S}}}\|^2$$

then

$$\lim_{n\to\infty} \boldsymbol{y}_{M^*_{\hat{S}}} = \boldsymbol{y}_{M_S}$$

We provide a sketch of the proof.

Proof. Since $\boldsymbol{\pi}(\boldsymbol{\beta}_S, \sigma_S^2)$ is proper, $\lim_{n\to\infty} P(M_S|D) = 1$, hence $\lim_{n\to\infty} \hat{\boldsymbol{y}}_B = \boldsymbol{y}_{M_S}$ also, $\lim_{n\to\infty} \overline{\boldsymbol{y}}_{M_S} = \boldsymbol{y}_{M_S}$, hence $\lim_{n\to\infty} \|\hat{\boldsymbol{y}}_B - \overline{\boldsymbol{y}}_{M_S}\|^2 = 0$. Since $y^*_{M_{\hat{S}}}$ minimizes $SC(M_{\hat{S}})$ then $\lim_{n\to\infty} y^*_{M_{\hat{S}}} = \hat{\boldsymbol{y}}_B$, then $\lim_{n\to\infty} \boldsymbol{y}_{M^*_{\hat{S}}} = \boldsymbol{y}_{M_S}$ □

Note that the optimal model that minimizes SC may not be unique.

4.3 Other loss functions

In this section we discuss other loss functions that explicitly include a penalization for not choosing the *true model*. One such loss function is the following.

$$L(M_{\hat{S}}, M_S) = \|G(\hat{S}) - G(S)\|^2 \tag{4.5}$$

where $G(S)$ is a p vector such that its i-th entry is equal to $\mathbb{I}_{i\in S}$. Suppose that the true model was generated by covariates X_1, X_2, X_3, and $X_1 = X_4$, $X_2 = X_5$ and $X_3 = X_6$. Defining $S = \{1,2,3\}$, $\hat{S} = \{4,5,6\}$ we obtain $L(M_{\hat{S}}, M_S) = \sqrt{6}$ although $M_{\hat{S}} = M_S$. Moreover, suppose that the true model is generated by covariates X_1, X_2 and X_3; suppose also that X_1 is highly correlated with X_3. Defining $S = \{1,2,3\}$ and $\hat{S} = \{1,2\}$ we obtain $L(M_{\hat{S}}, M_S) = 1$ even though we obtain the same prediction with a smaller number of covariates than with the true model.

Proposition 4.4. *The model which includes all variables with $p(X_i|D) > \frac{1}{2}$ is optimal under the loss function L of equation (4.5). This model is called the median probability model (MPM) by Barbieri and Berger (2002).*

Proof. For the sake of simplicity, we will denote by $p(X_i|D)$ the probability that X_i belongs to Model M_S. Note that

$$\begin{aligned} E\left(L(M_{\hat{S}}, M_S)\right) &= \sum_{i=1}^{p} E\left(\mathbb{I}_{i\in\hat{S}} - 2\mathbb{I}_{i\in\hat{S}}\mathbb{I}_{i\in S} + \mathbb{I}_{i\in S}\right) \\ &= \sum_{i=1}^{p} \left[\mathbb{I}_{i\in\hat{S}} - 2\mathbb{I}_{i\in\hat{S}} p(X_i|D) + p(X_i|D)\right]. \end{aligned}$$

Since the p terms of this sum do not affect each other, we only need to maximize each one in order to maximize $E\left(L(M_{\hat{S}}, M_S\right)$. If $p(X_i|D) \leq \frac{1}{2}$, we maximize $\mathbb{I}_{i\in\hat{S}} - 2\mathbb{I}_{i\in\hat{S}} p(X_i|D) + p(X_i|D)$ when $\mathbb{I}_{i\in\hat{S}} = 0$. In the case $p(X_i|D) > \frac{1}{2}$, the optimum value is obtained when $\mathbb{I}_{i\in\hat{S}} = 1$. □

The Kullback-Leibler (KL) divergence is a widely used criterion to measure the information loss when the distribution function Q is used to approximate P. Suppose that q and p are two densities with respect to Lebesgue measure, then the KL divergence between p and q is defined by

$$KL(p\|q) = \int \ln\left(\frac{p(x)}{q(x)}\right) p(x)dx$$

For the variable selection problem, Laud and Ibrahim (1995) used the following criterion based on the KL to choose the covariates that should be included in the model

$$K_{M_{\hat{S}}} = KL\left(M_{S_0}\|M_{\hat{S}}\right) + KL\left(M_{\hat{S}}\|M_{S_0}\right). \tag{4.6}$$

They compare all the models with the full model (M_{S_0}) and select the one that minimizes $K_{M_{\hat{S}}}$. This approach does not include the probability of each model and does not intend to minimize the expected value of $K_{M_{\hat{S}}}$. Based on this method, we can consider the following loss function.

Definition 4.5. *Let*

$$L\left(M_{\hat{S}}, M_S\right) = A \cdot KL\left(M_S\|M_{\hat{S}}\right) + B \cdot KL\left(M_{\hat{S}}\|M_S\right) + C|\hat{S}| \tag{4.7}$$

where $A, B, C \geq 0$.

In our definition, we compare the estimated model $M_{\hat{S}}$ with the *true model* M_S instead of the full model M_{S_0}. Using the sequential decision settings discussed before, we can calculate the optimum $M^*_{\hat{S}}$ which minimizes the expected loss $L\left(M_{\hat{S}}, M_S\right)$. This procedure is general and can be applied, for example, to generalized linear models, although it presents technical challenges such as the estimation of $\boldsymbol{\beta}_{\hat{S}}$ given $M_{\hat{S}}$.

4.4 Other theoretic decision approaches

Hahn and Carvalho (2015) provide an explicit Bayesian decision-theoretic perspective which we will briefly review here. Let $\bar{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ denote the density of the posterior of the parameters $\boldsymbol{\beta}$ and σ^2 by $\pi(\boldsymbol{\beta}, \sigma^2|D)$ and $\lambda \geq 0$. Let

$$f(\bar{Y}) = \int f(\bar{Y}|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2|D)d(\boldsymbol{\beta}, \sigma^2) \tag{4.8}$$

then the loss function

$$L(\hat{M}_S, \bar{Y}) = \lambda|\hat{S}| + n^{-1}\|\hat{y}_{\hat{M}_S} - \bar{Y}\|^2 \tag{4.9}$$

leads to minimizing

$$L(\hat{M}_S) = \lambda|\hat{S}| + n^{-1}\|\hat{y}_{\hat{M}_S} - \boldsymbol{X}\bar{\boldsymbol{\beta}}\|^2, \tag{4.10}$$

where $\bar{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|D)$.

They solve the optimization problem using combinatorial programming.

Our approach is different from the work of Hahn and Carvalho (2015); they do not use a probability measure over the models and so the loss function they employ compares only the predicted value of the full model with the prediction of the estimated model. In our case, we compare the predicted value of the estimated model with the true value using the probability induced by the proposed hierarchical model. In some sense, their approach can be seen as a particular case of our model when the probability of the full model is close to one ($p(M_{S_0}|D) \approx 1$). Since we have a probability measure over all possible models, we can calculate the marginal probability of each covariate ($p(X_i|D) = \sum_{\{M_S|i\in S\}} p(M_S|D)$), which can guide us about the relative importance of each covariate. This is unfeasible in their approach. From a practical point of view, there are also important differences. When the number of covariates is greater than the number of observations ($p > n$), their approach is very restrictive, requiring very informative distributions for the existence of $\pi(\boldsymbol{\beta}, \sigma^2|D)$ under the full model, as opposed to our proposal which does not assume any condition on the prior distribution of $(\boldsymbol{\beta}^S, \sigma_S^2)$ with $|S| \leq p-1$ (which can lead to a different model estimate).

Under the squared loss function, in orthogonal design or sequence of nested models, Barbieri and Berger (2002) have shown that the MPM is optimal. As pointed out by Hahn and Carvalho (2015), the MPM is defined via marginal quantities which can be misleading when strong dependence among predictors is present. Our approach overcomes this difficulties because it is general and does not impose any restriction on the models. We also include a term which penalizes complexity and multicollinearity, allowing us to find the set of uncorrelated variables which performs the best.

4.5 Computation of to $\hat{\boldsymbol{y}}_B$

In order to obtain $\hat{\boldsymbol{y}}_B$, we can use Gibbs Sampling on the space of the 2^k possible models. Let U_i be an auxiliary random quantity that reflects whether or not the variable X_i belongs to the model $M_{\hat{S}}$:

$$U_i = \begin{cases} 1 & \text{if} \quad i \in M_{\hat{S}}, \\ 0 & \text{if} \quad i \notin M_{\hat{S}}, \end{cases}$$

for $i = 1, ..., p$.

Each model $\left(U_1^s, ..., U_p^s\right)$, $s \geq 1$, is generated by using Gibbs sampling.

Starting from an initial value $\left(U_1^0, ..., U_p^0\right)$, the r-th element at step s, U_r^s, is generated from the conditional distribution

$$U_r \mid U_1^s, ..., U_{r-1}^s, U_{r+1}^{s-1}, ..., U_p^{s-1}; D$$

for $r = 1, ..., p$.

To avoid unnecessary calculations, it is enough to consider the following ratio

$$R_r = \frac{P\left(U_r = 1 \mid V_r^s; D\right)}{P\left(U_r = 0 \mid V_r^s; D\right)}$$

for $r = 1, ..., n-1$, in which $V_r^s = \left\{U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1}\right\}$,

$$P\left(U_r = 1 \mid V_r^s; D\right) = \frac{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 1, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid D\right)}{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid D\right)}$$

and

$$P\left(U_r = 0 \mid V_r^s; D\right) = \frac{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 0, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid D\right)}{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid D\right)}.$$

Then

$$R_r = \frac{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 1, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid D\right)}{P\left(U_1^s = u_1, ..., U_{r-1}^s = u_{r-1}, U_r = 0, U_{r+1}^{s-1} = u_{r+1}, ..., U_{n-1}^{s-1} = u_{n-1} \mid D\right)}.$$

Hence

$$R_r = \frac{p(D|M_{\hat{S}_1})p(M_{\hat{S}_1})}{p(D|M_{\hat{S}_2})p(M_{\hat{S}_2})},$$

where $\hat{S}_1 = \{i : u_i^s = 1, i < r\} \cup \{r\} \cup \{i : u_i^{s-1} = 1, i > r\}$ and $\hat{S}_2 = \{i : u_i^s = 1, i < r\} \cup \{i : u_i^{s-1} = 1, i > r\}$ and consequently, the criterion of choosing the values U_i^s, $i = 1, ..., k$, becomes:

$$U_r^s = \begin{cases} 1, & \text{if} \quad R_r \geq \frac{1-u}{u} \\ 0, & \text{otherwise} \end{cases}$$

for $r = 1, ..., p$, in which $u \sim Unif\,(0, 1)$.

4.6 Variable selection

We now provide an algorithm to minimize the loss function of equation (4.5) based on Proposition 4.2.

Algorithm 4.1 Variable selection

Step 1: Set $j = 1, S_1 = \{1, \ldots, k\}$ and evaluate $SC(M_{S_1})$.

Step 2: Set $j = 2$. Obtain $l \in S_1$ such that if $S_2 = S_1 - \{l\}$, $SC(M_{S_2})$ is minimum.

Step 3: If $SC(M_{S_2}) > SC(M_{S_1})$ or $S_2 = \varnothing$, set $\hat{S}^* = S_1$ and STOP. Otherwise go to step 4.

Step 4: Set $j = j + 1$. Obtain $l \in S_{j-1}$ such that if $S_j = S_{j-1} - \{l\}$, $SC(M_{S_j})$ is minimum.

Step 5: If $SC(M_{S_j}) > SC(M_{S_{j-1}})$ or $S_j = \varnothing$, set $\hat{S}^* = S_{j-1}$ and STOP. Otherwise go to step 4.

4.7 Correlation measures

The following example shows the consequences of multicollinearity in a multiple regression analysis. Let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\mu, \Sigma)$ with $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0.999999 \\ 0.999999 & 1 \end{pmatrix}$. Let $\epsilon \sim N(0, 0.1)$. We define the random variable $Y = 3X_1 + 2X_2 + \epsilon$.

We simulated 50 times these variables and performed a classical linear regression analysis. We obtained the following estimations for β_1 and β_2:

$\hat{\beta}_1 = 6.841$ and $\hat{\beta}_2 = -1.812$. We repeated this experiment 20 times. Table 4.1 shows the estimates of $\hat{\beta}_i$ for each experiment.

Simulation	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 + \hat{\beta}_2$
1	6.841	-1.812	5.029
2	5.735	-0.72	5.015
3	-6.031	11.029	4.998
4	3.354	1.649	5.003
5	-11.445	16.44	4.995
6	-2.369	7.363	4.994
7	9.461	-4.486	4.975
8	-7.581	12.56	4.979
9	2.653	2.362	5.015
10	15.451	-10.426	5.025
11	0.688	4.328	5.016
12	7.893	-2.888	5.005
13	-1.005	5.979	4.974
14	4.015	0.982	4.997
15	0.718	4.302	5.02
16	-4.945	9.945	5
17	-1.587	6.59	5.003
18	-5.79	10.786	4.996
19	-4.071	9.071	5
20	12.944	-7.937	5.007

Table 4.1: Estimated coefficients in the multicollinearity experiment

This simple experiment shows that multicollinearity can produce estimators of the coefficients with large variance and different signs. It also suggests a heuristic explanation of why multicolinearity can be very problematic. Note that $\hat{\beta}_1 + \hat{\beta}_2 \approx 5$. Since $X_1 \approx X_2$, and the true model is $Y = 3X_1 + 2X_2 + \epsilon$, there are several combinations of $\hat{\beta}_1$ and $\hat{\beta}_2$ such that $\hat{\beta}_1 + \hat{\beta}_2 \approx 5$. We need a real number $\psi(\boldsymbol{X}_{\hat{S}})$ which can be used in equation (4.3) that measures and the correlation of a group of random variables. We can find meaningful collinearity measures in Willan and Watts (1978). In fact, these authors show that $|C|^{\frac{1}{2}}$ is the volume generated by the standardized variables where C is the correlation matrix of the random variables considered. Note that

$$0 \leq |C|^{\frac{1}{2}} \leq 1$$

In Figures 4.3 and 4.4 we show examples of the measure $|C|^{\frac{1}{2}}$ for two and three dimensions respectively.

Example 4.6. *Let $X_i \sim N(0,1)$. We construct the following random variables:*

$$\begin{aligned} Z_1 &= 3X_3 + X_1 \\ Z_2 &= -3X_3 + 2X_2 \\ Z_3 &= 10X_1 + 2X_4 \\ Z_4 &= -2X_4 + X_5 \\ Z_5 &= -X_5 + 5X_6 \\ Z_6 &= -5X_6 + 20X_2 \end{aligned}$$

Now consider the model $Y = X_1 + 2X_2 + \epsilon$. Clearly $Y = Z_1 + Z_2 + \epsilon$ and $Y = Z_3 + Z_4 + Z_5 + Z_6 + \epsilon$. We want to perform variable selection with Z_i as the independent variables. For model M_{S_1} with $S = \{1, 2\}$, we obtain the following correlation matrix:

	Z_1	Z_2
Z_1	1	-0.97
Z_2	-0.97	1

In this model, $|C|^{\frac{1}{2}} = 0.266$. On the other hand, in model M_{S_2} with $S_2 = \{3, 4, 5, 6\}$, the following correlation matrix is obtained:

	Z_3	Z_4	Z_5	Z_6
Z_3	1	-0.175	0	0
Z_4	-0.175	1	-0.087	0
Z_5	0	-0.087	1	-0.237
Z_6	0	0	-0.237	1

In this model, $|C|^{\frac{1}{2}} = 0.958$*. This example shows the importance of considering a correlation measure in the loss function. In fact, model* M_{S_1} *has only two variables but they are highly correlated, which can cause several problems, whereas model* M_{S_2} *has four variables with low correlation which leads to robust estimations.*

Other correlation measures based on copulas can be found in Schmid et al. (2010).

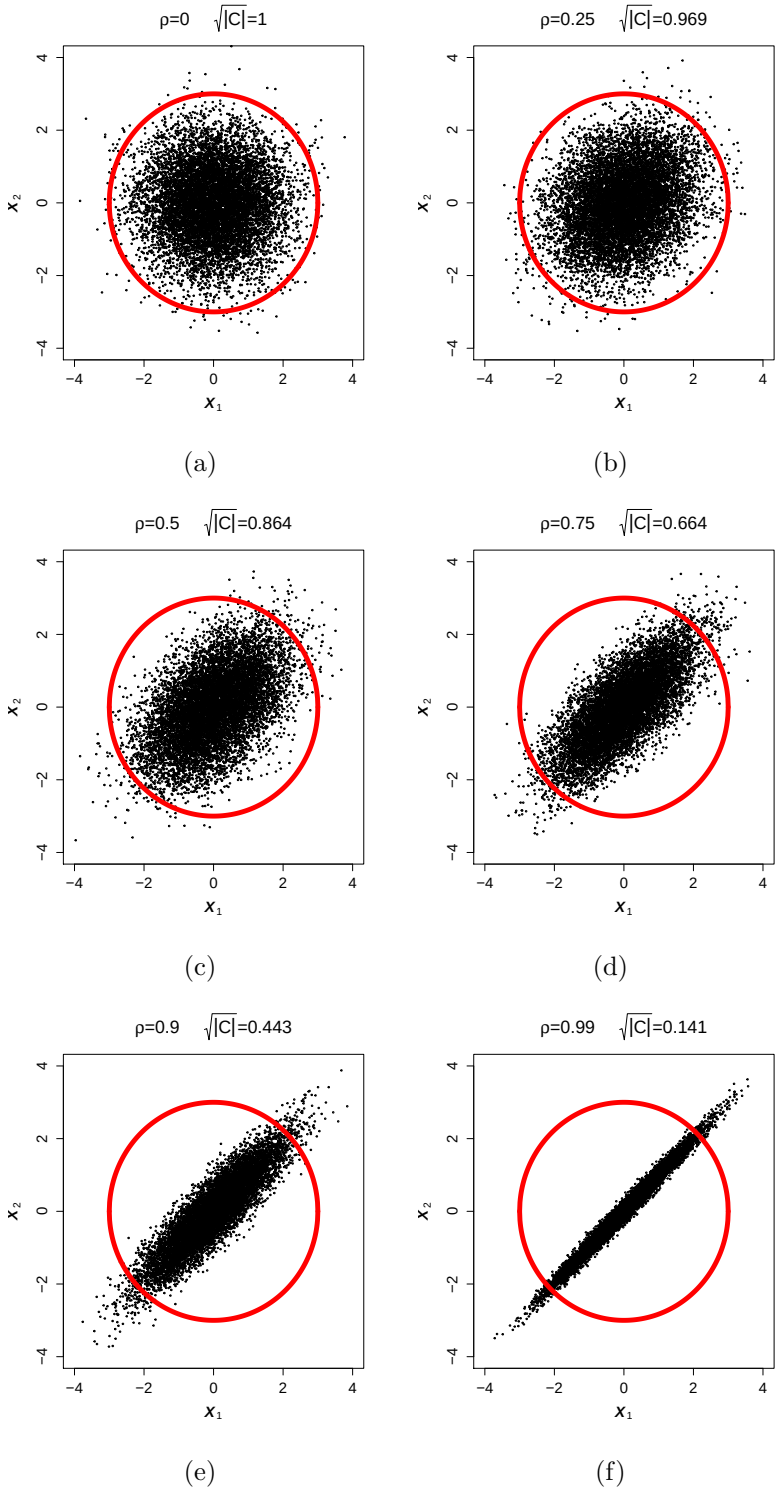
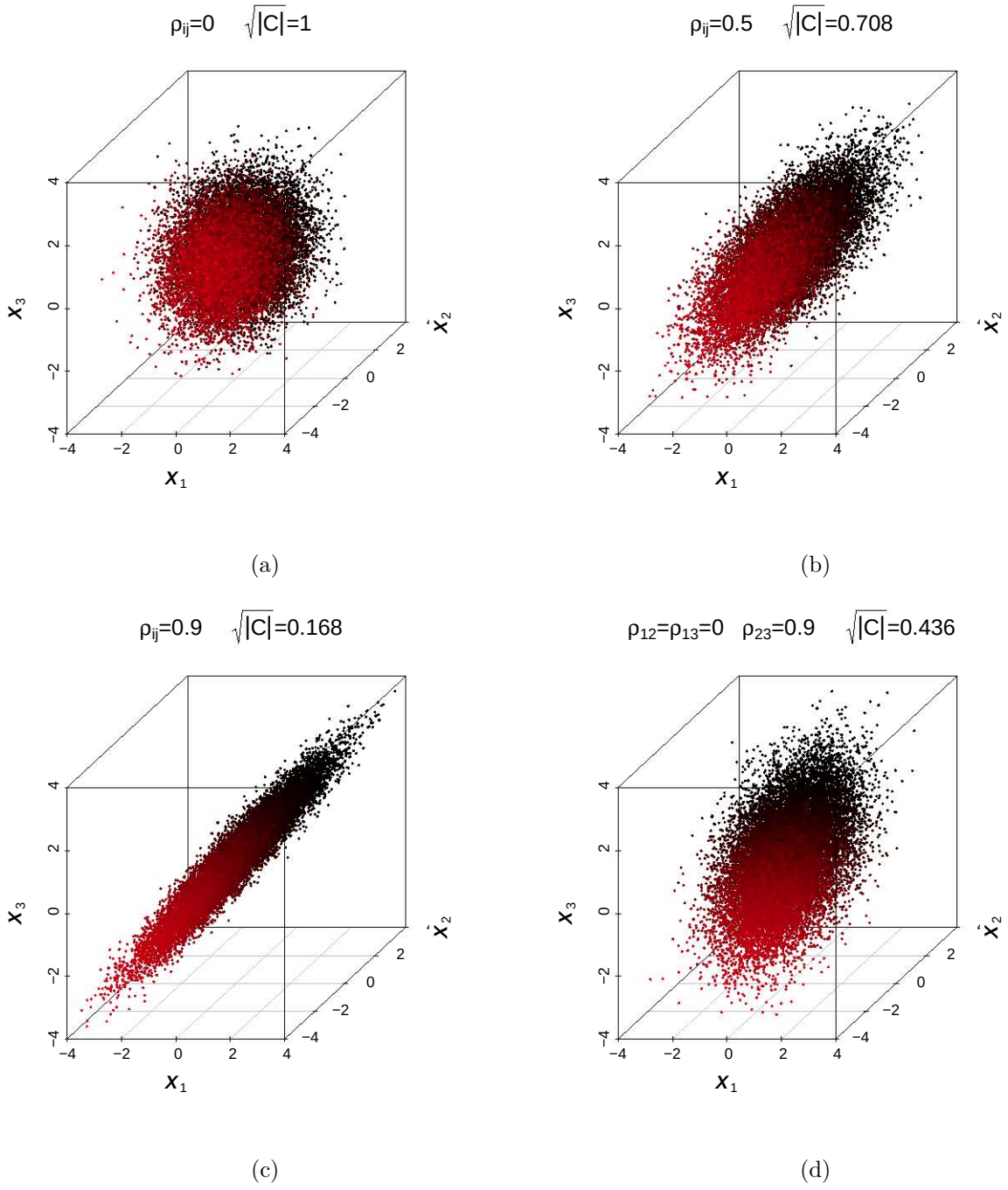


Figure 4.3: $|C|^{\frac{1}{2}}$ measure for bivariate normal distribution with different correlations.



(a) (b) (c) (d)

Figure 4.4: $|C|^{\frac{1}{2}}$ measure for multivariate normal distribution with different correlation matrices

4.8 Discussion

In this chapter, we proposed a new decision setting for variable selection problem in regression analysis with a loss function suitable for prediction. We proved a result that allowed to find the model which minimizes the expected loss function. We also provided a Gibbs sampler algorithm that helped us to optimize this loss function. In order to understand the difficulties that can arise in variable selection for regression analysis, we provided several illustrative examples.

When we suspect that the true model does not belong to our list of entertained model (for instance, if we think that the relationship between the dependent variable and independent variable might not be linear and we consider only linear functions) we could adopt the $\mathcal{M}$-open framework discussed in (Gutiérrez-Peña and Walker, 2001, 2005) and use a nonparametric model such as the Gaussian process with the Kullback-Leibler divergence as the loss function. We would then be choosing the linear model which is closest (on average) to the true model.

Chapter 5

Discussion

If you want others to be happy, practice compassion. If you want to be happy, practice compassion.

Dalai Lama XIV

5.1 Conclusions

In this thesis, we have proposed suitable loss functions for each of the three statistical problems we considered. For each case, we provided algorithms that can be regarded as applications and/or an extensions of the methodology proposed by Quintana and Iglesias (2003). Although their work has been extensively cited, mainly because they establish a link between PPMs and the DP, to the best of our knowledge the present thesis is the first work that exploits their proposal in order to make inferences based on loss functions in different settings other than clustering in a parametric framework.

In what follows, we will describe the contributions of each chapter of the thesis, including remarks that may be helpful in future work.

In Chapter 2, we used the Dirichlet process to extend PPMs to the nonparametric case, which does not impose any particular parametric form of the distribution function. We thus provided a methodology that uses loss functions to detect change points and that can also be applied to other models such as the nonparametric hidden Markov models. We also defined a simple loss function that can detect changes in the mean. This is very useful in many applications such as the analysis of aCGH data, where one

needs to process the human genome to detect abnormalities in the DNA that cause a specific disease. We also introduced more complex loss functions which are adequate to detect changes in the distribution. A general form was presented in Section 2.1; based on Anderson–Darling and Cramér–von Mises statistics, we used three different combinations of weights and measures that allowed us to detect changes in the tails as well as skewness more efficiently. For the loss functions proposed, we implemented a procedure to determine the value of the parameter γ, which appears in similar loss functions but in different settings such as variable selection. In many applications, missing data appears frequently. In order to tackle this difficulty, we took advantage of the random partition structure of our model to estimate the distribution function of each missing value without using multiple imputation, which is computationally less efficient. This procedure can also be used in the parametric product partition models for change-point analysis in order to estimate the "gold standards" of Theorems 2.9 and 2.10, we provided an exact computational procedure which is feasible for less than 300 observations, and a Gibbs sampler for larger data sets.

We focused our attention on the NPPMs for change-point detection purposes. However, NPPMs can also be used in a general clustering analysis. In fact, there is an interesting relationship with the nested Dirichlet process (NDP), in the same way as the PPMs are related with the Dirichlet process: by integrating out the Dirichlet process in the NDP, we obtained a particular case of our model. This result has also practical interest because we were able to profit from several simulation algorithms and apply then it to nonparametric outlier detection. This is so because the cohesion functions in this case can promote fewer clusters with larger amounts of data in each one.

We have shown empirically that methods based on loss functions may perform better than methods that use the marginal probability, which is a popular criterion. This is specially true in the case of the NPPMs, in which this criterion detected change points very poorly because of the flexibility of the model. Moreover, we have shown that our proposal performs better than parametric and nonparametric models recently discussed in the literature, such as PELT and ECP with different data sizes. We also applied our methodology to two real data sets in genetics and finance.

We used the low-level programming language Fortran in Ubuntu's operating system to implement the several methodologies discussed in Chapter 2. This allowed us to perform exhaustive simulations to obtain reasonable results. We used the R software

(R Core Team, 2016) as a GUI (graphical user interface) which allowed us to call the dynamic libraries compiled in GNU FORTRAN (gfortran) to analyze the output of the simulations. We also used OpenMP in gfortran to implement a parallelization of the simulations.

In Chapter 3, we argued for the use of more sophisticated methods such as non-parametric Bayesian statistics in finite population sampling, and stated a natural relationship between these two branches of statistics. We discussed a novel approach to Bayesian stratified sampling that allowed us to take advantage of prior knowledge about the structure of the population. This approach provides a methodology for collapsing strata in the post-stratification context. For this purpose, we induced the partition over the range of the covariates instead of the target variables. Since the inference is challenging and limited by the number of covariates, any improvement in the algorithm could allow us to consider more complex problems.

Finally, in Chapter 4, we studied the variable selection problem in regression analysis and constructed a loss function which includes a penalization correlations between the explanatory variables included in the model, in addition to the penalization on the number of variables. The construction of this loss function was carefully justified. We showed an example where one model had two highly correlated significant variables, while a second model contained five significant variables with low correlation. This example shows why we need to include a term in the loss function that penalizes correlation besides the size or complexity of the model. We also proved a result (Proposition 4.2) that allowed us to devise a methodology to select the model with minimum expected loss. A Gibbs sampler was described to obtain the "gold standards" required by Proposition 4.2. This methodology can be used in generalized linear models as well.

5.2 Future work

Theorem 2.10 will still hold if we use random measures other than the Dirichlet process to define NPMMs. However, practical concerns will arise, such as the calculation of the predictive distribution of the random measure employed for each block of each simulated partition. Since the number of partitions may be very large, this can result in an insurmountable problem. In the case of the Dirichlet and Poisson-Dirichlet

processes, this can be done because closed-form expressions of their multivariate predictive distributions are known. In the case of mixtures of Dirichlet process we can deal with this issue using variational methods, introduced in Bayesian nonparametric statistics by researchers on machine learning (see Blei and Jordan, 2006). Variational methods can provide an adequate approximation in much fewer iterations than MCMC strategies; see Bishop (2006, chap. 10) for an introduction to this methods in parametric Bayesian models. The calculation of the predictive distribution for most of the other random measures is computationally very demanding. A natural extension of the NPPMs would be the inclusion of covariates, in the same way as Park and Dunson (2010) or Müller and Quintana (2010); Müller et al. (2011). Although the generalization to several dimensions is straightforward, it was not developed here. On the other hand, while NPPMs can be defined in arbitrary dimensions; they have practical limitations due to the number of calculations needed. Therefore, it would be convenient to develop simulation algorithms that can be parallelized to profit from the number of CPU's available. In the NPPM defined in equation (2.15), we used the DP with dispersion parameter α_i and centered at G_i which is a parametric distribution function with parameter $\boldsymbol{\theta}_i$. It would be convenient to use a prior distribution on $\boldsymbol{\theta}_i$ and α_i. Algorithm 1.1 finds a local optimum. Alternatively, we could use other approaches such as agglomerative strategies or Simulated Annealing (SA)(Kirkpatrick et al., 1983). Although these algorithms do not ensure convergence to the global optimum either, they do not get stuck in local optima and can improve the solutions obtained.

Concerning our model for finite population sampling, an efficient computation implementation is also needed, perhaps using parallel programing, due to the number of possible stratifications. For instance, if we have a set of five discrete covariates, each one with 3 levels, we will have $3^5 = 243$ basic strata. Using the Bell number in equation (1.1), we have $B_{243} > 10^{275}$, which is the number of possible ways in which we can stratify our population using the five covariates. Clearly, this is computationally challenging.

Appendix A

Bayesian clustering and product partition models

Quintana and Iglesias (2003) developed the following result to obtain the partition with the minimum expected loss.

Let

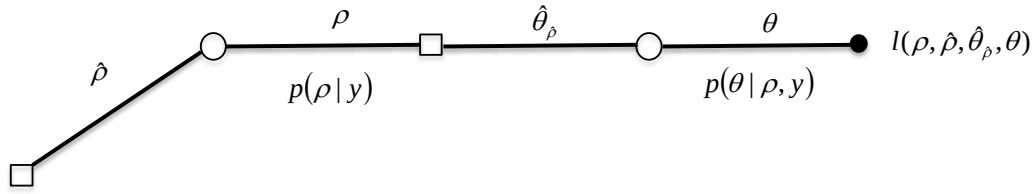
$$l\left(\rho, \hat{\rho}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\hat{\rho}}\right) = \gamma \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\hat{\rho}} \right\|^2 + (1-\gamma)\,|\hat{\rho}| \tag{A.1}$$

Theorem A.1. *Let* $\hat{\boldsymbol{\theta}}_B(\boldsymbol{y}) = E(\boldsymbol{\theta}|\boldsymbol{y})$ *and* $\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y}) = E(\boldsymbol{\theta}|\rho, \boldsymbol{y})$*, then the expected loss minimization criterion leads to the choice* $\hat{\rho}^*$ *that minimizes*

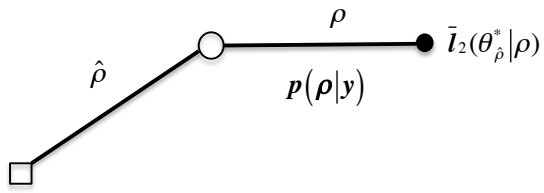
$$SC(\hat{\rho}) = \gamma \|\hat{\boldsymbol{\theta}}_B(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y})\|^2 + (1-\gamma)|\hat{\rho}| \tag{A.2}$$

where the vector $\hat{\boldsymbol{\theta}}_{\hat{\rho}}$ *is the estimate of the vector* $\boldsymbol{\theta}$ *associated with the estimated partition* $\hat{\rho}$*, and* $|\hat{\rho}|$ *denotes the cardinality of* $\hat{\rho}$*. Here,* $0 \leq \gamma \leq 1$ *is a complexity-cost parameter. The choice of loss function implies a trade-off (controlled by the user-defined quantity* γ*) between the optimal estimator of* $\boldsymbol{\theta}$ *and model simplicity, by which we mean a model with a low number of clusters or strata.*

Proof. Quintana and Iglesias (2003) For a sequential decision problem, Bernardo and Smith (1994) state that one has to first solve the final *nth* stage by minimizing the appropriate loss function, then one has to solve the $(n-1)th$ stage by minimizing the expected loss function conditional on making the optimal choice at the *nth* stage; and so on, working backwards progressively, until the optimal first stage option has been obtained. To visualize the decision tree of our problem see Figure A.1.



(a) Second stage



(b) First stage

Figure A.1: Decision tree of theorem A.1

To solve the optimization of the second stage:

$$\begin{aligned}\bar{l}_2(\hat{\boldsymbol{\theta}}_{\hat{\rho}}|\rho) &= \int l(\rho, \hat{\rho}, \hat{\boldsymbol{\theta}}_{\hat{\rho}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\rho, \boldsymbol{y}) d\boldsymbol{\theta} \\ &= \gamma \int \|\hat{\boldsymbol{\theta}}_{\hat{\rho}} - \boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\rho, \boldsymbol{y}) d\boldsymbol{\theta} + (1-\gamma)|\hat{\rho}|.\end{aligned}$$

Let $\hat{\boldsymbol{\theta}}^*_{\hat{\rho}} = \text{argmin}_{\hat{\boldsymbol{\theta}}_{\hat{\rho}}}\{\bar{l}_2(\hat{\boldsymbol{\theta}}_{\hat{\rho}}|\rho)\}$, then the loss of choosing $\hat{\rho}$ when ρ is the real state of the world would be $\bar{l}_2(\hat{\boldsymbol{\theta}}^*_{\hat{\rho}}|\rho)$. Now we will solve the first stage (see Figure A.1b). The expected loss is given by

$$\begin{aligned}\bar{l}(\hat{\rho}) &= \int \bar{l}_2(\hat{\boldsymbol{\theta}}^*_{\hat{\rho}}|\rho) p(\rho|\boldsymbol{y}) d\rho \\ &= \gamma \int \left\{ \int \|\hat{\boldsymbol{\theta}}^*_{\hat{\rho}} - \boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\rho, \boldsymbol{y}) d\boldsymbol{\theta} \right\} p(\rho|\boldsymbol{y}) d\rho + (1-\gamma)|\hat{\rho}| \\ &= \gamma \int \|\hat{\boldsymbol{\theta}}^*_{\hat{\rho}} - \boldsymbol{\theta}\|^2 \int p(\boldsymbol{\theta}|\rho, \boldsymbol{y}) p(\rho|\boldsymbol{y}) d\rho d\boldsymbol{\theta} + (1-\gamma)|\hat{\rho}| \\ &= \gamma \int \|\hat{\boldsymbol{\theta}}^*_{\hat{\rho}} - \boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} + (1-\gamma)|\hat{\rho}|\end{aligned}$$

Since $\hat{\boldsymbol{\theta}}^*_{\hat{\rho}}$ is the Bayes estimate under a quadratic loss function,

$$\hat{\boldsymbol{\theta}}^*_{\hat{\rho}} = E(\boldsymbol{\theta}|\rho, \boldsymbol{y}) = \hat{\boldsymbol{\theta}}_{\rho}(\boldsymbol{y})$$

Finally,

$$\begin{aligned}
\int \|\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y}) - \boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} &= \int \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_{\rho,i}(\boldsymbol{y}) - \boldsymbol{\theta}_i)^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\
&= \int \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_{\rho,i}(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}_{B,i}(\boldsymbol{y}) + \hat{\boldsymbol{\theta}}_{B,i}(\boldsymbol{y}) - \boldsymbol{\theta}_i)^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\
&= \int \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_{\rho,i}(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}_{B,i}(\boldsymbol{y}))^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\
&\quad + 2 \int \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_{\rho,i}(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}_{B,i}(\boldsymbol{y}))(\hat{\boldsymbol{\theta}}_{B,i}(\boldsymbol{y}) - \boldsymbol{\theta}_i) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\
&\quad + \int \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_{B,i}(\boldsymbol{y}) - \boldsymbol{\theta}_i)^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\
&= \|\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}_B(\boldsymbol{y})\|^2 + 0 + \mathrm{tr}(V(\boldsymbol{\theta}|\boldsymbol{y}))
\end{aligned}$$

where $tr(A)$ denotes the trace of a given matrix A. Then

$$\gamma \int \|\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y}) - \boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} + (1-\gamma)|\hat{\rho}| = \gamma \|\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y}) - \hat{\boldsymbol{\theta}}_B(\boldsymbol{y})\|^2 + \gamma \mathrm{tr}(V(\boldsymbol{\theta}|\boldsymbol{y})) + (1-\gamma)|\hat{\rho}|$$

The proof concludes by noting that the second term in the last expression does not depend on $\hat{\rho}$. □

The above result shows that the optimal choice $\hat{\rho}^*$ will be the partition for which the resulting estimate $\hat{\boldsymbol{\theta}}_\rho(\boldsymbol{y})$ is closest to $\hat{\boldsymbol{\theta}}_B(\boldsymbol{y})$.

A. BAYESIAN CLUSTERING AND PRODUCT PARTITION MODELS

Appendix B

Distributions and Related Results

B.1 Distributions

Definition B.1. *Multivariate normal distribution. If* $\boldsymbol{X} \in \mathbb{R}^n$ *has a multivariate normal density, conventionally written* $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\lambda})$ *then*

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\lambda}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\lambda}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

where $E(\boldsymbol{X}) = \boldsymbol{\mu}$ *and* $Var(\boldsymbol{X}) = \boldsymbol{\lambda}^{-1}$ *provided* $\boldsymbol{\lambda}$ *is positive definite.*

Remark B.2. *We will use the notation* $N(\boldsymbol{\mu}, \boldsymbol{\lambda})$ *and* $N_n(\boldsymbol{\mu}, \boldsymbol{\lambda})$ *indistinctively for simplicity purposes.*

The next definition concerns to the multivariate noncentral t-distribution

Definition B.3. *Multivariate t density. If* $\boldsymbol{X} \in \mathbb{R}^n$ *has a multivariate t density, conventionally written* $\boldsymbol{X} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{V})$*, then*

$$p(\boldsymbol{x}) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{n}{2}}} |\boldsymbol{V}|^{-\frac{1}{2}} \left(1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}} \tag{B.1}$$

$$p(x) \quad \propto \quad \left(1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}} \tag{B.2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ *is a location parameter.* $\boldsymbol{V}$ *is a* $n \times n$*, symmetric, positive definite matrix and* $\nu > 0$ *is a degrees of freedom parameter. If* $\nu > 1$ *then* $E(\boldsymbol{X}) = \boldsymbol{\mu}$*; otherwise it is undefined.*

If $\nu > 2$ then $Var(\boldsymbol{X}) = \frac{\nu}{\nu-2}\boldsymbol{V}$; otherwise it is undefined.
Note that if $n = 1$, $\mu = 0$, and $\boldsymbol{V} = 1$, then equation (B.1) becomes the pdf of the univariate Student's t distribution with ν degrees of freedom. See Kotz and Nadarajah (2004) for further properties and applications.

The following definition extends the definition of the normal-gamma distribution.

Definition B.4. *Multivariate normal-gamma distribution. A continuous random vector $\boldsymbol{X} = (X_1, ..., X_n)$ and a random quantity τ have a joint multivariate normal-gamma distribution of dimension n, with parameters $\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta$ where $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\lambda}$ a $n \times n$ symmetric, positive-definite matrix, $\alpha > 0$ and $\beta > 0$ if the joint probability density of $\boldsymbol{X}$ and τ, $Ng_n(\boldsymbol{x}, \tau|\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta)$ is*

$$Ng_n(\boldsymbol{x}, \tau|\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta) = N_n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\lambda}\tau)\, Ga(\tau|\alpha, \beta) \tag{B.3}$$

Remark B.5. *As in the case of the multivariate normal distribution, we will use indistinctively $Ng_n(\boldsymbol{x}, \tau|\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta)$ and $Ng(\boldsymbol{x}, \tau|\boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta)$*

The skew normal distribution is a generalization of the normal distribution that can have non zero skewness and is defined as follows.

Definition B.6. *The density of a skew normal random variable has the density of the form*

$$f(x|\xi, \omega, \alpha) = 2\phi(\frac{x-\xi}{\omega})\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right) \tag{B.4}$$

where ϕ and Φ are the density and the distribution function, respectively, of the standard normal random variable. ξ is the location, ω is the scale and α is the shape parameter which determines the skewness.

B.2 Proofs of selected propositions

The following result is very useful to complete quadratic forms.

Lemma B.7. *Let $\boldsymbol{z}$,$\boldsymbol{b}$ and $\boldsymbol{c}$ be an $n \times 1$ vectors, A an $n \times n$ symmetric and invertible matrix. Then*

$$\boldsymbol{z}^t A\boldsymbol{z} + 2\boldsymbol{b}^t\boldsymbol{z} + \boldsymbol{c} = \left(\boldsymbol{z} + A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{z} + A^{-1}\boldsymbol{b}\right) + \boldsymbol{c} - \boldsymbol{b}^t A^{-1}\boldsymbol{b}$$

Proof.

$$\begin{aligned}\left(\boldsymbol{z}+A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{z}+A^{-1}\boldsymbol{b}\right)+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b} &= \boldsymbol{z}^tA\boldsymbol{z}+\boldsymbol{z}^tAA^{-1}\boldsymbol{b}+\boldsymbol{b}^t\left(A^{-1}\right)^tA\boldsymbol{z}\\ &\quad +\boldsymbol{b}^t\left(A^{-1}\right)^tAA^{-1}\boldsymbol{b}+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}\end{aligned}$$

Since A is symmetric, we have that $(A^{-1})^t = A^{-1}$, hence

$$\left(\boldsymbol{z}+A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{z}+A^{-1}\boldsymbol{b}\right)+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b} = \boldsymbol{z}^tA\boldsymbol{z}+2\boldsymbol{b}^t\boldsymbol{z}+\boldsymbol{c}$$

□

Proposition B.8. *Let* $\boldsymbol{z}$,$\boldsymbol{b}$ *and* $\boldsymbol{c}$ *be an* $n\times 1$ *vectors,* A *an* $n\times n$ *symmetric and invertible matrix. Then*

$$\boldsymbol{z}^tA\boldsymbol{z}-2\boldsymbol{b}^t\boldsymbol{z}+\boldsymbol{c}=\left(\boldsymbol{z}-A^{-1}\boldsymbol{b}\right)^tA\left(\boldsymbol{z}-A^{-1}\boldsymbol{b}\right)+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}$$

Proof.

$$\begin{aligned}\left(\boldsymbol{z}-A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{z}-A^{-1}\boldsymbol{b}\right)+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b} &= \boldsymbol{z}^tA\boldsymbol{z}-\boldsymbol{z}^tAA^{-1}\boldsymbol{b}-\boldsymbol{b}^t\left(A^{-1}\right)^tA\boldsymbol{z}\\ &\quad +\boldsymbol{b}^t\left(A^{-1}\right)^tAA^{-1}\boldsymbol{b}+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}\\ &= \boldsymbol{z}^tA\boldsymbol{z}-2\boldsymbol{b}^t\boldsymbol{z}+\boldsymbol{c}\end{aligned}$$

□

Lemma B.9. *Let* H *be an* $n\times n$ *matrix with* $H_{ij}=1\ \forall i,j$ *then*

$$HH = nH \tag{B.5}$$

Proof.

$$\begin{aligned}(HH)_{ij} &= \sum_{k=1}^{n} H_{ik}H_{km}\\ &= n\ \forall i,j.\end{aligned}$$

□

Lemma B.10. *Let* H *be an* $n\times n$ *matrix with* $H_{ij}=1\ \forall i,j$, $\boldsymbol{I}_n$ *the* $n\times n$ *identity matrix and* b_0 *a positive real number then*

$$\left(\boldsymbol{I}_n-\frac{H}{n+b_0}\right)^{-1}=\left(\boldsymbol{I}_n+\frac{H}{b_0}\right)$$

Proof.

$$\begin{aligned}\left(\boldsymbol{I}_n+\frac{H}{b_0}\right)\left(\boldsymbol{I}_n-\frac{H}{n+b_0}\right) &= \left(\boldsymbol{I}_n+\frac{H}{b_0}\right)\left(\boldsymbol{I}_n-\frac{H}{n+b_0}\right)\\ &= \boldsymbol{I}_n-\frac{H}{n+b_0}+\frac{H}{b_0}-\frac{HH}{b_0\left(n+b_0\right)}\end{aligned}$$

Using Lemma B.9

$$\begin{aligned}&= \boldsymbol{I}_n+\frac{(n+b_0)H-b_0H-nH}{b_0(b+b_0)}\\ &= \boldsymbol{I}_n\end{aligned}$$

□

Proposition B.11. *Let* $X_i|\mu,\tau \overset{i.i.d}{\sim} N\left(\mu,\tau\right)$ *for* $1\leq i\leq n$ *and* $\mu\sim N(a_0,b_0\tau)$ *where* τ *and* $b_0\tau$ *are the respective precisions. Then* $\boldsymbol{X}|\tau\sim N\left(\boldsymbol{a_0},\tau\left(\boldsymbol{I}_n-\frac{H}{n+b_0}\right)\right)$ *where* $\boldsymbol{X}=(X_1,...,X_n)$, $\boldsymbol{a_0}\in\mathbb{R}^n$, $\boldsymbol{a_0}=(a_0,...,a_0)$, H *an* $n\times n$ *matrix with* $H_{ij}=1$ *for* $1\leq i\leq n$ *and* $1\leq j\leq n$.

Proof.

$$\begin{aligned}p\left(\boldsymbol{x}\right) &= \int p\left(\boldsymbol{x}|\mu\right)p(\mu)d\mu\\ &\propto \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t\tau\boldsymbol{I}_n\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\}\exp\left\{-\frac{b_0\tau}{2}\left(\mu-a_0\right)^2\right\}d\mu\\ &\propto \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t\tau\boldsymbol{I}_n\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\}\exp\left\{-\frac{b_0\tau}{2}\left(\mu^2-2\mu a_0+a_0^2\right)\right\}d\mu\end{aligned}$$

where $\boldsymbol{x}=(x_1,...,x_n)^t$ and $\boldsymbol{\mu}=(\mu,\mu,\mu,...,\mu)^t$

$$\begin{aligned}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t\tau\boldsymbol{I}_n\left(\boldsymbol{x}-\boldsymbol{\mu}\right) &= \boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x}-2\boldsymbol{\mu}^t\tau\boldsymbol{I}_n\boldsymbol{x}+\boldsymbol{\mu}^t\tau\boldsymbol{I}_n\boldsymbol{\mu}\\ &= \boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x}-2\boldsymbol{\mu}^t\tau\boldsymbol{I}_n\boldsymbol{x}+n\tau\mu^2\\ &= \boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x}-2\mu\tau(x_1+x_2+...+x_n)+n\tau\mu^2\end{aligned}$$

Hence

$$\begin{aligned}p\left(\boldsymbol{x}\right)&\propto \int \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x}-2\mu(\tau(x_1+..+x_n)+a_0b_0\tau)+\left(n\tau+b_0\tau\right)\mu^2+a_0^2b_0\tau\right]\right\}d\mu\\ &\propto \int \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x}+\left(n\tau+b_0\tau\right)\left(\mu^2-\frac{2\mu}{n+b_0}\left(x_1+..+x_n+a_0b_0\right)\right)+a_0^2b_0\tau\right]\right\}d\mu\\ &\propto \int \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x}+\left(n\tau+b_0\tau\right)\left(\mu^2-\frac{2\mu}{n+b_0}\left(x_1+..+x_n+a_0b_0\right)\right)\right.\right.\\ &\qquad\left.\left.+\left(n\tau+b_0\tau\right)\left(\left(\frac{x_1+..+x_n+a_0b_0}{n+b_0}\right)^2-\left(\frac{(x_1+..+x_n)+a_0b_0}{n+b_0}\right)^2\right)\right]\right\}d\mu\end{aligned}$$

$$
\begin{aligned}
&\propto \int \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} + (n\tau+b_0\tau)\left(\mu - \frac{(x_1+..+x_n)+a_0b_0}{n+b_0}\right)^2\right.\right.\\
&\qquad\left.\left. - (n\tau+b_0\tau)\left(\frac{(x_1+..+x_n)+a_0b_0}{n+b_0}\right)^2\right]\right\}d\mu\\
&\propto \int \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} - \frac{\tau\left((x_1+..+x_n)+a_0b_0\right)^2}{n+b_0}\right.\right.\\
&\qquad\left.\left. + (n\tau+b_0\tau)\left(\mu - \frac{(x_1+..+x_n)+a_0b_0}{n+b_0}\right)^2\right]\right\}d\mu\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} - \frac{\tau\left((x_1+..+x_n)+a_0b_0\right)^2}{n+b_0}\right]\right\}\\
&\qquad\times \int \exp\left\{-\frac{1}{2}\left[(n\tau+b_0\tau)\left(\mu - \frac{(x_1+..+x_n)+a_0b_0}{n+b_0}\right)^2\right]\right\}d\mu\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} - \frac{\tau\left((x_1+..+x_n)+a_0b_0\right)^2}{n+b_0}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} - \frac{\tau\left((x_1+..+x_n)^2+2a_0b_0\left(x_1+..+x_n\right)+a_0^2b_0^2\right)}{n+b_0}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} - \frac{\tau\left(\boldsymbol{x}^tH\boldsymbol{x}+2a_0b_0\left(x_1+..+x_n\right)\right)}{n+b_0}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\boldsymbol{I}_n\boldsymbol{x} - \tau\boldsymbol{x}^tH\boldsymbol{x} - \frac{2b_0\tau\boldsymbol{a_0}^tH\boldsymbol{x}}{n(n+b_0)}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{x}^t\tau\left(\boldsymbol{I}_n - \frac{H}{n+b_0}\right)\boldsymbol{x} - \frac{2b_0\tau\boldsymbol{a_0}^tH\boldsymbol{x}}{n(n+b_0)}\right]\right\}.
\end{aligned}
$$

Let $A = \tau\left(\boldsymbol{I}_n - \frac{H}{n+b_0}\right)$ and $\boldsymbol{b} = \frac{b_0\tau H\boldsymbol{a_0}}{n(n+b_0)}$, Using Lemma B.7 we obtain

$$
p\left(\boldsymbol{x}\right) \propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{x}-A^{-1}\boldsymbol{b}\right)^tA\left(\boldsymbol{x}-A^{-1}\boldsymbol{b}\right)\right]\right\}
$$

Using Lemma B.10,

$$
\begin{aligned}
A^{-1} &= \left(\tau\left(\boldsymbol{I}_n - \frac{H}{n+b_0}\right)\right)^{-1} &\text{(B.6)}\\
&= \tau^{-1}\left(\boldsymbol{I}_n + \frac{H}{b_0}\right) &\text{(B.7)}
\end{aligned}
$$

Therefore

$$
A^{-1}\boldsymbol{b} = \left(\tau^{-1}\boldsymbol{I}_n + \frac{H}{\tau b_0}\right)\left(\frac{b_0\tau H\boldsymbol{a_0}}{n\left(n+b_0\right)}\right)
$$

$$\begin{aligned} &= \frac{b_0 H \boldsymbol{a_0}}{n\left(n+b_0\right)} + \frac{b_0 \tau H H \boldsymbol{a_0}}{b_0 \tau n\left(n+b_0\right)} \\ &= \frac{b_0 H \boldsymbol{a_0} + H H \boldsymbol{a_0}}{n\left(n+b_0\right)}. \end{aligned}$$

Using again Lemma B.9

$$\begin{aligned} &= \frac{\left(b_0 \boldsymbol{I}_n + n \boldsymbol{I}_n\right) H \boldsymbol{a_0}}{n\left(n+b_0\right)} \\ &= \boldsymbol{a_0}. \end{aligned}$$

Finally

$$\boldsymbol{X}|\tau \sim N\left(\boldsymbol{a_0}, \tau\left(\boldsymbol{I}_n - \frac{H}{n+b_0}\right)\right).$$

□

Proposition B.12. *In the multivariate normal-gamma distribution the marginal density of* $\boldsymbol{X}$ *is a multivariate t-distribution* $t_{2\alpha}\left(\boldsymbol{\mu}, \alpha^{-1}\beta\boldsymbol{\lambda^{-1}}\right)$

Proof. Begin by noting that the marginal density of $\boldsymbol{X}$ is

$$p\left(\boldsymbol{x}\right) = \int_0^\infty p\left(\boldsymbol{x}|\tau\right) p\left(\tau\right) d\tau$$

$$\begin{aligned} p\left(\boldsymbol{x}\right) &\propto \int_0^\infty |\tau\lambda|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\tau\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right) \tau^{\alpha-1} \exp\left(-\beta\tau\right) d\tau \\ p\left(\boldsymbol{x}\right) &\propto \int_0^\infty \tau^{\frac{n}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\tau\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right) \tau^{\alpha-1} \exp\left(-\beta\tau\right) d\tau \\ &\propto \int_0^\infty \tau^{\alpha-\frac{n}{2}-1} \exp\left(-\tau\left[\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)+\beta\right]\right) d\tau \end{aligned}$$

Notice that $\tau^{\alpha-\frac{n}{2}-1} \exp\left(-\tau\left[\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)+\beta\right]\right)$ is the kernel of a gamma density $Ga\left(\tau|a,b\right)$ with $a = \alpha + \frac{n}{2}$ and $b = \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)+\beta$. Hence

$$\begin{aligned} p(\boldsymbol{x}) &\propto \frac{\Gamma\left(a\right)}{b^a} \\ &\propto \left(\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)+\beta\right)^{-\frac{2\alpha+n}{2}} \\ &\propto \left(\frac{1}{2\alpha}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\alpha\beta^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)+1\right)^{-\frac{2\alpha+n}{2}} \end{aligned}$$

which is the kernel of a multivariate t-student $t_{2\alpha}\left(\boldsymbol{\mu}, \alpha^{-1}\beta\boldsymbol{\lambda}^{-1}\right)$ □

We will calculate the prior predictive distribution for the normal-gamma model.

Proposition B.13. *Let* $\boldsymbol{X} = (X_1, ..., X_n)$ *with* $X_i \overset{iid}{\sim} N(\mu, \tau)$ *with* $i = 1, ..., n$ *and* $\boldsymbol{\theta} = (\mu, \tau) \sim NG(a_0, b_0, \alpha_0, \beta_0)$ *then*

$$\boldsymbol{X} \sim t_{2\alpha}\left(\boldsymbol{a_0}, \alpha_0^{-1}\beta_0\left(\boldsymbol{I}_n + \frac{H}{b_0}\right)\right) \tag{B.8}$$

Proof.

$$\begin{aligned} p(\boldsymbol{x}) &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_0^{\infty} \int_{-\infty}^{\infty} p(\boldsymbol{x} \mid \mu, \tau) p(\mu \mid \tau) p(\tau) d\mu d\tau \\ &= \int_0^{\infty} \left[\int_{\infty}^{\infty} p\left(\boldsymbol{x} \mid \mu, \tau\right) p\left(\mu \mid \tau\right) d\mu\right] p(\tau) d\tau \end{aligned}$$

We will also use of the fact that $p\left(\mu, \tau\right) = p\left(\mu \mid \tau\right) p(\tau)$, with

$$\begin{aligned} \mu \mid \tau, a_0, b_0 &\sim N\left(a_0, b_0\tau\right) \\ \tau \mid \alpha_0, \beta_0 &\sim Gamma\left(\alpha_0, \beta_0\right) \end{aligned}$$

$\int_{\infty}^{\infty} p\left(\boldsymbol{x} \mid \mu, \tau\right) p\left(\mu \mid \tau\right) d\mu = p\left(\boldsymbol{x} \mid \tau\right)$

$$p(\boldsymbol{x} \mid \tau) = N\left(\boldsymbol{a_0}, \tau\left(\boldsymbol{I}_n - \frac{H}{n + b_0}\right)\right)$$

Using Proposition B.12 and Lemma B.10, we obtain

$$\boldsymbol{X} \sim t_{2\alpha}\left(\boldsymbol{a_0}, \alpha_0^{-1}\beta_0\left(\boldsymbol{I}_n + \frac{H}{b_0}\right)\right). \tag{B.9}$$

□

Appendix C

Prior Predictive of the Normal Regression Model

C.1 Normal-Gamma regression model

Definition C.1. *We first reparametrize the model by introducing the precision parameter* $\tau = \frac{1}{\sigma^2}$*, then*

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\beta},\tau &\sim N(\boldsymbol{X\beta},\tau\boldsymbol{I}_n) \\ \boldsymbol{\beta}|\tau &\sim N_{k+1}\left(\boldsymbol{\beta}_0,\tau\boldsymbol{V}\right) \\ \tau &\sim Gamma(\alpha,\beta) \end{aligned}$$

We need to calculate

$$p(\boldsymbol{y}) = \int\int p(\boldsymbol{y}|\boldsymbol{\beta},\tau)p(\boldsymbol{\beta}|\tau)p(\tau)d\boldsymbol{\beta}d\tau$$

Note that

$$\begin{aligned} p(\boldsymbol{y}) &= \int\int p(\boldsymbol{y}|\boldsymbol{\beta},\tau)p(\boldsymbol{\beta}|\tau)p(\tau)d\boldsymbol{\beta}d\tau \\ &\propto \int (2\pi)^{-\frac{n}{2}}\left|\tau\boldsymbol{I_n}\right|^{\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{X\beta}\right)^t\tau\boldsymbol{I_n}\left(\mathbf{y}-\boldsymbol{X\beta}\right)\right\} \\ &\qquad \times (2\pi)^{-\frac{k+1}{2}}\left|\tau\boldsymbol{V}\right|^{\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta_0}\right)^t\tau\boldsymbol{V}\left(\boldsymbol{\beta}-\boldsymbol{\beta_0}\right)\right\} p\left(\tau\right)d\boldsymbol{\beta}d\tau \\ &\propto \int \left|\tau\boldsymbol{I_n}\right|^{\frac{1}{2}}\left|\tau\boldsymbol{V}\right|^{\frac{1}{2}} p(\tau) exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}-\boldsymbol{y}^t\boldsymbol{X\beta}-\left(\boldsymbol{X\beta}\right)^t\mathbf{y}\left(\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X\beta}\right)\right.\right. \end{aligned}$$

$$
\begin{aligned}
&+\boldsymbol{\beta}^t\boldsymbol{V}\boldsymbol{\beta}-\boldsymbol{\beta}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{\beta}_0^t\boldsymbol{V}\boldsymbol{\beta}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\big]\big\}\,d\boldsymbol{\beta}d\tau\\
\propto&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}exp\left\{-\frac{\tau}{2}\left[-\boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta}-(\boldsymbol{X}\boldsymbol{\beta})^t\,\mathbf{y}+\left(\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}\right)\right.\right.\\
&+\boldsymbol{\beta}^t\boldsymbol{V}\boldsymbol{\beta}-\boldsymbol{\beta}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{\beta}_0^t\boldsymbol{V}\boldsymbol{\beta}\big]\big\}\,d\boldsymbol{\beta}d\tau\\
\propto&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}exp\left\{-\frac{\tau}{2}\left[\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{y}\right.\right.\\
&+\boldsymbol{\beta}^t\boldsymbol{V}\boldsymbol{\beta}-\boldsymbol{\beta}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{\beta}_0^t\boldsymbol{V}\boldsymbol{\beta}\big]\big\}\,d\boldsymbol{\beta}d\tau\\
\propto&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}exp\left\{-\frac{\tau}{2}\left[\boldsymbol{\beta}^t\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{V}\right)\boldsymbol{\beta}-2\boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta}\right.\right.\\
&-{\boldsymbol{\beta_0}}^t\boldsymbol{V^t}\boldsymbol{\beta}-\boldsymbol{\beta}_0^t\boldsymbol{V}\boldsymbol{\beta}\big]\big\}\,d\boldsymbol{\beta}d\tau\\
\propto&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}exp\left\{-\frac{\tau}{2}\left[\boldsymbol{\beta}^t\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{V}\right)\boldsymbol{\beta}-2\boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta}\right.\right.\\
&\left.\left.-2\boldsymbol{\beta}_0^t\boldsymbol{V}\boldsymbol{\beta}\big]\big\}\,d\boldsymbol{\beta}d\tau\right.\\
\propto&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}\\
&\times exp\left\{-\frac{\tau}{2}\left[\boldsymbol{\beta}^t\left(\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{V}\right)\boldsymbol{\beta}-2\left(\boldsymbol{y}^t\boldsymbol{X}+\boldsymbol{\beta}_0^t\boldsymbol{V}\right)\boldsymbol{\beta}\right]\right\}d\boldsymbol{\beta}d\tau
\end{aligned}
$$

Using Proposition B.8

$$
\begin{aligned}
p(\boldsymbol{y})\;\propto\;&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}\\
&\times exp\left\{-\frac{\tau}{2}\left[\left(\boldsymbol{\beta}-A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{\beta}-A^{-1}\boldsymbol{b}\right)+\boldsymbol{c}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}\right]\right\}d\boldsymbol{\beta}d\tau
\end{aligned}
$$

with $\boldsymbol{A}=\boldsymbol{X}^t\boldsymbol{X}+\boldsymbol{V}\;\boldsymbol{b^t}=\boldsymbol{y}^t\boldsymbol{X}+\boldsymbol{\beta}_0^t\boldsymbol{V}$ *and* $\boldsymbol{c}=0$, *therefore* $\boldsymbol{b}=\boldsymbol{X}^t\boldsymbol{y}+\boldsymbol{V}\boldsymbol{\beta}_0$

$$
\begin{aligned}
p(\boldsymbol{y})\;\propto\;&\int\tau^{\frac{n+k+1}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}\\
&\times exp\left\{-\frac{\tau}{2}\left[\left(\boldsymbol{\beta}-A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{\beta}-A^{-1}\boldsymbol{b}\right)-\boldsymbol{b}^tA^{-1}\boldsymbol{b}\right]\right\}d\boldsymbol{\beta}d\tau\\
\propto\;&\int\tau^{\frac{n}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}\right]\right\}\tau^{\frac{k+1}{2}}\\
&\times exp\left\{-\frac{\tau}{2}\left[\left(\boldsymbol{\beta}-A^{-1}\boldsymbol{b}\right)^t A\left(\boldsymbol{\beta}-A^{-1}\boldsymbol{b}\right)-\boldsymbol{b}^tA^{-1}\boldsymbol{b}\right]\right\}d\boldsymbol{\beta}d\tau\\
\propto\;&\int\tau^{\frac{n}{2}}p(\tau)exp\left\{-\frac{\tau}{2}\left[\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}\right]\right\}d\tau
\end{aligned}
$$

We now define $D=\boldsymbol{y}^t\boldsymbol{y}+{\boldsymbol{\beta_0}}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}$

$$
p(\boldsymbol{y})\;\propto\;\int\tau^{\frac{n}{2}}p(\tau)exp\left\{-\frac{\tau}{2}D\right\}d\tau
$$

$$
\begin{aligned}
p(\boldsymbol{y}) &\propto \int \tau^{\frac{n}{2}} \tau^{\alpha-1} exp\left\{-\tau\beta\right\} exp\left\{-\frac{\tau}{2}D\right\} d\tau \\
p(\boldsymbol{y}) &\propto \int \tau^{\alpha+\frac{n}{2}-1} exp\left\{-\tau(\beta+\frac{D}{2}\right\} d\tau
\end{aligned}
$$

This is the kernel of a Gamma distribution $Ga\left(a,b\right)$ *with* $a=\alpha+\frac{n}{2}$ *and* $b=\beta+\frac{D}{2}$, *hence*

$$
\begin{aligned}
p(\boldsymbol{y}) &\propto b^{-a} \\
&\propto \left(\frac{\boldsymbol{y}^t\boldsymbol{y}+\boldsymbol{\beta_0}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}+2\beta}{2}\right)^{-\alpha-\frac{n}{2}}
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{b}^tA^{-1}\boldsymbol{b} &= \left(\boldsymbol{y}^t\boldsymbol{X}+\boldsymbol{\beta}_0^t\boldsymbol{V}\right)A^{-1}\left(\boldsymbol{X}^t\boldsymbol{y}+\boldsymbol{V}\boldsymbol{\beta}_0\right) \\
&= \boldsymbol{y}^t\boldsymbol{X}A^{-1}\boldsymbol{X}^t\boldsymbol{y}+\boldsymbol{y}^t\boldsymbol{X}A^{-1}\boldsymbol{V}\boldsymbol{\beta}_0+\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{X}^t\boldsymbol{y}+\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{V}\boldsymbol{\beta}_0 \\
&= \boldsymbol{y}^t\boldsymbol{X}A^{-1}\boldsymbol{X}^t\boldsymbol{y}+2\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{X}^t\boldsymbol{y}+\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{V}\boldsymbol{\beta}_0,
\end{aligned}
$$

then

$$
\begin{aligned}
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\boldsymbol{y}+\boldsymbol{\beta_0}^t\boldsymbol{V}\boldsymbol{\beta_0}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}+2\beta}{2}\right)^{-\alpha-\frac{n}{2}} \\
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\boldsymbol{y}-\boldsymbol{b}^tA^{-1}\boldsymbol{b}+\boldsymbol{\beta_0}^t\boldsymbol{V}\boldsymbol{\beta_0}+2\beta}{2}\right)^{-\alpha-\frac{n}{2}} \\
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\boldsymbol{y}-\boldsymbol{y}^t\boldsymbol{X}A^{-1}\boldsymbol{X}^t\boldsymbol{y}-2\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{X}^t\boldsymbol{y}-\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{V}\boldsymbol{\beta}_0+\boldsymbol{\beta_0}^t\boldsymbol{V}\boldsymbol{\beta_0}+2\beta}{2}\right)^{-\alpha-\frac{n}{2}} \\
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\left(I-\boldsymbol{X}A^{-1}\boldsymbol{X}^t\right)\boldsymbol{y}-2\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{X}^t\boldsymbol{y}-\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{V}\boldsymbol{\beta}_0+\boldsymbol{\beta_0}^t\boldsymbol{V}\boldsymbol{\beta_0}+2\beta}{2}\right)^{-\alpha-\frac{n}{2}}
\end{aligned}
$$

$E=-\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{V}\boldsymbol{\beta}_0+\boldsymbol{\beta_0}^t\boldsymbol{V}\boldsymbol{\beta_0}+2\beta=\boldsymbol{\beta_0}^t\left(I-\boldsymbol{V}A^{-1}\right)\boldsymbol{V}\boldsymbol{\beta_0}+2\beta$

hence

$$
\begin{aligned}
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\left(I-\boldsymbol{X}A^{-1}\boldsymbol{X}^t\right)\boldsymbol{y}-2\boldsymbol{\beta}_0^t\boldsymbol{V}A^{-1}\boldsymbol{X}^t\boldsymbol{y}+E}{2}\right)^{-\alpha-\frac{n}{2}} \\
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\boldsymbol{F}\boldsymbol{y}-2\boldsymbol{h}^t\boldsymbol{y}+E}{2}\right)^{-\alpha-\frac{n}{2}}.
\end{aligned}
$$

with
$\boldsymbol{F} = \left(\boldsymbol{I} - \boldsymbol{X}A^{-1}\boldsymbol{X}^t\right)$ and $\boldsymbol{h}^t = \boldsymbol{\beta}_0^t \boldsymbol{V}A^{-1}\boldsymbol{X}^t$ then $\boldsymbol{h} = \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{V}\boldsymbol{\beta_0}$

$$
\begin{aligned}
p(\boldsymbol{y}) &\propto \left(\frac{\boldsymbol{y}^t\boldsymbol{F}\boldsymbol{y} - 2\boldsymbol{h}^t\boldsymbol{y} + E}{2}\right)^{-\alpha-\frac{n}{2}} \\
&\propto \left(\frac{\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right)^t \boldsymbol{F}\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right) + E - \boldsymbol{h}^t\boldsymbol{F}^{-1}\boldsymbol{h}}{2}\right)^{-\alpha-\frac{n}{2}} \\
&\propto \left(\frac{\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right)^t \boldsymbol{F}\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right) + E - \boldsymbol{h}^{\boldsymbol{t}}\boldsymbol{F}^{-1}\boldsymbol{h}}{2}\right)^{-\frac{2\alpha+n}{2}} \\
&\propto \left(\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right)^t \boldsymbol{F}\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right) + G\right)^{-\frac{2\alpha+n}{2}} \\
&\propto \left(\frac{1}{2\alpha}\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right)^t \frac{2\alpha\boldsymbol{F}}{G}\left(\boldsymbol{y} - \boldsymbol{F}^{-1}\boldsymbol{h}\right) + 1\right)^{-\frac{2\alpha+n}{2}}
\end{aligned}
$$

with $G = E - \boldsymbol{h}^{\boldsymbol{t}}\boldsymbol{F}^{-1}\boldsymbol{h}$
Hence, $\boldsymbol{Y} \sim t_\nu\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with $\nu = 2\alpha$, $\boldsymbol{\mu} = \boldsymbol{F}^{-1}\boldsymbol{h}$ and $\boldsymbol{\Sigma^{-1}} = \frac{2\alpha\boldsymbol{F}}{G}$

Bibliography

Barbieri, M. M. and Berger, J. O. (2002). Optimal Predictive Model Selection. *Ann. Statist*, 32:870–897. 91, 98, 100

Barry, D. and Hartigan, J. A. (1992). Product Partition Models for Change Point Problems. *The Annals of Statistics*, 20(1):260–279. 2, 6, 8, 9, 17, 18, 23, 37, 48

Barry, D. and Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, 88(421):309–319. 5, 6, 8, 17, 18, 37, 40, 41, 48

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, 1 edition. 31, 113

Bethlehem, J. G. and Keller, W. J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics*, 3(2):141–153. 73, 76

Binder, D. A. (1978). Bayesian Cluster Analysis. *Biometrika*, 65(1):31–38. 1, 86

Binder, D. A. (1981). Approximations to Bayesian Clustering Rules. *Biometrika*, 68(1):275–285. 1

Binder, D. A. (1982). Non-Parametric Bayesian Models for Samples from Finite Populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):388–393. 71

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, first edition. 112

BIBLIOGRAPHY

Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1(2):353–355. 11, 13, 70

Blasi, P. D., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:212–229. 17

Blei, D. M. and Jordan, M. I. (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144. 112

Bormetti, G., De Giuli, M. E., Delpini, D., and Tarantola, C. (2012). Bayesian Value-at-Risk with product partition models. *Quantitative Finance*, 12(5):769–780. 2, 6, 8, 26

Breidt, F. J. and Opsomer, J. D. (2000). Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics*, 28(4):1026–1053. 76

Brodsky, E. and Darkhovsky, B. S. (2010). *Non-Parametric Statistical Diagnosis: Problems and Methods.* Springer, first edition. 37

Campirán García, E. and Gutiérrez-Peña, E. (2018). Nonparametric product partition models for multiple change-points analysis. *Communications in Statistics - Simulation and Computation*, 0(0):1–26. 2

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. *Journal of the American Statistical Association*, 88(421):268–277. 77

Chen, J. and Gupta, A. K. (2011). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance.* Birkhäuser Boston, second edition. 36, 37, 60

Ciampi, A., Rich, B., Dyachenko, A., Antoniano, I., Murie, C., and Nadon, R. (2007). Locally Linear Regression and the Calibration Problem for Micro-Array Analysis. In *Selected Contributions in Data Analysis and Classification*, pages 549–555. Springer. 79

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons, 3rd edition. 68, 71

Crowley, E. M. (1997). Product Partition Models for Normal Means. *Journal of the American Statistical Association*, 92(437):192–198. 6, 8

Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis.* Wiley, first edition. 37

Dahl, D. B. (2009). Modal Clustering in a Class of Product Partition Models. *Bayesian Analysis*, 4(2):243–264. 6, 8

Dalenius, T. and Hodges, J. L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, 54(285):88–101. 71

de Finetti, B. (1972). *Probability, Induction and Statistics (Probability & Mathematical Statistics).* John Wiley & Sons Ltd. 25

De Giuli, M. E., Maggi, M. A., and Tarantola, C. (2010). Bayesian outlier detection in Capital Asset Pricing Model. *Statistical Modelling*, 10(4):375–390. 6, 26

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Wiley, 1 edition. 71

Deville, J. C. and Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382. 75, 76

Ecley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models.* Cambridge University Press. 37

Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 31(2):195–233. 70

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2). 8

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230. 2, 5, 10, 27, 70

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2010). A Probability for Classification Based on the Dirichlet Process Mixture Model. *Journal of Classification*, 27(3):389–403. 42

Gutiérrez-Peña, E. and Walker, S. G. (2001). A Bayesian predictive approach to model selection. *Journal of Statistical Planning and Inference*, 93(1):259–276. 108

Gutiérrez-Peña, E. and Walker, S. G. (2005). Statistical Decision Problems and Bayesian Nonparametric Methods. *International Statistical Review*, 73(3):309–330. 108

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110:435–448. 1, 3, 66, 91, 99, 100

Hartigan, J. A. (1990). Partition models. *Communications in Statistics (Theory and Methods)*, 19(8):2745–2756. 5, 8, 23

Hegarty, A. and Barry, D. (2008). Bayesian disease mapping using product partition models. *Statistics in Medicine*, 27(19):3868–3893. 8, 81

Holmes, C. C., Denison, D. G. T., Ray, S., and Mallick, B. K. (2005). Bayesian prediction via partitioning. *Journal of Computational and Graphical Statistics*, 14(4):811–830. 81

Hsu, D. A. (1979). Detecting Shifts of Parameter in Gamma Sequences with Applications to Stock Price and Air Traffic Flow Analysis. *Journal of the American Statistical Association*, 74(365):31–40. 57

Hupe, P. (2011). *GLAD: Gain and Loss Analysis of DNA*. R package version 2.20.0. 61

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating Mixtures of Regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79. 1

Jordan, C., Livingstone, V., and Barry, D. (2007). Statistical modelling using product partition models. *Statistical Modelling*, 7(3):275–295. 8, 81

Kehagias, A., Nicolaou, A., Petridis, V., and Fragkou, P. (2004). Text Segmentation by Product Partition Models and Dynamic Programming. *Mathematical and Computer Modeling*, 39:209–217. 6, 9

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500):1590–1598. 1, 37, 48

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, 220(4598):671–680. 112

Kotz, S. and Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press, Cambridge, New York, Madrid. 118

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. *Proceeding of the section on survey researh methods. American Statistical Association*, pages 280–285. 76

Lau, J. W. and Green, P. J. (2007). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics*, 16:526–558. 1

Laud, P. W. and Ibrahim, J. G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57:247–262. 99

Leon-Novelo, L. G., Bekele, B. N., Müller, P., Quintana, F. A., and Wathen, K. (2012). Borrowing Strength with Nonexchangeable Priors over Subpopulations. *Biometrics*, 68(2):550–558. 6

Little, R. J. A. (1993). Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association*, 88(423):1001–1012. 72, 73

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556. 68, 69, 70

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12(1). 13

Lo, A. Y. (1986). Bayesian Statistical Inference for Sampling a Finite Population. *The Annals of Statistics*, 14(3):1226–1233. 71

Loschi, R. H. (2002). An analysis of the influence of some prior specifications in the identification of change points via product partition model. *Computational Statistics and Data Analysis*, 39(4):477–501. 9

Loschi, R. H. and Cruz, F. R. B. (2005). Extension to the product partition model: computing the probability of a change. *Computational Statistics & Data Analysis*, 48(2):255–268. 2, 9, 17, 18, 37

Loschi, R. H., Cruz, F. R. B., Iglesias, P. L., and Arellano-Valle, R. (2003). A Gibbs sampling scheme to the product partition model: an application to change-point problems. *Computers & Operations Research*, 30(3):463–482. 9, 15, 48

Loschi, R. H., Pontel, J. G., and Cruz, F. R. B. (2010). Multiple change-point analysis for linear regression models. *Chilean Journal of Statistics*, 1(2):93–112. 6, 9

Martínez, A. F. and Mena, R. H. (2014). On a Nonparametric Change Point Detection Model in Markovian Regimes. *Bayesian Analysis*, 9(4):823–858. 42

Matteson, D. S. and James, N. A. (2014). A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, 109(505):334–345. 48

Mena, R. H. and Ruggiero, M. (2016). Dynamic density estimation with diffusive Dirichlet mixtures. *Bernoulli*, 22(2):901–926. 66

Mira, A. and Petrone, S. (1996). Bayesian hierarchical non-parametric inference for change-point problems. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 693–703. Oxford University Press, USA. 37

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric Methods in Survey Sampling. In *New Developments in Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, chapter 24, pages 203–210. 77

Monteiro, J. V. D., Assunção, R. M., and Loschi, R. H. (2011). Product partition models with correlated parameters. *Bayesian Analysis*, 6(4):691–726. 9, 17

Müller, P. and Nieto-Barajas, L. (2008). The nested dirichlet process: Commentary. *Journal of the American Statistical Association*, 103(483):1146–1147. 28, 30

Müller, P. and Quintana, F. A. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808. 6, 9, 81, 112

Müller, P., Quintana, F. A., and Rosner, G. L. (2011). A Product Partition Model With Regression on Covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278. 6, 9, 81, 112

Nelson, D. and Meeden, G. (1998). Pólya Posterior Quantile Estimation for Stratified Populations. Technical report. 79

Park, J. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226. 9, 112

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In Al, E., editor, *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30 of *Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, California. 7

Puelz, D., Hahn, P. R., and Carvalho, C. M. (2016). Variable Selection in Seemingly Unrelated Regressions with Random Predictors. 1

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136(8):2407–2429. 5, 7, 8

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574. 2, 3, 5, 9, 12, 18, 23, 26, 31, 89, 95, 109, 113

Quintana, F. A., Iglesias, P. L., and Bolfarine, H. (2005a). Bayesian Identification Of Outliers And Change-Points In Measurement Error Models. *Advances in Complex Systems*, 8(04):433–449. 2, 17, 18, 26, 37

Quintana, F. A., Iglesias, P. L., and Galea-Rojas, M. (2005b). Bayesian robust estimation of systematic risk using product partition models. *Applied Financial Economics Letters*, 1(5):313–320. 6, 26

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 3, 111

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850. 48

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The Nested Dirichlet Process. *Journal of the American Statistical Association*, 103(483):1131–1154. 2, 11, 26, 27, 29, 30

Rueda, M. and Sánchez-Borrego, I. (2009). A predictive estimator of finite population mean using nonparametric regression. *Computational Statistics*, 24(1):1–14. 76

Rueda, M., Sánchez-Borrego, I., Arcos, A., and Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71(1):33–44. 78

Schmid, F., Schmidt, R., Blumentritt, T., Gaiber, S., and Ruppert, M. (2010). *Copula-Based Measures of Multivariate Association*, volume 198 of *Lecture Notes in Statistics*, pages 209–236. Springer Berlin Heidelberg. 105

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650. 10

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature genetics*, 29(3):263–264. 61

Tarantola, C., Consonni, G., and Dellaportas, P. (2008). Bayesian clustering for row effects models. *Journal of Statistical Planning and Inference*, 138(7):2223–2235. 5, 6, 9

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525. 61

Willan, A. R. and Watts, D. G. (1978). Meaningful Multicollinearity Measures. *Technometrics*, 20(4):407+. 104

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, pages 1434–1447. 17

Yau, C. and Holmes, C. C. (2013). A decision-theoretic approach for segmental classification. *The Annals of Applied Statistics*, 7(3):1814–1835. 1, 19, 37

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57. 37, 66

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, 7(2):1–38. 57

Zheng, H. and Little, R. J. A. (2003). Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples. *Journal of Official Statistics*, 19(2):99–107. 79

Zheng, H. and Little, R. J. A. (2004). Penalized Spline Nonparametric Mixed Models for Inference About a Finite Population Mean from Two-Stage Samples. *Survey Methodology*, 30:209–218. 79