Universidad Nacional Autónoma de México

Programa de Maestría y Doctorado en Ciencias Matemáticas
y de la Especialización en Estadística Aplicada

Usages of random combinatorial structures in Statistics; a Bayesian nonparametric approach

Tesis

que para optar por el grado de

Doctor en Ciencias

presenta

Asael Fabian Martínez Martínez

tutor principal

Dr. Ramsés H. Mena Chávez, IIMAS-UNAM

comité tutor

Dr. Raúl Rueda Díaz del Campo, IIMAS-UNAM
Dr. Gerónimo Uribe Bravo, IMATE-UNAM

México, D.F., Octubre de 2015

ABSTRACT

The present thesis is about statistical clustering studied under a Bayesian nonparametric approach. I started with the general clustering problem, exploring the underlying space where it takes place, and identified two particular situations which arise when the grouping conditions change. Therefore, I concentrated on the study of some theoretical properties of these situations, under a combinatorial and a statistical framework, as well as on their practical application in Statistics.

The specific situations I studied can be encoded by different combinatorial classes: set partitions, segmentations and ordered set partitions. After providing a brief account of their properties and studying existent and novel probabilistic models defined over them, I dedicated a space to propose statistical models.

The class of set partitions is adequate for the general clustering problem, and there is a vast quantity of literature about it. In particular, I restricted my study to the theory of random partitions and their relation with mixture models. While random partitions are suitable to model discrete observations, mixture models are a natural extension for the continuous case. Regarding this, I explored a recent mixture model and generalized it. The results constitute an appealing alternative to standard nonparametric models.

For the class of segmentations, I found that it appears when the phenomena under study evolves in such a way that the observed data are indexed by some ordered set. Therefore, the admissible groupings are those formed by consecutive observations with respect to such a set. Given this and based on the random partition theory, I developed a novel model for change-point detection.

Finally, I studied the relationship between the class of ordered set partitions and the label-switching problem. This problem appears frequently in sampling schemes based on mixture models, and there are many methodologies dealing with it. Thus, I presented a proposal based on this class.

RESUMEN

La presente tesis versa sobre el problema de clustering estudiado bajo un enfoque Bayesiano no paramétrico. Comenzando con este problema general y explorando el espacio donde tiene lugar, dos situaciones específicas fueron identificadas, que surgen cuando las reglas de agrupamiento cambian. Así, me concentré en el estudio de algunas de sus propiedades teóricas, bajo un enfoque combinatorio y estadístico, y en su aplicación práctica en Estadística.

Dichas situaciones de clustering se pueden codificar como clases combinatorias: *set partitions*, *segmentations* y *ordered set partitions*. Luego de dar una breve exposición de sus propiedades y estudiar existentes y nuevos modelos probabilísticos, dediqué un espacio para proponer modelos estadísticos.

La clase *set partitions* es adecuada para el problema general de clustering y existe una gran cantidad de literatura al respecto. En particular, me concentré en la teoría de particiones aleatorias y su relación con modelos de mezclas. Mientras que las particiones aleatorias modelan observaciones discretas, los modelos de mezclas son una extensión natural para el caso continuo. A este respecto, exploré un modelo de mezclas reciente y lo generalicé. Los resultados representan una alternativa a los modelos no paramétricos estándar.

Sobre la clase *segmentations*, encontré que aparece cuando el fenómeno bajo estudio evoluciona de tal forma que los datos observados son indexados por un conjunto ordenado. Por tanto, los agrupamientos permitidos son aquellos que contienen observaciones consecutivas con respecto a tal conjunto. Dado esto y basado en la teoría de particiones aleatorias, desarrollé un novedoso modelo para detección de puntos de cambio.

Finalmente, estudié la relación entre la clase *ordered set partitions* y el problema de *label-switching*. Este problema aparece frecuentemente en esquemas de simulación basados en modelos de mezclas, y existen muchas técnicas para resolverlo. Así, presento una propuesta basada en esta clase.

To Manuel, Mercedes
Haziel, Elianai, Eliel
Matías & Nicolás

ACKNOWLEDGEMENTS

I would like to thank to everyone who helped me—one way or another—to make this achievement possible:

To my parents, Manuel and Mercedes, for their unconditional support

To Ramsés, for introducing me into this fascinating world of research

To the Postgraduate Program of Mathematics and to CONACyT, project 241195, for their support

Last—but always being the first one—thanks Abba, for creating all the things

כה אמר ה' נתן שמש לאור יומם
חקת ירח וכוכבים לאור לילה
רגע הים ויהמו גליו ה' צבאות שמו

...אם ימדו שמים מלמעלה ויחקרו מוסדי ארץ למטה
גם אני אמאס בכל זרע ישראל על כל אשר עשו נאם ה'

ירמיהו, לא"לד,לו

CONTENTS

INTRODUCTION

This thesis studies statistical cluster modeling under a Bayesian nonparametric approach. We aim to understand the underlying space where clustering takes place, as well as to identify particular cluster situations—arising when the grouping conditions change—in order to provide adequate probabilistic models dealing with the resulting problems.

Clustering is a problem where the principal objective is to group a collection of objects into subsets—or clusters—in such a way that objects within each cluster are more similar to one another than those assigned to different clusters (Hastie et al. 2009). A clustering model is completely specified after providing some method to measure such similarity. Broadly speaking, these models can be categorized according to how similarity is modeled: (i) *unsupervised learning*—term used in Machine Learning—where the main rationale for grouping is based on distance metrics and algorithmic approaches, and (ii) *statistical clustering* which makes use of stochastic methods to infer about the data grouping structure. The counterpart of unsupervised learning is called *supervised learning*, and the main difference is that the number of clusters is known in the latter, either they are previously fixed or are obtained after a training procedure. The analogous term for supervised learning under the statistical perspective is known as *discriminant analysis* or *classification*. The importance of this is that—historically—classification and statistical clustering were developed almost at the same time and, in some cases, referring to the same problem.

Early statistical works on these topics date back to Pearson (1894) and Fisher (1936). Fisher focused on the classification of observations into one of two groups through the usage of a *discriminant function*. Based on Fisher's work, we can find, for example, the approaches of Welch (1939), von Mises (1944) and Rao (1947) who defined different forms to achieve correct classifications. With regard to clustering, Zubin (1938) and Tryon (1939) can be considered the pioneers on this area. A more complete historical review is provided, for example, by Hershberger (2005).

Under a probabilistic perspective, clustering methodologies have been extensively

studied by means of special random objects, called *random partitions*. First studies—motivated by applications in population genetics—can be traced back to Ewens' paper on sampling theory of neutral alleles (Ewens 1972), and—in a more formal way—to Kingman's works on partition structures (Kingman 1978*a*,*b*). Following this approach, special attention has been placed on diverse representations of such structures, bringing processes of fragmentation and coalescence (Kingman 1982) and models for species sampling problems (Pitman 1996), among others. Random partitions are also important in Bayesian Nonparametric Statistics. Theoretically, random partition distributions can be obtained via de Finetti's representation theorem, so this provides a natural way to address the problem of clustering under a full Bayesian framework; see, for example, Hjort et al. (2010) for an up-to-date survey about this connection and some applications. Hence, we take as a starting point the random partition theory for the different problems we study along this work.

The fact that we confine all the work presented here to a Bayesian perspective entails some challenges. It is well known that—in order to perform any kind of analysis—every Bayesian model requires to assign a *prior* probability distribution for all the unknown elements. For the case of clustering models, the principal uncertainty is about the grouping structure. In other words, we need to define probability distributions over spaces which describe all the possible ways to put together any set of objects given certain clustering rules; it is important, thus, knowing well such spaces. We start studying the clustering problem modeled by random partitions, and afterwards, we modify its clustering rules in order to analyze the implications of these changes and investigate the possible applications of the resulting grouping structure. With the purpose of illustrate the different clustering scenarios we are dealing with throughout this thesis, we make use of some urn models next.

The first scenario corresponds to the one modeled by random partitions. Using an urn model and fixing the number of urns, it consists on—given a set of n labeled balls and m unlabeled urns—thrown these balls into the different urns; we cannot leave empty urns. A ball label is formed by two parts: an identifier and some information, i.e. a sampled value. The clustering from this scheme is obtained by the urn contents—in terms of the balls information—after ordering the urns according to their least ball identifier. Since the number of urns is not known *a priori*, the space should contain all the groupings when the number of urns ranges from 1 to n.

Based on this urn scheme, our next scenario uses unlabeled balls. The rest of the procedure remains unchanged. In this case, the resulting clustering is obtained by presenting

the urn contents and labeling the balls left to right from 1 to n. This scheme resembles Feller's model known as *stars and bars* (Feller 1968).

Our last scenario is represented by the first urn model but using identified urns. Identifying urns changes how the clustering structure is presented, since it is done precisely according to the urn identifier. Following this last scheme, there are some approaches resulting by allowing empty urns, however, we will not pay much attention to them.

These three scenarios constitute our central object of study, and—in order to give a mathematical treatment of all of them—we present in Chapter I the combinatorial classes encoding each of such scenarios. As already said, we start with random-partition-based clustering problems, mainly because their popularity in applications has allowed to have many literature on the topic. So we provide—in addition to some combinatorial characteristics of this problem—a survey on various of the methodologies used to define probability distributions for random partitions. Afterwards, in this same chapter, we settle the combinatorial classes and some of their characteristics for the second and third scenarios. Further chapters deal with each scenario in a more detailed form.

The Chapter II is devoted to the study of segmentation models, arising from the second scenario. In particular, we focus on a proposal dealing with change-point detection problems. In this part, we perform a full Bayesian analysis providing different constructions for probability distributions—specific for this cases—as well as a Markov chain Monte Carlo sampling scheme. A simulation study is performed together with an application using a real dataset. Most of the work included in this chapter can be also found in Martínez & Mena (2014). We conclude this chapter giving some insights about possible future works on the topic.

Turning back to random partitions, in Chapter III, we focus on the relationship between mixture models and clustering problems. Indeed, mixture models are natural candidates for clustering under a Bayesian nonparametric approach due to their construction. Thus, we will explain how cluster modeling takes place under these models. Moreover, this relationship broadens our panorama in many directions. On the one hand, mixture models can be considered as an extension of random partitions in the sense that they allow to cluster continuous observations. Then, by modifying some features of the mixture model, we could obtain new methods or approaches applicable to random partition models. On the other hand, mixture modeling theory has its own issues—like the *label-switching* problem, therefore, some of them can be handled if we focus on the underlying clustering structure of these models and work with it. In this same chapter, we put special attention to this first direction exploring particular cases of mixture models

with decreasing weights.

Finally, in Chapter IV, we continue studying mixture models. The first part presents some applications of our third clustering scenario, in particular, we study the problem of how observations can be allocated into mixture components. This problem is useful since it leads us to the central topic of the chapter, which is the second direction explained in the previous paragraph. Regarding this topic, we study why label-switching appears in mixture models and provide some ideas about how to deal with it.

We conclude the thesis with some final remarks about our current work on each of these topics, as well as possible future work and some related extensions. We also present, in the appendices, specific simulation schemes used in different chapters.

CHAPTER I

COMBINATORIAL STRUCTURES IN CLUSTER MODELING

In this first chapter, we introduce different combinatorial structures appearing in cluster modeling. With the help of different clustering scenarios, we illustrate how the random object of interest is affected as the grouping conditions change.

We start with, perhaps, the simplest scenario where a series of items is clustered into nonempty groups. All the possibilities for grouping them are encoded by the combinatorial class called *set partitions*, which is very well known in different probabilistic and statistical models since it allows us to define *random partitions*. Throughout Section 1—in addition to provide some properties of set partitions—we survey the theory of random partitions which has become a main tool in Bayesian nonparametric models.

In the rest of the chapter, we present other clustering scenarios where set partitions are not good candidates to encode the problem of interest. Section 2 gives an example where observed data are indexed by an ordered set and we want to cluster them in such a way that only consecutive indices can be put together. The possible groupings are degenerated—if we compare them with set partitions—because of this additional constrain. We call *segmentations* to the combinatorial class encoding this scenario. Section 3 presents a third clustering scenario. The first part studies a similar scenario that the one of set partitions, but—in this case—the groups are labeled or identified somehow. As we can imagine, the possible groupings increase; they are encoded by the combinatorial class known as *ordered set partitions*. This scenario is often stated slightly different, resulting actually a classification problem, but of interest because of their relationship with ordered set partitions. The last part of this section studies the combinatorial class called *words* and such a relationship. Hence, the material presented in this chapter serves as motivation as well as theoretical background for the rest of the thesis.

In order to derive the expressions regarding combinatorial classes—specification, generating function and counting sequence, among others—we make use of the *symbolic method*, a combinatorial counting technique developed by Flajolet & Sedgewick (2009). To establish some notation, we denote by $\mathcal{C}$ any combinatorial class, its generating function is a formal power series $z \mapsto C(z)$, and its counting sequence corresponds to a sequence of nonnegative integers $(C_1, \dots, C_n, \dots)$. Note that a combinatorial class is a generic description, but if we want to indicate the set of items X we are working with to obtain the class $\mathcal{C}$, it will be denoted by $\mathcal{C}_X$. A further reference about this topic can be found in the book by Flajolet & Sedgewick (2009).

Moreover, since combinatorial classes arise in many different—and apparently unrelated—contexts, there exists an *On-line Encyclopedia of Integer Sequences* collecting diverse information about them. Therefore—in the cases where there is an entry on this encyclopedia for the class under study—we indicate it using the label EIS followed by its id number. An on-line version of this encyclopedia is found in `oeis.org`.

§1 Set partitions

The combinatorial class *set partitions* has become an attractive topic of study in Mathematics and other sciences, for example, Combinatorics (Mansour 2013), Population Genetics (Lijoi et al. 2007*a*), Economy (Aoki 2003), Probability (Berestycki 2009) and Statistics (Hjort et al. 2010). We confine ourselves to their role in Probability and Statistics, focusing on the study of *random partitions* and their applications in Bayesian clustering models. In order to do it, we first illustrate the kind of problems where there appear.

Clustering scenario 1. Suppose there is a finite number of items—each one uniquely identified, for example, though a label—and we want to gather them in groups, called *blocks*, in such a way that each item belongs to only one block. There cannot be empty blocks.

For the case of having three items, $\{x_1, x_2, x_3\}$, all the possible groupings satisfying the above specification are

$$\{\{x_1, x_2, x_3\}\},\ \{\{x_1\}, \{x_2, x_3\}\},\ \{\{x_1, x_2\}, \{x_3\}\},\ \{\{x_1, x_3\}, \{x_2\}\},\ \{\{x_1\}, \{x_2\}, \{x_3\}\}.$$

Thus, for example, $\{\{x_1, x_2, x_3\}\}$ indicates that all the observations belong to the same single block and $\{\{x_1\}, \{x_2\}, \{x_3\}\}$ that each one belongs to its own block.

Notice that this scenario has an implicit assumption: there is no order restriction

in the items, neither within groups nor among them, so—from the example above—groupings like $\{\{x_1, x_2\}, \{x_3\}\}$, $\{\{x_2, x_1\}, \{x_3\}\}$, $\{\{x_3\}, \{x_1, x_2\}\}$ and $\{\{x_3\}, \{x_2, x_1\}\}$ are treated as the same and only one of them is used. Therefore, this kind of grouping structure can be considered as the simplest problem in cluster analysis.

In Combinatorics, the combinatorial class which encodes the grouping structure under this scenario is known as *set partitions*. Using the symbolic method, it is constructed through a labeled combinatorial class as follows. Assuming, first, that the number of blocks is fixed; all the possible ways of grouping any given set of items in exactly k blocks are encoded by the class

$$\mathcal{P}^{(k)} = \mathbf{Set}_k(\mathbf{Set}_{\geq 1}(\mathcal{Z})), \tag{1}$$

for $\mathcal{Z}$ an atomic class. Hence, set partitions—denoted by $\mathcal{P}$—are obtained by putting together all the elements of $\mathcal{P}^{(k)}$ as k varies, that is

$$\mathcal{P} = \bigcup_{k\geq 0} \mathcal{P}^{(k)} = \mathbf{Set}(\mathbf{Set}_{\geq 1}(\mathcal{Z})).$$

Each element of $\mathcal{P}$ is called a *partition* since its blocks are a partition of the whole set of items X, that is, for any $\{\pi_1, \dots, \pi_k\} \in \mathcal{P}_X$, three conditions hold: (i) $\pi_j \neq \varnothing$ for all $j = 1, \dots, k$, (ii) $\pi_i \cap \pi_j = \varnothing$ for all $i \neq j$, and (iii) $\pi_1 \cup \cdots \cup \pi_k = X$.

From the above specification, the counting sequence of $\mathcal{P}$ is obtained from its *exponential generating function* (EGF), which is given by

$$P(z) = e^{e^z - 1} = \frac{1}{e} \sum_{j=0}^{\infty} \frac{e^{jz}}{j!},$$

(formula first deduced by Dobinski 1877) and its nth coefficient is, therefore,

$$P_n = n! \, [z^n] P(z) = \frac{1}{e} \sum_{j=0}^{\infty} \frac{j^n}{j!}.$$

The resulting counting sequence $(P_0, P_1, \dots)$ is known as *Bell numbers* (EIS A000110)—usually denoted by $(B_n)_{n\geq 0}$ instead of $(P_n)_{n\geq 0}$—and it satisfies the recurrence relation

$$P_{n+1} = \sum_{k=0}^{n} \binom{n}{k} P_k,$$

with $P_0 = 1$. A list of the first Bell numbers is shown in Table 1. These numbers are named

	k											
n	0	1	2	3	4	5	6	7	8	9	10	B_n
0	1											1
1	0	1										1
2	0	1	1									2
3	0	1	3	1								5
4	0	1	7	6	1							15
5	0	1	15	25	10	1						52
6	0	1	31	90	65	15	1					203
7	0	1	63	301	350	140	21	1				877
8	0	1	127	966	1 701	1 050	266	28	1			4 140
9	0	1	255	3 025	7 770	6 951	2 646	462	36	1		21 147
10	0	1	511	9 330	34 105	42 525	22 827	5 880	750	45	1	115 975

TABLE 1: (First columns) Stirling numbers of the second kind, $S(n,k)$, displayed in a triangular array for $0 \le k \le n \le 10$; note that $S(n,k) = 0$ for all $k > n$. (Last column) List of the first Bell numbers, B_n, for $n = 0, \ldots, 10$.

after Eric Temple Bell because of his studies on Bell polynomials—where he called them *exponential numbers*—but, actually, they appear in previous papers (e.g. Dobinski 1877). On the other hand, an asymptotic form for Bell numbers is given by

$$P_n \sim \frac{n!\, e^{e^r-1}}{r^n \sqrt{2\pi r(r+1)e^r}},$$

where r is the positive root of the equation $re^r = n+1$.

The subsets $\mathcal{P}^{(k)}$ of $\mathcal{P}$, $k = 0, 1, 2, \ldots$, are also of interest. From the specification given in Equation (1), the EGF of $\mathcal{P}^{(k)}$ is given by

$$P^{(k)}(z) = \frac{1}{k!}\left(e^z - 1\right)^k,$$

from where—using the binomial theorem—its nth-coefficient is

$$P_n^{(k)} = n!\,[z^n]\,\frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}(-1)^k e^{(k-j)z} = \frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}(-1)^k(k-j)^n.$$

For a fixed k, they have the asymptotic form $P_n^{(k)} \sim k^n/k!$. The counting sequence $(P_n^{(k)})_{n\ge 0}$ is known as *Stirling numbers of the second kind* (EIS A008277), usually denoted as $S(n,k)$. These numbers were introduced by Stirling (1717) and became an important tool in the calculus of finite differences and in probability theory. Table 1 shows the Stirling numbers

for different values of n. Using them, Bell numbers can be computed easily since $B_n = \sum_{k=0}^{n} S(n,k)$.

Although set partitions can be obtained from any set of items, in most of the literature, probabilistic and statistical studies of set partitions deal only with the set partitions of the set $[n] := \{1, \dots, n\}$, denoted by $\mathcal{P}_{[n]}$. This is done only to simplify notation. The set partitions of $[n]$ are exactly the same—in quantity and form—as the set partition of any other arbitrary set of n items. In other words, both classes encode the same clustering structure. On the other hand—as already said—set partitions do not impose any restriction in the order among the blocks neither among the elements within each block; for example, the partitions $\{\{1,2\},\{3\}\}$, $\{\{2,1\},\{3\}\}$, $\{\{3\},\{1,2\}\}$ and $\{\{3\},\{2,1\}\}$ could be thought as different elements of $\mathcal{P}_{[3]}$, but they are considered as the same. Thus, as a convention, partitions will be listed *in order of appearance*, which means that, for any partition $\{\pi_1, \dots, \pi_k\}$, the elements of each block π_j will be listed in increasing order, and the blocks will be listed in such a way that $\pi_{1,1} < \pi_{2,1} < \cdots < \pi_{k,1}$ for $\pi_{j,1} = \min \pi_j$. Thus, $\{\{1,2\},\{3\}\}$ is the only element of $\mathcal{P}_{[3]}$ that will be used from all the examples given above.

Having said that, $\mathcal{P}_{[n]}$ corresponds to the induced support for the clustering scenario described at the beginning of the section, and any related probabilistic model or statistical inference should be based on it.

1.1 Probability distributions over set partitions

Once set partitions have been introduced and their role in clustering has been established, we are able to define probability models based on them.

Definition 1. A *random partition* Π_n is a $\mathcal{P}_{[n]}$-valued random variable.

Before studying probability distributions for random partitions, notice that a random variable implicitly defined in random partitions, denoted hereafter by K_n, is the one modeling the number of blocks. This is due to the fact that—under the framework assumed in the whole thesis—when we ask for some probability $\mathbf{P}(\Pi_n = \pi)$, for $\pi \in \mathcal{P}_{[n]}$, we are actually asking for the probability of the event $[K_n = k] \cap [\Pi_n = \pi]$, since any partition π is formed by k groups for some positive $k \leq n$. Hence, the marginal distribution of K_n is obtained by marginalizing Π_n as follows

$$\mathbf{P}(K_n = k) = \sum_{\pi \in \mathcal{P}_{[n]}^{(k)}} \mathbf{P}(\Pi_n = \pi), \quad k = 1, 2, \dots, n,$$

where $\mathcal{P}_{[n]}^{(k)}$ is—as stated before—the set partitions of $[n]$ having exactly k blocks. From this fact, we can say that the random-partition approach for clustering modeling has an advantage over other approaches which requires a fixed number of blocks in order to perform clustering.

Additionally, the random variable K_n is very useful in the study of random partitions. The set $\mathcal{P}_{[n]}$ is not an ordered set, so it is sometimes difficult to have a clear understanding of how some distribution for Π_n behaves. Some light is shed by analyzing the behavior of K_n, for example, its probability distribution or its asymptotic results.

In the following sections, some approaches to define probability distributions over $\mathcal{P}_{[n]}$ are discussed. Special attention is placed in distributions known as *exchangeable partition probability functions*, therefore, after presenting the general approaches, different families of this kind of distributions are explained.

1.1.1 Distributions based on exchangeable sequences of random variables

A first approach to construct probability distributions for random partitions is based on sequences of exchangeable random variables. Let $(X_n)_{n\geq 1}$ be an infinite sequence of random variables taking values in some complete and separable space $\mathbb{X}$ endowed with the Borel σ-algebra $\mathcal{X}$. It is said that $(X_n)_{n\geq 1}$ is *exchangeable* if, for all $n \geq 1$, the distribution of $(X_1, X_2, \dots, X_n)$ is the same of $(X_{\rho(1)}, X_{\rho(2)}, \dots, X_{\rho(n)})$ for every finite permutation ρ of $[n]$. Due to de Finetti's representation theorem (de Finetti 1937), the assumption of exchangeability is equivalent to assume the existence of a probability distribution Q on $\mathcal{M}(\mathbb{X})$—the space of probability measures on $\mathbb{X}$—such that, for every $n \geq 1$ and for all $A_i \in \mathcal{X}$, $i = 1, \dots, n$,

$$\mathbf{P}[X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n] = \int_{\mathcal{M}(\mathbb{X})} \prod_{i=1}^{n} \tilde{p}(A_i) Q(\mathrm{d}\tilde{p}). \tag{2}$$

The distribution Q is known as the *de Finetti measure* of the sequence $(X_n)_{n\geq 1}$, and it represents the prior distribution for the random probability measure (RPM) $\tilde{p}$. This justifies Bayesian inference models, and whenever Q is infinite-dimensional, one speaks of a Bayesian nonparametric inferential problem.

From de Finetti's theorem, it is also said that the sequence $(X_n)_{n\geq 1}$ is exchangeable if and only if it is a mixture of sequences of independent and identically distributed (i.i.d.) random variables. This property is expressed as conditionally independence, and some-

times it is written in a hierarchical form as

$$X_i|\tilde{p} \overset{\text{i.i.d.}}{\sim} \tilde{p}, \quad i \geq 1$$
$$\tilde{p} \sim Q.$$

Further studies of exchangeability can be found, for example, in Aldous (1985), Schervish (1995) and Kallenberg (2005).

Focusing on Bayesian nonparametric problems, there has been special interest in distributions Q which select discrete probability measures almost surely (a.s.). In general, RPMs selected by these distributions are represented as

$$\tilde{p}(\cdot) = \sum_{j=1}^{\infty} w_j \delta_{\xi_j}(\cdot), \tag{3}$$

where $w_j > 0$ for all j, with $\sum_{j\geq 1} w_j = 1$ a.s., and $(\xi_j)_{j\geq 1}$ is a sequence of i.i.d. $\mathbb{X}$-valued random variables—independent of $(w_j)_{j\geq 1}$—from a non-atomic distribution ν_0.

After giving this background, probability distributions for random partitions are constructed as follows. Suppose a sample $(X_1, \ldots, X_n)$ of size n is taken conditionally i.i.d. given $\tilde{p}$. Therefore, due to the discrete nature of $\tilde{p}$, there is a positive probability of having ties within the sample, that is $\mathbf{P}[X_i = X_j] > 0$ for $i \neq j$. In other words, the sample contains K_n distinct values, say $(X_1^*, \ldots, X_{K_n}^*)$, appearing with frequencies $(N_1, \ldots, N_{K_n})$. Clearly $K_n \leq n$ and $\sum_k N_k = n$ a.s. From (2), the joint distribution of the random vector $(K_n, N_1, \ldots, N_{K_n})$ is given by

$$\mathbf{P}(K_n = k, N_1 = n_1, \ldots, N_{K_n} = n_k) = \int_{\mathbb{X}^k} \mathbf{E}_Q\big(\tilde{p}^{n_1}(\mathrm{d}x_1) \cdots \tilde{p}^{n_k}(\mathrm{d}x_k)\big), \tag{4}$$

and it can be read as the probability of observing k distinct values in a sample from $\tilde{p}$ of size n, each one having frequency n_j.

Notice that the distinct values of the sample induce a partition $\{A_1, \ldots, A_k\}$ of $[n]$, where the blocks are such that $A_j = \{i : X_i = X_j^*\}$ for $j = 1, \ldots, k$. Hence, there is a one-to-one correspondence between each element of $\mathcal{P}_{[n]}$ and each possible realization of the random vector $(K_n, N_1, \ldots, N_{K_n})$ and we can define a probability distribution for some random partition Π_n from (4).

This construction for random-partition distributions has, however, an undetermined element: the distribution Q. There exists a vast literature about how to construct Q, and according to the approach selected, different expressions are obtained. Section 1.2

provides a study of some of such approaches. But before, two properties of (4) are studied: *consistency* and *exchangeability*, which have played an important role in such approaches.

1.1.2 Consistent partition distributions

Kingman (1978*a*,*b*)—who developed the concept of *partition structures* to refer to what we called random partitions—was the first in studying the property of *consistency*.

REMARK 1.1. For any $\pi \in \mathcal{P}_{[n+1]}$, $n \geq 1$, we said that $\pi_{|[n]}$ is the *restriction* of π to $\mathcal{P}_{[n]}$, and it is defined as

$$\pi_{|[n]} := \{\alpha \cap [n] \,:\, \alpha \cap [n] \neq \varnothing,\, \alpha \in \pi\}\,.$$

DEFINITION 2. A sequence $(\nu_n)_{n\geq 1}$ of $\mathcal{P}_{[n]}$-valued probability distributions is *consistent* if

$$\nu_n(\pi) = \nu_{n+1}(R_{n+1}^{-1}\pi), \qquad \pi \in \mathcal{P}_{[n]}, \tag{5}$$

for all $n \geq 1$; where $R_{n+1} : \mathcal{P}_{[n+1]} \to \mathcal{P}_{[n]}$ is the function induced by the operation of restriction defined above. Moreover, a sequence of random partitions $(\Pi_n)_{n\geq 1}$ such that $\Pi_n \sim \nu_n$ for $n \geq 1$, with $(\nu_n)_{n\geq 1}$ a sequence of consistent distributions, is said to be *consistent in distribution*.

This property essentially says that ν_n is the marginal distribution obtained from ν_{n+1} by deleting the element n from the set $[n]$, and it guarantees the existence of a random partition of $\mathbb{N}$ whose finite restrictions are distributed as ν_n. On the other hand, as already said, probability distributions constructed using (4) are all consistent; see Aldous (1985) for a proof.

As a side note, the definition of restriction provides an algorithmic construction for set partitions of $[n+1]$ based on set partitions of $[n]$, $n \geq 1$. Indeed, given $\pi \in \mathcal{P}_{[n]}$, the inverse image of R_{n+1} is the set

$$e(\pi) = \left\{\{\alpha_1, \ldots, \alpha_j \cup \{n+1\}, \ldots, \alpha_k\} \,:\, \alpha_j \in \pi,\, j = 1, \ldots, k\right\} \cup \{\pi \cup \{\{n+1\}\}\}\,,$$

with $k = \#\pi$ the number of blocks in π. Therefore, the set $\left\{e(\pi) \,:\, \pi \in \mathcal{P}_{[n]}\right\}$ corresponds to the set partitions of $[n+1]$. Figure 1 illustrates this approach.

1.1.3 Exchangeable partition probability functions

The second characteristic of the construction in (4) is known as *exchangeability*. This term has been used, for example, in Kingman (1982) and Aldous (1985). However, it is

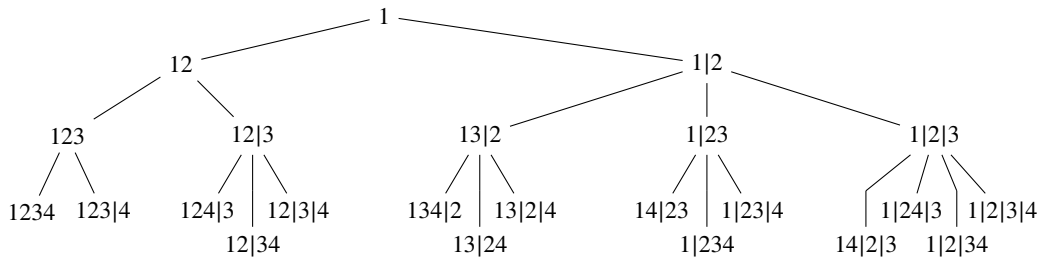


FIGURE 1: Tree showing a construction of set partitions. All the elements at the same depth, d, form the set $\mathcal{P}_{[d+1]}$. Partitions are displayed in a simplified way, so, for example, 13|2|4 corresponds to the partition $\{\{13\},\{2\},\{4\}\}$.

important to clarify that it does not refer to the property of exchangeability discussed previously, but to the symmetry in the arguments of the distribution.

REMARK 1.2. For any positive integer n, an *integer composition of n* is a sequence of positive integers $(n_1,\dots,n_k)$ such that $n_1+\cdots+n_k=n$. Let

$$\Delta_{n,k} := \left\{(n_1,\dots,n_k) \,:\, n_j \geq 1, \sum_{j=1}^{k} n_j = n\right\} \tag{6}$$

be the set of all integer compositions of n with exactly k summands.

DEFINITION 3. A random partition Π_n is *exchangeable* if, for any partition $\pi=\{\pi_1,\dots,\pi_k\}\in\mathcal{P}_{[n]}$,

$$\mathbf{P}(\Pi_n=\pi)=\mathbf{P}(\Pi_n=\rho(\pi)), \tag{7}$$

for all permutation ρ in the symmetric group acting on $[n]$. Equivalently,

$$\mathbf{P}(\Pi_n=\{\pi_1,\dots,\pi_k\})=\Pi_k^{(n)}(n_1,\dots,n_k), \tag{8}$$

for some symmetric function $\Pi_k^{(n)} : \Delta_{n.k}\to[0,1]$, with $n_j := \#\pi_j$, $j=1,\dots,k$, and where $\#\pi_j$ denotes the number of elements in π_j. Furthermore, the function $\Pi_k^{(n)}$ is called an *exchangeable partition probability function* (EPPF).

The concept of EPPF—introduced by Pitman (1995)—constitutes an important tool for the study of random partitions. Indeed, most of the literature about this topic has been focused on exchangeable random partitions. On the other hand, based on de Finetti's representation theorem, we can see that all induced random partitions are exchangeable, since it is clear that Equation (4) is a symmetric function with respect to $(n_1,\dots,n_k)$. Thus,

we can defined an EPPF by setting

$$\Pi_k^{(n)}(n_1, \dots, n_k) := \mathbf{P}(K_n = k, N_1 = n_1, \dots, N_{K_n} = n_k).$$

This characteristic makes EPPFs useful in Bayesian analyses when they are used as prior distributions since: (i) the probability of any partition does not depend on the possible values of the samples, but only on the size of each block, and (ii) the symmetry avoids to favor any particular partition from all of them having the same block sizes.

Additionally, it is important to say that—when working with exchangeable random partitions—the property of consistency (Equation 5) is satisfied if the *addition rule* holds for the EPPF, that is

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \Pi_{k+1}^{(n+1)}(n_1, \dots, n_k, 1) + \sum_{j=1}^{k} \Pi_k^{(n+1)}(n_1, \dots, n_j + 1, \dots, n_k).$$

Note, however, that not all exchangeable random partition is consistent. See, for example, Zhou & Walker (2014); consider also the following example.

EXAMPLE 1. Let Π_n be a random partition uniformly distributed, that is $\mathbf{P}(\Pi_n = \pi) = B_n^{-1}$ for any $\pi \in \mathcal{P}_{[n]}$ and where B_n is the nth Bell number. Clearly, Π_n is exchangeable. To see that it is not consistent, take the case $n = 2$ and $(n_1, n_2) = (1, 1)$. According to the addition rule, it would be satisfied that

$$\Pi_2^{(2)}(1, 1) = \Pi_3^{(3)}(1, 1, 1) + 2\Pi_2^{(3)}(2, 1),$$

however, the left side equals 1/2 whereas the right side equals 3/5.

1.2 Families of exchangeable partition probability functions

As explained in the previous sections—in order to define probability distributions for random partitions—we have to specify the de Finetti measure Q in the exchangeability-based construction given in Equation (4) or, equivalently, we have to provide an EPPF. In this section, different methodologies to build EPPFs are presented; some of them make use of specific forms of Q and others define the distribution in terms of the EPPF itself. A first approach is based on normalized random measures with independent increments, whereas a second one—known as Gibbs-type priors—is defined through the EPPF. Since these constructions are based on completely random measures, a brief description of this general class of measures is given first. Further details can be found, for example, in Lijoi

& Prünster (2010) and in the references therein.

DEFINITION 4. Let $\mathbb{M}$ denote the space of σ-finite measures on $(\mathbb{X}, \mathcal{X})$ with corresponding σ-algebra $\mathcal{M}$. Let $\tilde{\mu}$ be a measurable mapping taking values in $(\mathbb{M}, \mathcal{M})$ such that the random variables $\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_n)$ are independent for any $A_1, \dots, A_n \in \mathcal{X}$ with $A_i \cap A_j = \varnothing$, $i \neq j$. Then, $\tilde{\mu}$ is called a *completely random measure* (CRM).

Completely random measures have been introduced by Kingman (1967) and one of their characteristics is that they are a.s. discrete. Indeed, any CRM $\tilde{\mu}$ in $\mathbb{X}$ can be written as the sum of two components, i.e. $\tilde{\mu} := \tilde{\mu}_c + \tilde{\mu}_f$, where

1. $\tilde{\mu}_c$ is a measure written as
$$\tilde{\mu}_c = \sum_{i=1}^{\infty} J_i \delta_{\xi_i},$$
with positive jumps $(J_i)_{i\geq 1}$ and $\mathbb{X}$-valued locations $(\xi_i)_{i\geq 1}$, where both sequences are random and independent each other; and

2. $\tilde{\mu}_f$ is a measure with random masses at fixed locations, $x_1, \dots, x_M \in \mathbb{X}$ for $M \geq 1$, that is
$$\tilde{\mu}_f = \sum_{i=1}^{M} V_i \delta_{x_i},$$
with the nonnegative random jumps $(V_i)_{i=1}^{M}$ independent each other and independent of $\tilde{\mu}_c$.

Furthermore, the measure $\tilde{\mu}_c$ is characterized by the Lévy-Khintchine representation, corresponding to

$$\mathbf{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\tilde{\mu}_c(\mathrm{d}x)\right\}\right] = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} \left(1 - e^{-sf(x)}\right) \nu(\mathrm{d}s, \mathrm{d}x)\right\},$$

where $f : \mathbb{X} \to \mathbb{R}$ is a measurable function such that $\int |f| \mathrm{d}\tilde{\mu}_c < \infty$ a.s. and ν is a measure on $\mathbb{R}^+ \times \mathbb{X}$ such that, for any $B \in \mathcal{X}$,

$$\int_B \int_{\mathbb{R}^+} \min\{s, 1\} \nu(\mathrm{d}s, \mathrm{d}x) < \infty.$$

The measure ν is known as the *Lévy intensity*. It characterizes $\tilde{\mu}_c$ since it contains all the information about the distributions of the jumps $(J_i)_{i\geq 1}$ and locations $(\xi_i)_{i\geq 1}$.

For our purposes, we only consider CRMs with non-fixed locations, that is we will work with CRMs such that $\tilde{\mu} = \tilde{\mu}_c$. Moreover, we restrict our study to Lévy intensities ν

which factorize the jump and location parts as

$$\nu(\mathrm{d}s, \mathrm{d}x) = \rho_x(\mathrm{d}s)\alpha(\mathrm{d}x), \tag{9}$$

where α is a measure on $(\mathbb{X}, \mathcal{X})$—often called base measure—and ρ_x is a transition kernel on $\mathbb{X} \times \mathcal{B}(\mathbb{R}^+)$, that is $x \mapsto \rho_x(A)$ is $\mathcal{X}$-measurable for any $A \in \mathcal{B}(\mathbb{R}^+)$ and ρ_x is a measure on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ for any $x \in \mathbb{X}$ (c.f. Epifani et al. 2003). In the case when $\rho_x = \rho$ for any x, then the distribution of the jumps of $\tilde{\mu}_c$ is independent of their locations and, then, ν and $\tilde{\mu}_c$ are called *homogeneous*, otherwise they are called *non-homogeneous*.

Two important CRMs are the gamma and the σ-stable processes, since they induce two canonical random probability measures widely used in applications: the Dirichlet (Ferguson 1973) and the normalized σ-stable (Kingman 1975) processes, respectively. The gamma process is a homogeneous CRM with Lévy intensity given by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{e^{-s}}{s}\mathrm{d}s\,\alpha(\mathrm{d}x).$$

The σ-stable process—also known as the *one-parameter Poisson-Dirichlet process*—is a homogeneous CRM with Lévy intensity

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)s^{1+\sigma}}\mathrm{d}s\,\alpha(\mathrm{d}x),$$

for a given $\sigma \in (0, 1)$.

1.2.1 Normalized random measures with independent increments

Normalized random measures with independent increments are a wide class of de Finetti measures. Based on the construction of the Dirichlet process (Ferguson 1973, Section 4) through a gamma process, they were first studied in its general form by Regazzini et al. (2003). Their approach is based on Lévy processes, however, a more general construction based on CRMs is presented here.

DEFINITION 5. Let $\tilde{\mu}$ be a CRM on $\mathbb{X}$ such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ a.s. Then, the random probability measure $\tilde{p} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$ is called a *normalized random measure with independent increments* (NRMI).

In order to have a well defined measure $\tilde{p}$, the denominator $\tilde{\mu}(\mathbb{X})$ is constrained to be positive and finite. These conditions are expressed in terms of the Lévy intensity (9), so it satisfies $\rho_x(\mathbb{R}^+) = \infty$ for every x and $0 < \alpha(\mathbb{X}) < \infty$.

At this point, we can introduce a first family of probability distributions for random partitions. Since NRMIs induce exchangeable random partitions, an EPPF can be obtained and it is given by

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \frac{1}{\Gamma(n)} \int_0^\infty u^{n-1} e^{-\psi(u)} \prod_{j=1}^k \int_{\mathbb{X}} \tau_{n_j}(u; x) \alpha(\mathrm{d}x) \mathrm{d}u, \tag{10}$$

for a positive, finite and non-atomic measure α, and where $\psi(u)$ is the Laplace exponent, i.e.

$$\psi(u) = \int_{\mathbb{X}} \int_{\mathbb{R}^+} (1 - e^{-ut}) \rho_x(\mathrm{d}t) \alpha(\mathrm{d}x),$$

and $\tau_m(u; x)$ is defined for any $m \geq 1$ as

$$\tau_m(u; x) := \int_{\mathbb{R}^+} s^m e^{-us} \rho_x(\mathrm{d}s).$$

Some examples of NRMIs are presented below.

EXAMPLE 2 (Dirichlet process; Ferguson 1973). The Dirichlet process can be constructed as an NRMI by taking the gamma process explained before. It is necessary that $0 < \theta := \alpha(\mathbb{X}) < \infty$ in order to have a well defined probability measure. Furthermore, the EPPF associated to this process is given by

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^k \Gamma(n_j),$$

where $(x)_{n\uparrow}$ stands for the Pochhammer symbol defined as $(x)_{n\uparrow} := \prod_{j=0}^{n-1}(x + j)$ with $(x)_{0\uparrow} = 1$. This expression coincides with the one obtained by Antoniak (1974).

EXAMPLE 3 (Normalized σ-stable process; Kingman 1975). This process, like the previous one, results from the normalization of the σ-stable CRM. In this case, the EPPF is written as

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \frac{\sigma^{k-1}\Gamma(k)}{\Gamma(n)} \prod_{j=1}^k (1 - \sigma)_{n_j - 1\uparrow},$$

for $0 < \sigma < 1$.

EXAMPLE 4 (Normalized generalized gamma process; Lijoi et al. 2007*b*). A third example of NRMI is based on the generalized gamma process which has Lévy intensity

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\exp(-\tau s) s^{-(1+\sigma)}}{\Gamma(1 - \sigma)} \mathrm{d}s \, \alpha(\mathrm{d}x),$$

for $\tau \geq 0$ and $\sigma \in (0,1)$. In this case, its EPPF is given by

$$\Pi_k^{(n)}(n_1,\dots,n_k) = \frac{\sigma^{k-1}e^{\beta}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right) \prod_{j=1}^{k} (1-\sigma)_{n_j-1\uparrow}.$$

where $\beta = \tau^{\sigma}/\sigma$ and $\Gamma(x;z) = \int_z^{\infty} t^{x-1}e^{-t}\mathrm{d}t$ is the incomplete gamma function.

This latter example is more general than the first two since it comprises different RPMs. For example, the normalized inverse Gaussian process (Lijoi et al. 2005) is obtained by setting $\sigma = 1/2$, the normalized σ-stable process arises when $\tau = 0$ and the Dirichlet process corresponds to the case $\tau = 1$ and $\sigma \to 0$.

1.2.2 Poisson-Kingman models and Gibbs-type random partitions

Poisson-Kingman models (Pitman 2003) provide a general construction for discrete random probability measures. A particular case of them are the so called *Gibbs-type priors*—introduced by Gnedin & Pitman (2005)—which is a generic term encompassing random probability measures and their underlying random partition. Starting with a short explanation of Poisson-Kingman models, we focus on Gibbs-type random partitions and on their most representative example: the two-parameter Poisson-Dirichlet process.

Consider a homogeneous CRM $\tilde{\mu}$ with Lévy measure $\nu(\mathrm{d}s,\mathrm{d}x) = \rho(\mathrm{d}s)\alpha(\mathrm{d}x)$ such that $\rho(\mathbb{R}^+) = \infty$ and α is non-atomic probability measure. Denote by $(J_{(i)})_{i\geq 1}$ the ranked jumps of $\tilde{\mu}$ and let $T = \sum_{i\geq 1} J_{(i)}$ be the total mass. Assume the probability distribution of T is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$.

DEFINITION 6. Let $P_{\rho,t}$ be the conditional distribution of the sequence of ranked random probabilities $(\tilde{p}_{(i)})_{i\geq 1}$ defined as $\tilde{p}_{(i)} := J_{(i)}/T$ generated from a CRM as explained above, conditioned on $T = t$. Let η be a probability distribution on $\mathbb{R}^+$. Then, the distribution $\int_{\mathbb{R}^+} P_{\rho,t}\eta(\mathrm{d}t)$ on the simplex $\left\{(p_1,p_2,\dots) : p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i\geq 1} p_i = 1\right\}$ is called a *Poisson-Kingman distribution with Lévy density ρ and mixing distribution η*.

For these distributions, conditioned on $T = t$, the EPPF is given by

$$\Pi_k^{(n)}(n_1,\dots,n_k|t) = \int_0^t \frac{f(t-v)}{t^n f(t)} v^{n+k-1} \int_{S_k} \prod_{i=1}^{k} \rho(vu_i) u_i^{n_i}\,\mathrm{d}u_1\cdots\mathrm{d}u_{k-1}\mathrm{d}v,$$

where f is the density function of T and $\mathcal{S}_k$ is the $k-1$ simplex, i.e.

$$\mathcal{S}_k = \left\{ (u_1, \dots, u_k) : u_i \geq 0, i = 1, \dots, k, \sum_{i=1}^{k} u_i = 1 \right\}.$$

Poisson-Kingman models, as already said, provide a general construction of RPMs. In fact, when the distribution of T corresponds to η, their probability measures coincides with NRMIs. On the other hand, these models generate a class of RPMs known as *Gibbs-type RPMs* which are characterized by their EPPF.

DEFINITION 7. Let $\tilde{p}$ be a discrete random probability measure as in (3). Then, it is said that $\tilde{p}$ is a *Gibbs-type probability measure* if, for all $1 \leq k \leq n$ and any integer composition $(n_1, \dots, n_k)$ of n, its EPPF can be represented as

$$\Pi_k^{(n)}(n_1, \dots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1\uparrow},$$

for some $\sigma \in [0, 1)$. Moreover, its underlying random partition is called a *Gibbs-type random partition*.

In order to define a probability measure, the weights $V_{n,k}$ have to satisfy the forward recursive equation

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1},$$

for any $n \geq 1$ and $1 \leq k \leq n$, with $V_{1,1} = 1$.

An alternative way to define Gibbs-type priors from Poisson-Kingman models is also provided by Gnedin & Pitman (2005). First, define a *σ-stable Poisson-Kingman model* as the Poisson-Kingman model with Lévy density

$$\rho(s) = \frac{\sigma s^{-1-\sigma}}{\Gamma(1-\sigma)}.$$

Then, $\tilde{p}$ is a Gibbs-type prior if and only if it is a σ-stable Poisson-Kingman model. In this case, the weights $V_{n,k}$ take the form

$$V_{n,k} = \int_0^1 \frac{\sigma^k t^{-n}}{\Gamma(n - k\sigma) f_\sigma(t)} \int_0^t s^{n-k\sigma-1} f_\sigma(t-s) \mathrm{d}s \eta(\mathrm{d}t),$$

where f_σ is the density of a σ-stable random variable.

EXAMPLE 5 (Two-parameter Poisson-Dirichlet process; Perman et al. 1992). The two-

parameter Poisson-Dirichlet process is a σ-stable Poisson-Kingman model with mixing distribution

$$\eta(\mathrm{d}t) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)} t^{-\theta} f_\sigma(t)\mathrm{d}t,$$

for some $\theta > -\sigma$ and f_σ the density of a σ-stable random variable. Alternatively, it can be defined as a Gibbs-type random partition with weights

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1\uparrow}}, \tag{11}$$

leading the EPPF

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1\uparrow}.$$

This process—better known as *Pitman-Yor process* within the machine learning community—is one of the most important Gibbs-type random partitions. As $\sigma \to 0$, the Dirichlet process is recovered, and, when $\theta = 0$, the normalized σ-stable process is obtained. Another feature of this process is that it is not an NRMI. Indeed, the only NRMI which also belongs to the class of Gibbs type priors, for $0 < \sigma < 1$, is the normalized generalized gamma process (Lijoi et al. 2008).

1.2.3 Product partition distributions

Product partition distributions, better known as product partition models, were proposed by Hartigan (1990) to tackle classification problems under a parametric Bayesian setting. However, this class of distributions is of interest since it intersects with some of the already studied approaches, as explained by Quintana & Iglesias (2003) and Lijoi et al. (2007*b*).

DEFINITION 8. Let Π_n a $\mathcal{P}_{[n]}$-valued random partition. A *product partition distribution* (PPD) is the probability distribution of Π_n given by

$$\mathbf{P}(\Pi_n = \{\pi_1, \ldots, \pi_k\}) = M \prod_{j=1}^{k} c(\pi_j),$$

for any $\{\pi_1, \ldots, \pi_k\} \in \mathcal{P}_{[n]}$, and where $c : [n] \to \mathbb{R}^+ \cup \{0\}$ is a *cohesion function* and M is

the normalizing constant, i.e.

$$M^{-1} = \sum_{\pi \in \mathcal{P}_{[n]}} \prod_{j=1}^{\#\pi} c(\pi_j),$$

with $\#\pi$ the number of blocks in the partition π.

A PPD will be exchangeable or consistent according to the cohesion function chosen. For example, if $c(\pi_j) = \theta\Gamma(\#\pi_j)$, for some $\theta > 0$ and where $\#\pi_j$ is the number of elements in π_j, we recover the EPPF induced by the Dirichlet process.

1.2.4 Sequential construction via predictive distributions

The last construction discussed in this section is based on predictive distributions. Consider a sample $(x_1, \dots, x_n)$ from an a.s. discrete RPM $\tilde{p}$. Then, we can ask for the value of a new observation X_{n+1}; there are two possibilities: (i) it is equal to one of the values previously observed, or (ii) it is a new distinct value. We can compute the probability of each event.

Notice that there is a stochastic process, $(X_i)_{i\geq 1}$, behind. Hence, each one of the possible trajectories of such stochastic process of length n, say $(x_1, x_2, \dots, x_n)$, induces a partition $\{\pi_1, \dots, \pi_k\} \in \mathcal{P}_{[n]}$ where the blocks are given by $\pi_j = \{i : x_i = x_j^*\}$, $j = 1, \dots, k$, where $\{x_1^*, \dots, x_k^*\}$ are the different states visited by the process. Furthermore, the random-partition distribution is computed as follows

$$\begin{aligned}\mathbf{P}(\Pi_n = \{\pi_1, \dots, \pi_k\}) &= \mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \prod_{i=2}^{n} \mathbf{P}(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}). \end{aligned} \tag{12}$$

This construction is related to the *species sampling models* introduced by Pitman (1996), and notice that it is also related to the property of consistency of random partitions. However, it important to say that it is not always possible to obtain expressions of the underlying random partition from a species sampling model. In fact, the probability distribution constructed following this approach may be not exchangeable; see Lee et al. (2013) for a related discussion.

NRMIs and Gibbs-type priors are two cases where we can find explicit expressions

for (12). In these, the predictive distribution has the form

$$\mathbf{P}(X_{n+1} \in \cdot | X_1, \dots, X_n) = w_0 \nu_0(\cdot) + \sum_{j=1}^{k} w_j \delta_{X_j^*}(\cdot), \tag{13}$$

where ν_0 is a non-atomic distribution such that $X_1 \sim \nu_0$. Specifically, if an NRMI is chosen, the weights $(w_0, \dots, w_k)$ are given by

$$w_0 = \frac{\alpha(\mathbb{R})\Pi_{k+1}^{(n+1)}(n_1, \dots, n_k, 1)}{n\Pi_k^{(n)}(n_1, \dots, n_k)}, \quad \text{and} \quad w_j = \frac{\Pi_k^{(n+1)}(n_1, \dots, n_j + 1, \dots, n_k)}{n\Pi_k^{(n)}(n_1, \dots, n_k)}, \quad j = 1, \dots, k,$$

with $\nu_0(\cdot) = \alpha(\cdot)/\alpha(\mathbb{R})$. On the other hand, for a Gibbs-type prior, they are given by

$$w_0 = \frac{V_{n+1,k+1}}{V_{n,k}}, \qquad \text{and} \quad w_j = \frac{V_{n+1,k}}{V_{n,k}}(n_j - \sigma), \quad j = 1, \dots, k.$$

for ν_0 a non-atomic distribution.

1.3 Probability distribution of the number of blocks

As already explained, the random variable modeling the number of blocks, K_n, is implicitly defined when working with random partitions and it represents an important tool in both, theory and applications. There are studies, for example, about the asymptotic growing rate of such a variable as the sample size increases. Under the context of species sampling models, on the other hand, inferences are made based on this random variable. In practice, the study of K_n allows to have a more clear picture about the prior clustering behavior of a random partition as well as to select the parameters of the partition distribution adequately.

For Gibbs-type priors, the distribution of K_n can be written as

$$\mathbf{P}(K_n = k) = \frac{V_{n,k}}{\sigma^k} G(n, k, \sigma), \qquad k = 1, \dots, n,$$

where $G(n, k, \sigma)$ denotes the *generalized Stirling number*—also known as *generalized factorial coefficient*—defined as

$$G(n, k, \sigma) = \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} (-1)^j (-j\sigma)_{n\uparrow}.$$

Example 2, continued. The distribution of K_n for the Dirichlet process with parameter

$\theta > 0$ is given by

$$\mathbf{P}(K_n = k) = \frac{c(n,k)\theta^k}{(\theta)_{n\uparrow}}, \qquad k = 1, \dots, n,$$

where $c(n,k)$, $k = 1, \dots, n$, denotes the *unsigned Stirling number of the first kind* (EIS A132393). Its expected value is

$$\mathbf{E}(K_n) = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}.$$

On the other hand, Korwar & Hollander (1973) obtain that

$$\frac{K_n}{\log n} \to \theta, \qquad n \to \infty.$$

EXAMPLE 3, CONTINUED. For the normalized σ-stable process, we have

$$\mathbf{P}(K_n = k) = \frac{\Gamma(k)}{\sigma\Gamma(n)} G(n,k,\sigma), \qquad k = 1, \dots, n,$$

and

$$\mathbf{E}(K_n) = \frac{(1+\sigma)_{n-1\uparrow}}{\Gamma(n)}.$$

EXAMPLE 4, CONTINUED. The normalized generalized gamma process is an NRMI probability measure but—as explained—it can also be expressed as a Gibbs-type prior. Thus, the distribution of K_n is in this case

$$\mathbf{P}(K_n = k) = \frac{e^\beta G(n,k,\sigma)}{\sigma\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right) \qquad k = 1, \dots, n.$$

EXAMPLE 5, CONTINUED. For the two-parameter Poisson-Dirichlet process, the distribution of K_n corresponds to

$$\mathbf{P}(K_n = k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma^k(\theta+1)_{n-1}} G(n,k,\sigma) \qquad k = 1, \dots, n.$$

Furthermore,

$$\mathbf{E}(K_n) = \frac{(\theta+\sigma)_{n\uparrow}}{\sigma(\theta+1)_{n-1\uparrow}} - \frac{\theta}{\sigma},$$

Regarding its asymptotic grown, Pitman (2006) shows that

$$\frac{K_n}{n^\sigma} \to Y_{\theta/\sigma} \text{ a.s.}, \quad n \to \infty,$$

where Y_q has density given by

$$f(y) = \frac{\Gamma(q\sigma + 1)}{\sigma\Gamma(q+1)} y^{q-1-1/\sigma} f_\sigma(y^{-1/\sigma}),$$

for $q \geq 0$ and where $f_\sigma(\cdot)$ is the density of a positive stable random variable with parameter σ.

§ 2 Segmentations

From the previous section, we can see that random partitions are a vast field for researching. However, in some situations, there are additional constrains which make set partitions, or random partitions, not adequate to be used in such scenarios. These constrains change the kinds of valid groupings, and, as a consequence, different random objects are required to tackle the problems of interest. In this and the next sections, we present other clustering scenarios and study the latent combinatorial structures.

Clustering scenario 2. Suppose there is a set of items indexed by some ordered set T and we want to group the items in such a way that the items in the same block are formed only by those having consecutive indices. Setting $T = \{t_1, t_2, \dots, t_n\}$, the valid groupings are such that, if items y_{t_i} and y_{t_j}, for $t_i < t_j$, belong to the same block, then all the items y_{t_l}, for all t_l such that $t_i < t_l < t_j$, also belong to the same block.

For the case $n = 3$, we have $\{y_{t_1}, y_{t_2}, y_{t_3}\}$, and their indices are such that $t_1 < t_2 < t_3$. Then, the valid groupings under this specification are

$$\{\{y_{t_1}, y_{t_2}, y_{t_3}\}\},\ \{\{y_{t_1}\}, \{y_{t_2}, y_{t_3}\}\},\ \{\{y_{t_1}, y_{t_2}\}, \{y_{t_3}\}\},\ \{\{y_{t_1}\}, \{y_{t_2}\}, \{y_{t_3}\}\}.$$

Throughout this work, we call *segmentations* to the set conforming this kind of groupings. This is done by analogy with the way groupings are made: starting with the complete set of items $\{y_{t_1}, y_{t_2}, \dots, y_{t_n}\}$, it is split into k nonempty segments, for some positive $k \leq n$. Each segment becomes a block. It is worth saying that there is a unique way to list segmentations, unlike set partitions where the order-of-appearance assumption is necessary.

Using the symbolic method, segmentations are encoded by an unlabeled combinatorial class, whose specification is

$$\mathcal{S} = \mathbf{Seq}(\mathbf{Seq}_{\geq 1}(\mathcal{Z})).$$

In this case, its counting sequence is given by the *ordinary generating function* (OGF)

$$S(z) = \frac{1-z}{1-2z},$$

from where its nth coefficient is $S_n = [z^n]S(z) = 2^{n-1}$, for $n \geq 1$, with the assumption $S_0 = 1$ (EIS A000079).

Similarly to set partitions, it is possible to obtain the number of segmentations having exactly k blocks. This is encoded as

$$\mathcal{S}^{(k)} = \mathbf{Seq}_k(\mathbf{Seq}_{\geq 1}(\mathcal{Z})),$$

which has counting sequence

$$S^{(k)}(z) = \frac{z^k}{(1-z)^k},$$

and nth coefficient given by $S_n^{(k)} = [z^n]S^{(k)}(z) = \binom{n-1}{k-1}$, for $n \geq 1$, and $S_0^{(0)} = 1$ (cf. EIS A007318). In this case, its asymptotic form is $S_n^{(k)} \sim n^{k-1}/(k-1)!$

As a side note, there is a relationship between the segmentations of $[n]$ with exactly k blocks, $S_{[n]}^{(k)}$, and the set of integer compositions $\Delta_{n,k}$ defined in Equation (6). Actually, we have that the cardinality of both sets is the same. On the other hand, we also have that the class of segmentations is a subset of set partitions, that is $\mathcal{S}_{\mathcal{X}} \subseteq \mathcal{P}_{\mathcal{X}}$ for any set of items $\mathcal{X}$; the contention is proper whenever $\#\mathcal{X} \geq 3$. These are important facts that will be used during the study of probability distributions over segmentations.

§ 3 ORDERED SET PARTITIONS

For the third clustering scenario, we generalize Scenario 1 as follows.

CLUSTERING SCENARIO 3. Suppose there is a finite set of labeled items which we want to gather them into nonempty groups. In this case, groups are also labeled.

As an illustration, consider the case where we have three items: $\{x_1, x_2, x_3\}$. Then, all the groupings satisfying the given specification are

$$\{\{x_1, x_2, x_3\}\},\quad \{\{x_1\}, \{x_2, x_3\}\}, \{\{x_2\}, \{x_1, x_3\}\}, \{\{x_3\}, \{x_1, x_2\}\},$$
$$\{\{x_1, x_2\}, \{x_3\}\}, \{\{x_1, x_3\}, \{x_2\}\}, \{\{x_2, x_3\}, \{x_1\}\},\quad \{\{x_1\}, \{x_2\}, \{x_3\}\},$$
$$\{\{x_1\}, \{x_3\}, \{x_2\}\}, \{\{x_2\}, \{x_1\}, \{x_3\}\}, \{\{x_2\}, \{x_3\}, \{x_1\}\},$$
$$\{\{x_3\}, \{x_1\}, \{x_2\}\}, \{\{x_3\}, \{x_2\}, \{x_1\}\}.$$

The fact that the blocks are now distinguishable by a label—or ordered—forces us to count partitions like $\{\{x_1\},\{x_2,x_3\}\}$ and $\{\{x_2,x_3\},\{x_1\}\}$ as different. The order in the elements within the same block, however, continues to be irrelevant. Therefore, as we expected, the number of groupings under this scenario is bigger than the ones of the first scenario.

In order to obtain the combinatorial class which models this last clustering scenario, the number of blocks is fixed first. Thus, for a fixed k, the labeled combinatorial class defined as

$$\mathcal{O}^{(k)} = \mathbf{Seq}_k(\mathbf{Set}_{\geq 1}(\mathcal{Z})),$$

encodes the desired groupings. Notice that the operator $\mathbf{Seq}_k$ allows to identified the blocks; see Flajolet & Sedgewick (2009) for a further discussion. The EGF of this class is

$$O^{(k)}(z) = (e^z - 1)^k, \tag{14}$$

with nth coefficient given by

$$O_n^{(k)} = n!\,[z^n]O^{(k)}(z) = \sum_{j=0}^{k} \binom{k}{j}(-1)^k(k-j)^n,$$

(EIS A131689). Table 2 shows the counting sequence $(O_n^{(k)})_{k=0}^n$ for some values of n.

Now, if we take all the possible values of k in $\mathcal{O}^{(k)}$, the resulting class encodes all the possible groupings as explained in Scenario 3. This class is known as *ordered set partitions* and can be written as

$$\mathcal{O} = \bigcup_{k\geq 0} \mathbf{Seq}_k(\mathbf{Set}_{\geq 1}(\mathcal{Z})) = \mathbf{Seq}(\mathbf{Set}_{\geq 1}(\mathcal{Z})).$$

The counting sequence $(O_n)_{n\geq 0}$ is, then, obtained from the EGF

$$O(z) = \frac{1}{2 - e^z} = \sum_{j=0}^{\infty} \frac{e^{jz}}{2^{j+1}},$$

by taking its nth coefficient, which is

$$O_n = n!\,[z^n]O(z) = \sum_{j=0}^{\infty} \frac{j^n}{2^{j+1}}.$$

This sequence is known as the *Fubini numbers* or also ordered Bell numbers (EIS A000670) and are commonly denoted by $(F_n)_{n\geq 1}$; Table 2 lists the first ten of them. An alternative and simpler form to compute Fubini numbers is given by the sum $O_n = \sum_{k=0}^{n} O_n^{(k)}$, similarly to Bell numbers. Moreover, their asymptotic form—found, for example, in Gross (1962)—is $O_n \sim n!/(2\log^{n+1} 2)$.

Fubini numbers received its name by Comtet (1974) because they count the number of different ways to rearrange the ordering of integrals in Fubini's theorem; however, they appear in an early work of Cayley (1859) who counts a certain class of plain trees.

Similarly to segmentations, we have that ordered set partitions are a superset of set partitions, i.e. $\mathcal{P}_{\mathcal{X}} \subseteq \mathcal{O}_{\mathcal{X}}$, where the contention is proper for all the non trivial cases $\#\mathcal{X} \in \{0, 1\}$. Indeed, we have the following relationship

$$O_n^{(k)} = k!\, S_n^{(k)} = k!\, S(n, k),$$

for all $0 \leq k \leq n$. Note that the term $k!$ appears in order to differentiate the blocks. This also gives another form to compute Fubini numbers in terms of Stirling numbers of the second kind, namely $O_n = \sum_{k=0}^{n} k!\, S(n, k)$. Furthermore, due to their relationship with set partitions—under some probabilistic model—we should be able to infer not only about the clustering structure of certain dataset, but also about the number of groups.

3.1 Words

Classification is a problem very close to clustering. In this, the main interest is knowing to which category each item belongs. As we will see, the relationship between these two problems can be explained by means of the relationship between their combinatorial structure. We start describing classification in the following scenario.

Clustering scenario 4. Suppose we have a fixed number of categories, or blocks, each one identified, for example, by a label. Given a set of items, we are interested in knowing how the items are allocated into the different blocks. Unlike previous cases, there could be empty groups.

As an example, consider the case where there are three items, $\{x_1, x_2, x_3\}$, and two blocks to allocate them. Then, the induced groupings are

$$\{\{x_1, x_2, x_3\}, \{\}\},\ \{\{x_1, x_2\}, \{x_3\}\},\ \{\{x_1, x_3\}, \{x_2\}\},\ \{\{x_1\}, \{x_2, x_3\}\},$$
$$\{\{x_2, x_3\}, \{x_1\}\},\ \{\{x_2\}, \{x_1, x_3\}\},\ \{\{x_3\}, \{x_1, x_2\}\},\ \{\{\}, \{x_1, x_2, x_3\}\}.$$

	$k!\,S(n,k)$											F_n
$n \setminus k$	0	1	2	3	4	5	6	7	8	9	10	
0	1											1
1	0	1										1
2	0	1	2									3
3	0	1	6	6								13
4	0	1	14	36	24							75
5	0	1	30	150	240	120						541
6	0	1	62	540	1 560	1 800	720					4 683
7	0	1	126	1 806	8 400	16 800	15 120	5 040				47 293
8	0	1	254	5 796	40 824	126 000	191 520	141 120	40 320			545 835
9	0	1	510	18 150	186 480	834 120	1 905 120	2 328 480	1 451 520	362 880		7 087 261
10	0	1	1 022	55 980	818 520	5 103 000	16 435 440	29 635 200	30 240 000	16 329 600	3 628 800	102 247 563

Table 2: (First columns) Counting sequences, read by row, of ordered set partitions with exactly k blocks, $k!\,S(n,k)$, displayed in a triangular array for $0 \le k \le n \le 10$; note that $k!\,S(n,k) = 0$ for all $k > n$. (Last column) List of the first Fubini numbers, F_n, for $n = 0, \ldots, 10$.

In this case, grouping $\{\{x_1, x_2\}, \{x_3\}\}$ indicates that items x_1 and x_2 belong to the first block and x_3 belongs to the second one, whereas in $\{\{x_3\}, \{x_1, x_2\}\}$, the items are grouped in the same way but they are assigned to opposite blocks. Additionally, groupings like $\{\{x_1, x_2, x_3\}, \{\}\}$ are valid now, since there can be empty blocks.

This new structure differentiates groups, but not the order of the items within each group. Thus, from the example, the groupings $\{\{x_1, x_2\}, \{x_3\}\}$ and $\{\{x_2, x_1\}, \{x_3\}\}$ are treated as the same, but both of them are different to $\{\{x_3\}, \{x_1, x_2\}\}$.

The combinatorial class modeling this kind of grouping is known as *words* because of its usage analyzing letters in words. Under this context, every word is represented slightly different to the way showed in the example. Given a set of k different letters $\mathcal{L} = \{w_1^*, \dots, w_k^*\}$—called *alphabet*—a word of length $n \geq 1$ is represented by a sequence of letters $(w_1, w_2, \dots, w_n)$. There could be repeated letters. Thus, in the example, the first two groupings can be written as the words $(1, 1, 1)$ and $(1, 1, 2)$, assuming the alphabet is $\{1, 2\}$.

Using the symbolic method, the words obtained from a finite alphabet $\mathcal{L}$ are encoded by the labeled class

$$\mathcal{W}^{\mathcal{L}} = \mathbf{Seq}_{|\mathcal{L}|}(\mathbf{Set}(\mathcal{Z})).$$

Setting $k = \#\mathcal{L}$, the EFG of the class is

$$W^{(k)}(z) = e^{kz},$$

and its nth coefficient is, therefore, $W_n^{(k)} = n![z^n]W^{(k)}(z) = k^n$. Table 3 displays the number of words for different values of n and k.

There is an interesting relationship between the class of words and ordered set partitions. Suppose we are interested in the class of words of a k-letter alphabet where each letter appears at least r times, $r \geq 0$. This is encoded by

$$\mathcal{W}_{\geq r}^{\mathcal{L}} = \mathbf{Seq}_{|\mathcal{L}|}(\mathbf{Set}_{\geq r}(\mathcal{Z})).$$

When $r = 0$, the previous class of words is obtained, but for the case $r = 1$, each letter of the alphabet is forced to appear at least once. In this second case, the EFG is given by

$$W_{\geq 1}^{(k)}(z) = (e^z - 1)^k,$$

which is actually the same generating function obtained by the class of ordered set par-

	k^n									
$n \setminus k$	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	1	4	9	16	25	36	49	64	81	100
3	1	8	27	64	125	216	343	512	729	1 000
4	1	16	81	256	625	1 296	2 401	4 096	6 561	10 000
5	1	32	243	1 024	3 125	7 776	16 807	32 768	59 049	100 000
6	1	64	729	4 096	15 625	46 656	117 649	262 144	531 441	1 000 000
7	1	128	2 187	16 384	78 125	279 936	823 543	2 097 152	4 782 969	10 000 000
8	1	256	6 561	65 536	390 625	1 679 616	5 764 801	16 777 216	43 046 721	100 000 000
9	1	512	19 683	262 144	1 953 125	10 077 696	40 353 607	134 217 728	387 420 489	1 000 000 000
10	1	1 024	59 049	1 048 576	9 765 625	60 466 176	282 475 249	1 073 741 824	3 486 784 401	10 000 000 000

TABLE 3: Counting sequences of words of length n with an alphabet of exactly k letters for $1 \leq k, n \leq 10$.

titions with exactly k blocks, $\mathcal{O}^{(k)}$. We have already established the role of ordered set partitions in clustering, and—in this section—we have seen how words can be used in classification; but what is more interesting is that—under this latter restriction—both problems become the same.

During this chapter, we have defined and studied different clustering scenarios with particular emphasis on the combinatorial structure they induce. This fact results crucial when we want to tackle any problem related to these scenarios. In further chapters, we examine different and new methodologies to define probability distributions for each structure, and study some specific problems where they can be applied.

CHAPTER II

SEGMENTATION MODELS; AN APPLICATION TO CHANGE-POINT DETECTION

The study of *segmentation models* is the main focus of this chapter. By *segmentation models* we refer to a particular class of clustering models where the data are indexed by covariates. These covariates are points of some ordered set and play an important role in the models since they restrict the set of the possible—or valid—groupings. In Chapter 1, the Clustering Scenario 1.2 illustrates how segmentation models differ from the more general clustering model. There, it was also studied how the combinatorial structure called segmentations, denoted by $\mathcal{S}$, encodes all the valid groupings for these models. Hence, throughout this chapter, we complete the study of this topic by providing different constructions of probability distributions over segmentations and illustrating their usage under the context of change-point detection.

After presenting three different constructions of $\mathcal{S}_{[n]}$-valued probability distributions in Section 1, the rest of the chapter studies a novel retrospective change-point detection model. In this particular kind of classification models, the covariates are certain time units and the main goal is to infer—in terms of time indices—about the abrupt changes within the data. Thus, Section 2 presents with detail such a model. In this section, we justify our approach to tackle this problem—focusing on the prior distribution and the likelihood function—give an MCMC algorithm to draw samples from the posterior distribution, and illustrate its performance with simulated and real datasets. Finally, in Sections 3 and 4, we expose some of the current and future work in this research field: the estimation of regime parameters and the prediction of change-points; both of them are topics of interest in change-point analysis.

It is worth mentioning that most of the material presented in Sections 1 and 2 has been already published in Martínez & Mena (2014).

§1 Probability distributions for random segmentations

Let T be an ordered set formed by n distinct points: $t_1, \ldots, t_n$, $n \geq 1$, such that $t_i < t_j$ for all $i < j$. As explained in Section 1.2, the combinatorial class of segmentations, $\mathcal{S}_T$, models all the possible groupings of T where the blocks in each grouping are formed by consecutive elements. To be precise, let $\{\tau_1, \ldots, \tau_k\} \in \mathcal{S}_T$, $1 \leq k \leq n$, and let $t_i, t_l \in \tau_j$, for some j and $i < l$, then all $t_r \in T$ such that $t_i < t_r < t_j$ also belong to τ_k; the case where the block has only one element is trivial. Similar to the case of random partitions, it is sufficient to only work with the segmentations of $[n]$, $\mathcal{S}_{[n]}$, instead of the segmentations of T, since both classes are isomorphic. Thus, we define a *random segmentation*, Υ_n, as an $\mathcal{S}_{[n]}$-valued random variable.

In the next sections, three different constructions of $\mathcal{S}_{[n]}$-valued probability distributions are provided, all of them based on EPPFs. It is worth noting that—in addition to providing the probability distribution for the random segmentation—the underlying probability distribution for the number of blocks, denoted by B_n, is also studied.

1.1 Degeneration of random-partition probability distributions

Based on the fact that $\mathcal{S}_{[n]} \subseteq \mathcal{P}_{[n]}$ for all n—where the contention is proper for $n > 2$—Fuentes-García et al. (2010*b*) restrict $\mathcal{P}_{[n]}$-valued distributions under a classification context. Their approach consists in assigning probability zero to all partitions which are not segmentations; the remaining values are, then, normalized in order to define a probability distribution. Hence, the distribution of a random segmentation Υ_n is given by

$$\mathbf{P}(\Upsilon_n = \tau) = M\mathbf{P}(\Pi_n = \tau), \qquad \tau \in \mathcal{S}_{[n]}, \tag{1}$$

for Π_n a random partition, and where M is the normalizing constant

$$M^{-1} = \sum_{\tau \in \mathcal{S}_{[n]}} \mathbf{P}(\Pi_n = \tau).$$

Assuming Π_n has EPPF $\Pi_k^{(n)}$, the distribution (1), denoted by p^*, can be written as

$$p^*(n_1, \ldots, n_k) = M\Pi_k^{(n)}(n_1, \ldots, n_k), \tag{2}$$

for any $(n_1, \dots, n_k) \in \Delta_{n,k}$, where $\Delta_{n,k}$ is the set of all integer compositions of n with exactly k summands (see Equation 1.6, page 13). Note that—similar to EPPFs—p^* is a symmetric function.

Regarding the distribution of the number of blocks, B_n, it is obtained by marginalizing p^* as follows

$$\mathbf{P}(B_n = k) = \sum_{\tau \in S^k_{[n]}} \mathbf{P}(\Upsilon_n = \tau) = \sum_{(n_1,\dots,n_k)\in\Delta_{n,k}} p^*(n_1, \dots, n_k), \qquad k = 1, \dots, n.$$

Notice that, if this distribution is compared with the corresponding distribution for K_n—where K_n models the number of blocks in a $\mathcal{P}_{[n]}$-valued approach—we have that

$$\begin{aligned}\mathbf{P}(K_n = k) &= \sum_{\pi \in \mathcal{P}^k_{[n]}} \mathbf{P}(\Pi_n = \pi) = \sum_{\pi \in S^k_{[n]}} \mathbf{P}(\Pi_n = \pi) + \sum_{\pi \in (\mathcal{P}^k_{[n]} \setminus S^k_{[n]})} \mathbf{P}(\Pi_n = \pi) \\ &= M^{-1}\mathbf{P}(B_n = k) + \sum_{\pi \in (\mathcal{P}^k_{[n]} \setminus S^k_{[n]})} \mathbf{P}(\Pi_n = \pi),\end{aligned}$$

from where

$$\mathbf{P}(B_n = k) = M\mathbf{P}(K_n = k) - \sum_{\pi \in (\mathcal{P}^k_{[n]} \setminus S^k_{[n]})} M\mathbf{P}(\Pi_n = \pi).$$

The relevance of this equality is that—even though p^* is based on EPPFs and it is a symmetric function—the underlying distribution for B_n changes drastically with respect to the one for K_n. Such a change becomes greater as $k \to n/\log n$, where the cardinality of $S^k_{[n]}$ reaches its maximum.

1.2 Markov-chain based approach

A second approach to obtain probability distributions for random segmentations is based on the Pólya urn scheme given in Equation (1.13). In such a scheme, we start from a sequence $(X_n)_{n\geq 1}$—sampled from a discrete random probability measure—which induces a Markov chain $(K_n)_{n=1}^{\infty}$ obtained by considering $\mathbf{P}(X_{n+1} = \text{“new”}|X_1, \dots, X_n)$ and its complement. Indeed, this Markov chain models the dynamic of the number of blocks in $(X_n)_{n\geq 1}$. The probability of observing a new value indicates the appearance of a new block, otherwise the number of blocks remains unchanged. Therefore, the one-step tran-

sition probabilities of the Markov chain $(K_n)_{n=1}^{\infty}$ are given by

$$\begin{aligned}\mathbf{P}(K_{n+1} = k+1 | K_n = k) &:= \mathbf{P}(X_{n+1} \neq x_j^* \text{ for all } j \leq k | X_1 = x_{i_1}^*, \dots, X_n = x_{i_n}^*),\\ \mathbf{P}(K_{n+1} = k | K_n = k) &:= \mathbf{P}(X_{n+1} = x_j^* \text{ for some } j \leq k | X_1 = x_{i_1}^*, \dots, X_n = x_{i_n}^*),\end{aligned}$$

where $\{x_1^*, x_2^*, \dots, x_k^*\}$ are the k distinct values observed in $(X_i)_{i=1}^n$, and $(i_1, \dots, i_n) \subset [k]$.

These probabilities can be written in terms of EPPFs as explained in Section 1.1.2.4. Let $\Pi_k^{(n)}$ be an EPPF and $(n_1, \dots, n_k) \in \Delta_{n,k}$, then

$$\mathbf{P}(K_{n+1} = k+1 | K_n = k) = \frac{\Pi_{k+1}^{(n+1)}(n_1, \dots, n_k, 1)}{\Pi_k^{(n)}(n_1, \dots, n_k)}, \tag{3}$$

$$\mathbf{P}(K_{n+1} = k | K_n = k) = \frac{\sum_{j=1}^k \Pi_k^{(n+1)}(n_1, \dots, n_j + 1, \dots, n_k)}{\Pi_k^{(n)}(n_1, \dots, n_k)}, \tag{4}$$

with $K_1 = 1$. Note that this Markov chain is a non-homogeneous birth process. Furthermore, each of its paths of length n induces a segmentation of $[n]$; see Figure 1 for an illustration. Hence, it is possible to define a probability distribution for a random segmentation Υ_n; specifically

$$\begin{aligned}\mathbf{P}(\Upsilon_n = \{\tau_1, \dots, \tau_k\}) &= \mathbf{P}(K_1 = i_1, K_2 = i_2, \dots, K_n = i_n)\\ &= \prod_{j=1}^{n-1} \mathbf{P}(K_{j+1} = i_{j+1} | K_j = i_j),\end{aligned} \tag{5}$$

for $\{\tau_1, \dots, \tau_k\} \in \mathcal{S}_{[n]}$, $1 \leq k \leq n$, and where the sequence $(i_1, \dots, i_n) \subset [k]$ is such that

$$\begin{aligned} i_1 = i_2 = \cdots = i_{n_1} &= 1,\\ i_{n_1+1} = i_{n_1+2} = \cdots = i_{n_1+n_2} &= 2,\\ &\vdots\\ i_{s_{k-1}+1} = i_{s_{k-1}+2} = \cdots = i_n &= k,\end{aligned} \tag{6}$$

with $s_j = n_1 + \cdots + n_j$, and where $n_j = \#\tau_j$ for $j = 1, \dots, k$. Notice that (5) depends on a segmentation only through its blocks' sizes $(n_1, \dots, n_k)$, thus—using (3) and (4)—it can be written as

$$p^\star(n_1, \dots, n_k) = \prod_{j=1}^{k-1} \kappa(s_j, j, j+1) \prod_{j=1}^{k} \prod_{l=1}^{n_j - 1} \kappa(s_{j-1} + l, j, j), \tag{7}$$

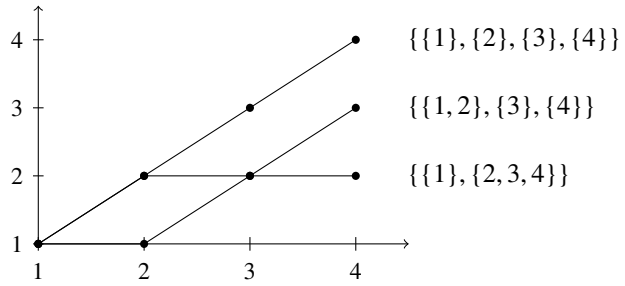


FIGURE 1: Examples of the possible paths for the Markov chain $(K_n)_{n\geq 1}$ until time $n = 4$. As shown in the right side of the plot, every path induces a segmentation of $\{1, 2, 3, 4\}$.

where $\kappa(r, k, k') := \mathbf{P}(K_{r+1} = k' | K_r = k)$ and with $s_0 = 0$. It is also important to note, however, that the distribution $p^\star$ is not symmetric because of the term s_j.

From the construction of the Markov chain $(K_n)_{n=1}^{\infty}$, the number of blocks, B_n, is immediately obtained from the marginal distribution of K_n. Similar to the previous construction, the probability distribution for B_n can be computed by marginalizing (7), leading to

$$\begin{aligned}\mathbf{P}(B_n = k) &= \sum_{(i_1,\ldots,i_n)} \prod_{j=1}^{n-1} \mathbf{P}(K_{j+1} = i_{j+1} | K_j = i_j) \\ &= \sum_{(n_1,\ldots,n_k)\in\Delta_{n,k}} \prod_{j=1}^{k-1} \kappa(s_j, j, j+1) \prod_{j=1}^{k} \prod_{l=1}^{n_j-1} \kappa(s_{j-1} + l, j, j), \end{aligned} \tag{8}$$

where the first sum is over all the sequences $(i_1, \ldots, i_n)$ like in (6)—that is, over all the sequences of positive integers such that $1 = i_1 \leq i_2 \leq \cdots \leq i_n = k \leq n$, and $i_j - i_{j-1} \leq 1$ for all $1 < j \leq n$—and for the second sum, recall that $s_j = n_1 + \cdots + n_j$ with $s_0 = 0$.

As a side note, it is worth saying that this approach results of interest since it could be considered as tantamount to the one obtained via infinite hidden Markov models, with the state Markov process restricted to have upper triangular transition matrices (see, for example, Ko et al. 2015, and the references therein). In species sampling problems, the study of the marginal statistic B_n could be also an appealing methodology to infer about the number of species (Lijoi et al. 2007*a*).

1.3 Degeneration based on integer compositions

The last approach studied in this chapter to construct $\mathcal{S}_{[n]}$-valued probability distributions is also a degeneration of $\mathcal{P}_{[n]}$-valued probability distributions. However, such a

degeneration is made according certain rules. The idea of this construction consists in taking the mass of all the partitions having block sizes $(n_1, \dots, n_k)$ regardless their order, and uniformly distribute the whole mass into all the segmentations with the same block sizes.

In order to obtain the corresponding probability distribution—given an integer composition $(n_1, \dots, n_k)$ of n—define the set

$$\Delta(n_1, \dots, n_k) := \left\{(n_{\rho(1)}, \dots, n_{\rho(k)}) \in \Delta_{n,k} \text{ for all the permutations } \rho \text{ of } [k]\right\};$$

notice that the cardinality of this set is not $k!$ since repeated elements are included only once; for example, consider $(1, 1, 2)$, then $\Delta(1, 1, 2) = \{(1, 1, 2), (1, 2, 1), (2, 1, 1)\}$. Let $\Pi_k^{(n)}$ be an EPPF. Due to the symmetry of $\Pi_k^{(n)}$, the probability of any element in $\Delta(n_1, \dots, n_k)$ is the same, so

$$\sum_{n' \in \Delta(n_1,\dots,n_k)} \Pi_k^{(n)}(n_1', \dots, n_k') = P_{n_1,\dots,n_k} \Pi_k^{(n)}(n_1, \dots, n_k),$$

where $P_{n_1,\dots,n_k}$ is the number of all the partitions having block sizes in $\Delta(n_1, \dots, n_k)$. This probability will be uniformly reassigned among all the segmentations with block sizes in the same set; denote by $S_{n_1,\dots,n_k}$ the number of such segmentations. Therefore, the probability distribution of interest is given by a function $p' : \Delta_{n,k} \to [0, 1]$ such that

$$p'(n_1, \dots, n_k) = \frac{P_{n_1,\dots,n_k}}{S_{n_1,\dots,n_k}} \Pi_k^{(n)}(n_1, \dots, n_k). \tag{9}$$

The numbers $P_{n_1,\dots,n_k}$ and $S_{n_1,\dots,n_k}$ can be found by means of the symbolic method and introducing *markers* in the specification of set partitions, $\mathcal{P}$, and segmentations, $\mathcal{S}$, given in the previous chapter (Flajolet & Sedgewick 2009, Chapter 3).

Let $(\chi_j)_{j \geq 1}$ be a sequence markers—where χ_j marks the number of blocks of size j—and consider the following labeled combinatorial class

$$\mathcal{P} = \mathbf{Set}(\chi_1 \mathbf{Set}_1(\mathcal{Z}) + \chi_2 \mathbf{Set}_2(\mathcal{Z}) + \cdots + \chi_j \mathbf{Set}_j(\mathcal{Z}) + \cdots).$$

The EGF of this class is a function of z and $u := (u_1, u_2, \dots)$ and it is given by

$$P(z, u) = \exp\left\{\sum_{j \geq 1} \frac{u_j z^j}{j!}\right\} = \prod_{j \geq 1} \sum_{r \geq 0} \frac{u_j^r}{r!} \frac{z^{jr}}{j!^r}.$$

For a fixed n, the number of partitions having m_1 blocks of size 1, m_2 blocks of size 2, and so on, equalizes $n!$ times the coefficient $[u_1^{m_1}][u_2^{m_2}]\cdots[u_n^{m_n}][z^n]$ of $P(z,u)$. Notice that it should be satisfied $n = \sum_{j\geq 1} jm_j$. Therefore, the number of partitions having block sizes in $\Delta(n_1,\ldots,n_k)$ is

$$P_{n_1,\ldots,n_k} = n!\,[u_1^{m_1}][u_2^{m_2}]\cdots[u_n^{m_n}][z^n]P(z,u) = \frac{n!}{\prod_{j=1}^{n} m_j!\; j!^{m_j}}. \tag{10}$$

Similarly, consider the following unlabeled class

$$\mathcal{S} = \mathbf{Seq}(\chi_1\mathbf{Seq}_1(\mathcal{Z}) + \chi_2\mathbf{Seq}_2(\mathcal{Z}) + \cdots + \chi_j\mathbf{Seq}_j(\mathcal{Z}) + \cdots),$$

where χ_j marks the number of blocks of size j; then, its OFG is

$$S(z,u) = \sum_{r\geq 0}\left\{\sum_{j\geq 1} u_j z^j\right\}^r.$$

In this case, it is necessary to find the kth coefficient of z such that it has m_1 blocks of size 1, m_2 blocks of size 2, and so on, in order to match it with the corresponding $P_{n_1,\ldots,n_k}$. Therefore, such a coefficient is given by

$$S_{n_1,\ldots,n_k} = [u_1^{m_1}][u_2^{m_2}]\cdots[u_n^{m_n}][z^k]S(z,u) = \frac{k!}{m_1!\cdots m_n!} = \binom{k}{m_1,\ldots,m_n}. \tag{11}$$

Note that $k = \sum_{j=1}^{n} m_j$ is the total number of blocks.

Dividing (10) by (11), we have

$$\frac{P_{n_1,\ldots,n_k}}{S_{n_1,\ldots,n_k}} = \frac{n!}{k!\prod_{j=1}^{n} j!^{m_j}}.$$

However—in order to express this quantity in terms of the block sizes $(n_1,\ldots,n_k)$ instead of the groups of size j $(m_1,\ldots,m_n)$—the following equivalence is used

$$\prod_{j=1}^{n} j!^{m_j} = \prod_{j=1}^{k} n_j!\,,$$

which can be proved as follows. Expanding the factorials in the right side, we have

$$\prod_{j=1}^{k} n_j! = 1^{a_1}2^{a_2}\cdots n^{a_n},$$

where $a_j = \sum_{i=1}^{k} \mathbf{1}(n_i \geq j)$, whereas the left side is written as

$$\prod_{j=1}^{n} j!^{m_j} = 1^{m_1+\cdots+m_n} 2^{m_2+\cdots+m_n} \cdots n^{m_n},$$

thus—since $m_j = \sum_{i=1}^{n} \mathbf{1}(n_i = j)$ for $j = 1, \ldots, n$—the exponent in each term is given by

$$\sum_{j=r}^{n} m_j = \sum_{j=r}^{n} \sum_{i=1}^{n} \mathbf{1}(n_i = j) = \sum_{i=1}^{n} \sum_{j=r}^{n} \mathbf{1}(n_i = j) = \sum_{i=1}^{n} \mathbf{1}(n_i \geq r) = a_r.$$

Hence, the probability distribution (9) for a random segmentation Υ_n is given by

$$\mathbf{P}(\Upsilon_n = \{\tau_1, \ldots, \tau_k\}) = p'(n_1, \ldots, n_k) = \binom{n}{n_1, \ldots, n_k} \frac{1}{k!} \Pi_k^{(n)}(n_1, \ldots, n_k), \tag{12}$$

for $\{\tau_1, \ldots, \tau_k\} \in \mathcal{S}_{[n]}$ and where $n_j = \#\tau_j$ for $j = 1, \ldots, n$.

The distribution of the number of blocks, B_n, is obtained as usual. However—given the way the EPPF was degenerated—there is a nice relation between B_n and K_n. Notice first that, for a fixed k, there is a finite collection of disjoint sets $\Delta(n_1, \ldots, n_k)$ whose union is $\Delta_{n,k}$. Let $\Delta_{n,k}^*$ be the set of the different sequences $(n_1, \ldots, n_k)$ generating such sets $\Delta(n_1, \ldots, n_k)$. Furthermore, denote by $\mathcal{P}_{n_1,\ldots,n_k}$ and $\mathcal{S}_{n_1,\ldots,n_k}$ the sets of all partitions and segmentations, respectively, having block sizes $(n_1, \ldots, n_k)$; clearly—from (10) and (11)—the cardinality of each set is $\#\mathcal{P}_{n_1,\ldots,n_k} = P_{n_1,\ldots,n_k}$ and $\#\mathcal{S}_{n_1,\ldots,n_k} = S_{n_1,\ldots,n_k}$. Therefore

$$\begin{aligned}
\mathbf{P}(B_n = k) &= \sum_{\tau \in \mathcal{S}_{[n]}^k} \mathbf{P}(\Upsilon_n = \tau) = \sum_{\Delta_{n,k}^*} \sum_{\tau \in \mathcal{S}_{n_1,\ldots,n_k}} \mathbf{P}(\Upsilon_n = \tau) \\
&= \sum_{\Delta_{n,k}^*} S_{n_1,\ldots,n_k} \mathbf{P}(\Upsilon_n = \tau^*) = \sum_{\Delta_{n,k}^*} P_{n_1,\ldots,n_k} \mathbf{P}(\Pi_n = \tau^*) \\
&= \sum_{\Delta_{n,k}^*} \sum_{\tau \in \mathcal{P}_{n_1,\ldots,n_k}} \mathbf{P}(\Pi_n = \tau) = \sum_{\tau \in \mathcal{P}_{[n]}^k} \mathbf{P}(\Pi_n = \tau) \\
&= \mathbf{P}(K_n = k),
\end{aligned}$$

where τ^* is any element of $\mathcal{S}_{n_1,\ldots,n_k}$, Π_n is a random partition with EPPF $\Pi_k^{(n)}$, and the second line follows from (9).

It is worth saying that—based on the notion of covariate proximity and covariate-dependent weights, under a context of Dirichlet process mixture models for curve fitting—Wade et al. (2014) work with a $\mathcal{S}_{[n]}$-valued distribution having the form (12) with the EPPF $\Pi_k^{(n)}$ induced by the Dirichlet process. They also show that the distributions of

π	p'	p^*	$p^\star$
$\{\{1,2,3,4\}\}$	0.029	0.041	0.029
$\{\{1\},\{2,3,4\}\}$	0.066	0.047	0.092
$\{\{1,2\},\{3,4\}\}$	0.039	0.019	0.046
$\{\{1,2,3\},\{4\}\}$	0.066	0.047	0.033
$\{\{1\},\{2\},\{3,4\}\}$	0.136	0.097	0.204
$\{\{1\},\{2,3\},\{4\}\}$	0.136	0.097	0.136
$\{\{1,2\},\{3\},\{4\}\}$	0.136	0.097	0.068
$\{\{1\},\{2\},\{3\},\{4\}\}$	0.392	0.557	0.392

(A) Distribution of Υ_4, $\mathbf{P}(\Upsilon_4 = \pi)$

k	p'	p^*	$p^\star$	p
1	0.029	0.041	0.029	0.029
2	0.171	0.112	0.171	0.171
3	0.408	0.290	0.408	0.408
4	0.392	0.557	0.392	0.392

(B) Distribution of B_4, $\mathbf{P}(B_4 = k)$

TABLE 1: Distribution of Υ_4 and B_4 under the different approaches to construct probability distributions on $\mathcal{S}_{[n]}$: degenerated case p^*, Markov-chain approach, $p^\star$, and integer-composition degenerated construction, p'. The parameters are $(\sigma, \theta) = (0.35, 2.7)$ in all cases. For the distribution of B_4, it is also shown the $\mathcal{P}_{[n]}$-valued case, p.

B_n and of K_n coincide for this specific EPPF.

As an illustration of the performance of each one of these constructions, Table 1 displays them for the case where the EPPF is induced by the two-parameter Poisson-Dirichlet process (Example 1.5). The corresponding distributions for Υ_n are, then, given by

$$p^*(n_1,\dots,n_k) = M \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1\uparrow} \tag{13a}$$

$$p^\star(n_1,\dots,n_k) = \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k}(s_{j-1}+1-j\sigma)_{n_j-1\uparrow}, \tag{13b}$$

$$p'(n_1,\dots,n_k) = \frac{n!}{k!} \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k} \frac{(1-\sigma)_{n_j-1\uparrow}}{n_j!}, \tag{13c}$$

where M is the normalizing constant of p^* given by

$$M^{-1} = \frac{1}{(\theta+1)_{n-1\uparrow}} \sum_{k=1}^{n} \prod_{i=1}^{k-1}(\theta+i\sigma) \sum_{(n_1,\dots,n_k)\in\Delta_{n,k}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1\uparrow}.$$

§ 2 A RETROSPECTIVE CHANGE-POINT DETECTION MODEL FOR MARKOVIAN DATA

One class of segmentation models are the well known *change-point detection* models. When observing a phenomenon evolving in time, there might be some abrupt variations in the measurements suggesting that something has altered its regular behavior. Change-point detection models study these kinds of variations, and the main goal is to detect the number and the location of such changes.

First works can be traced back to Page (1954), Shiryaev (1963), Chernoff & Zacks (1964), Kander & Zacks (1966) and Hinkley & Hinkley (1970). However, the interest in these models continues because of their attractiveness in many different fields. For example, we can find applications in genetics (in DNA segmentation, Braun et al. 2000, and phylogenetic recombination detection, Minin et al. 2007), signal processing (for instance, in EEG analysis, Kaplan & Shishkin 2000, and signal segmentation, Punskaya et al. 2002), environmental time series (for detecting changes in wind speed and direction, Dobigeon & Toumeret 2007, and in hydrometeorological data, Perreault et al. 2000), econometrics (detecting changes in variance for stock prices, Chen & Gupta 1997, and analyzing inflation dynamics, Jochmann 2010) among others.

From a methodological perspective, the literature on this topic is also vast, either from a frequentist or a Bayesian perspective. First models dealt mainly with the detection of a single change-point (Chernoff & Zacks 1964, Kander & Zacks 1966, Hinkley & Hinkley 1970, Smith 1975) where such a change point was induced by a change in the regime parameter within a parametric family modeling sequential observations. Later on, the model of Smith (1975) was extended to a nonparametric regime distribution by Muliere & Scarsini (1985) and Mira & Petrone (1996) who made use of the Dirichlet process as the regime model.

For these first change-point detection models, we can see that the underlying support of the clustering is a subset of segmentations, specifically all those with at most two blocks. However, the generalization to multiple change points—even though it requires more demanding mathematical and computational treatments—has also been studied. For instance, within the Bayesian parametric framework, Green (1995) makes use of reversible jump Markov chain Monte Carlo (MCMC) techniques to draw posterior inferences on the number of change points. In addition, and under a different rationale, Barry & Hartigan (1992, 1993) proposed product partition models (cf. Section 1.1.2.3) and adjusted them properly in order to apply them in this scenario; works of Loschi et al. (2003) and Loschi & Cruz (2005) follow this methodology. There are also Bayesian non-

parametric approaches tackling this problem, for example, Quintana & Iglesias (2003), Park & Dunson (2010) and Müller & Quintana (2011), by means of random-partition methodologies.

Nowadays, multiple-change-point detection models seem to prevail, and a new category emerged. All the previous approaches are now known as *retrospective* models, since they estimate the change-points from a given dataset. But there is a different approach—called *on-line* methods—where change-points locations are estimated as new observations are available. This kind of approach is commonly used in Computer Science, from where the term *on-line* is borrowed; examples of on-line change-point detection models are Fearnhead & Liu (2007) and Caron et al. (2012). Nevertheless, both approaches are used in practice.

Hence, in this section, we present our novel retrospective model for detecting multiple change points, which can be found in Martínez & Mena (2014). Starting with a series of dependent data $y = (y_{t_1}, y_{t_2}, \dots, y_{t_n})$ observed at times $t_1, t_2, \dots, t_n$, such that $t_1 < t_2 < \cdots < t_n$, our interest is in the posterior distribution

$$\mathbf{P}(\Upsilon_n | y) \propto \mathbf{P}(\Upsilon_n)\mathbf{P}(y|\Upsilon_n), \tag{14}$$

where Υ_n is a random segmentation modeling all the possible change-point locations.

Notice that a random segmentation is a good candidate for modeling these locations. Consider the combinatorial class of all segmentations of the set T, $\mathcal{S}_T$, where T is the ordered set formed by the time points $t_1, \dots, t_n$ where the data were observed. Let $\{\tau_1, \dots, \tau_k\} \in \mathcal{S}_T$ and let $t_{i,1} = \min(\tau_i)$, $i = 1, \dots, k$. Then, from the definition of $\mathcal{S}_T$ (Section 1.2), every t_j such that $t_{i,1} \leq t_j < t_{i+1,1}$ for $i = 1, \dots, k-1$, and $t_{k,1} \leq t_j \leq t_n$, belongs to the set τ_i. Therefore, the time constrain is preserved by every segmentation, and furthermore, every possible sequence of change-point locations is encoded by only one segmentation. This differs from other partition-based approaches, like the ones mentioned above, which make use of random partitions to model the locations.

Providing some terminology and notation, it is said that all the observations (y_{t_j}) associated to the points $t_j \in \tau_i$—for τ_i a block in some segmentation $\{\tau_1, \dots, \tau_k\}$—belong to the same *regime*. Moreover, the points $(t_{1,1}, t_{2,1}, \dots, t_{k,1})$, $1 \leq k \leq n$, defined in the previous paragraph, correspond to the *change-point locations* of the dataset. However, throughout this work, it will be assumed that the point $t_{1,1} = t_1$ is not a change-point; this allows us to have the case of no change points, which corresponds to the segmentation $\{\tau_1\} = \{T\}$. As a consequence, the number of change points will range from zero to $n-1$.

On the other hand—even though Υ_n takes values in $\mathcal{S}_{[n]}$—the class of segmentations, $\mathcal{S}_{[n]}$, is isomorphic to the set of all integer compositions of $[n]$, i.e. $\cup_{k=1}^{n}\Delta_{n,k}$, therefore, we will use both sets interchangeably when working with Υ_n. Finally, a sequence of change-point locations will be denoted by $\langle t_1^*, \dots, t_k^*\rangle$, where $t_j^* = t_{j+1,1}$ for $j = 1, \dots, k-1$—or equivalently, $t_j^* = t_{s_j+1}$ where $s_j = n_1 + \cdots + n_j$ and $n_j = \#\tau_j$—and $\langle\rangle$ will indicate the case of no change points.

2.1 Prior distribution and likelihood function

Once Model (14) has been introduced and it makes sense for detecting change-point locations, it is necessary to provide specific forms to the prior $\mathbf{P}(\Upsilon_n)$ and the likelihood $\mathbf{P}(y|\Upsilon_n)$.

Regarding the prior for Υ_n, we will work with the third construction, p', given in Section 1.3. The reasons for this choice are that: (i) since we are working with a retrospective model, the symmetry of p' allows to be *non informative* in the sense that the same prior probability will be assigned to any segmentation having the same blocks' sizes—feature not obtained with $p^\star$ for example—and (ii) the fact that the distribution for the number of blocks is the same to the one in an EPPF approach inherits all the existing results to our model, which is not true for p^*.

The likelihood function $\mathbf{P}(y|\Upsilon_n)$ has also some elements to consider. Our approach, like many others in change-point detection, is *model-based*: data are modeled by some stochastic process with law F which is constant by regimes; thus, the ability for detecting certain kind of changes is determined by this law. Approaches like product partition models (PPMs) introduced by Barry & Hartigan (1992, 1993), work with a conditional independent-and-identically-distributed law, that is, there is a finite-dimensional parameter x, not observed, of F such that—conditional on it—observations within a regime are independent and identically distributed according to $F(x)$; moreover, observations in different regimes are independent. The detection problem is, then, restated as finding the time locations where parameter x changes. Then, the likelihood function is written as

$$\begin{aligned}\mathbf{P}(y|\Upsilon_n = (n_1, \dots, n_k), x = (x_1, \dots, x_k)) &= \prod_{j=1}^{k} F(y_{t_{j,1}}, \dots, y_{t_{j,n_j}}; x_j) \\ &= \prod_{j=1}^{k}\prod_{i=1}^{n_j} F(y_{t_{j,i}}; x_j),\end{aligned} \tag{15}$$

where $y_{j,i}$ denotes the ith observation in the jth regime, that is, $y_{j,i} := y_{t_{s_j+i}}$ for $j = 1, \dots, k$ and $i = 1, \dots, n_j$. In some cases, parameter x is integrated out, so (15) becomes

$$\mathbf{P}(y|\Upsilon_n = (n_1, \dots, n_k)) = \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i=1}^{n_j} F(y_{t_{j,i}}; x_j) \nu_0(\mathrm{d}x_j), \tag{16}$$

where ν_0 is the prior distribution of x_j.

Our choice for $\mathbf{P}(y|\Upsilon_n)$ has the inter-regime independence property, like PPMs, but it uses a different stochastic process to modeling data: we consider a strictly stationary Markovian process, so, data are dependent within regimes. The choice of a specific process is bounded by the nature of the phenomenon under study, but it is worth saying that—although more complex process could be considered—the model-based nature of these problems suggests that these processes do not necessarily need to be very complicated; sudden changes, jumps or other kinds of distributional variations could simply be part of a different regime under the selected model.

Hence—assuming the Markovian process has invariant distribution $\pi(y_t; x)$ and transition density $p(y_0, y_t; x)$, where x denotes generically the driving parameter—the likelihood function (15) is written as

$$\mathbf{P}(y|\Upsilon_n = (n_1, \dots, n_k), x = (x_1, \dots, x_k)) = \prod_{j=1}^{k} \pi(y_{j,1}; x_j) \prod_{l=1}^{n_j - 1} p(y_{j,l}, y_{j,l+1}; x_j). \tag{17}$$

Some examples of strictly stationary Markovian processes having explicit invariant and transition densities are studied next.

Ornstein-Uhlenbeck process. Assuming the nature of the phenomena producing y evolves in continuous time, the above specification is completed by considering an Ornstein-Uhlenbeck process (Uhlenbeck & Ornstein 1930). In particular, this process can be also seen as the solution to the stochastic differential equation

$$\mathrm{d}Y_t = -\alpha(Y_t - \mu)\mathrm{d}t + \sqrt{\frac{2\alpha}{\lambda}}\, \mathrm{d}W_t,$$

for $\mu \in \mathbb{R}$, $\alpha, \lambda > 0$ and where $(W_t)_{t \geq 0}$ denotes the standard Brownian motion. This process, apart from being stationary, reversible and Markovian is also Gaussian, allowing to study the independent case through a suitable choice of its autocorrelation. Following Karatzas & Shreve (1988), it can be easily seen that the invariant and transition densities

for this process, setting $\phi := e^{-\alpha}$, are given by

$$\begin{aligned}\pi(y_t; \mu, \lambda) &= \mathrm{N}(y_t; \mu, 1/\lambda)\\ p(y_0, y_t; \mu, \lambda, \phi) &= \mathrm{N}(y_t; y_0\phi^t + \mu(1-\phi^t), (1-\phi^{2t})/\lambda),\end{aligned}$$

where $\mathrm{N}(y; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2. Thus, its mean and variance are given by $\mathbf{E}[Y_t] = \mu$ and $\mathrm{Var}[Y_t] = 1/\lambda$ for all t, respectively. Moreover, its correlation is given by $\mathrm{Cor}(Y_s, Y_t) = \phi^{t-s}$ for all $0 \leq s \leq t$, with $0 < \phi < 1$. Therefore, the parameter ϕ controls the correlation, and the case where observations are independent and normally distributed is obtained when $\phi \to 0$. To refer to the law of such a Gaussian process, the notation $\mathrm{OU}(\mu, \lambda, \phi)$ will be used.

Assuming data y are observed at times $t_1, \ldots, t_n$, the likelihood function of Υ_n and $x_j = (\mu_j, \lambda_j, \phi_j)$, $1 \leq j \leq n$, is given by

$$\mathbf{P}(y|\Upsilon_n = (n_1, \ldots, n_k), x = (x_1, \ldots, x_k)) = \prod_{j=1}^{k} \frac{\lambda_j^{n_j/2}}{(2\pi)^{n_j/2}} \exp\left\{ -\frac{\lambda_j}{2}(\mathbf{y}_j - \boldsymbol{\mu}_j)'\mathbf{S}_{\phi,j}(\mathbf{y}_j - \boldsymbol{\mu}_j)' \right\},$$

for $\mathbf{y}_j = (y_{j,1}, y_{j,2}, \ldots, y_{j,n_j})$, $\boldsymbol{\mu}_j = (\mu_j, \ldots, \mu_j) \in \mathbb{R}^{n_j}$, and where $\mathbf{S}_{\phi,j} \in \mathbb{R}^{n_j \times n_j}$ is such that

$$\mathbf{S}_{\phi,j} = \begin{pmatrix} \frac{1}{1-\phi^{2\delta_1}} & \frac{-\phi^{\delta_1}}{1-\phi^{2\delta_1}} & 0 & \cdots & 0 & 0 \\ \frac{-\phi^{\delta_1}}{1-\phi^{2\delta_1}} & \frac{1-\phi^{2\delta_1^2}}{(1-\phi^{2\delta_1})(1-\phi^{2\delta_2})} & \frac{-\phi^{\delta_2}}{1-\phi^{2\delta_2}} & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & \frac{1-\phi^{2\delta_{n_j-2}^2}}{(1-\phi^{2\delta_{n_j-2}})(1-\phi^{2\delta_{n_j-1}})} & \frac{-\phi^{\delta_{n_j-1}}}{1-\phi^{2\delta_{n_j-1}}} \\ 0 & 0 & 0 & \cdots & \frac{-\phi^{\delta_{n_j-1}}}{1-\phi^{2\delta_{n_j-1}}} & \frac{1}{1-\phi^{2\delta_{n_j-1}}} \end{pmatrix},$$

with $\delta_i = t_{i+1} - t_i$ and $\delta_i^2 = t_{i+2} - t_i$. The model is completely specified after a prior ν_0 is assigned to parameters $(\mu_j, \lambda_j, \phi_j)$, $i = 1, \ldots, k$, either to integrate them out or to infer about them.

Cox-Ingersoll-Ross model. The Ornstein-Uhlenbeck process is a good choice for modeling $\mathbb{R}$-valued data. However, under certain contexts, it could be more convenient or necessary using a different stochastic process. For example, a Cox-Ingersoll-Ross (CIR) model (Cox et al. 1985) can be applied in interest-rate problems or others having $\mathbb{R}^+$-valued observations. A CIR model can be seen as the solution to the following stochastic

differential equation

$$\mathrm{d}Y_t = c\left(\frac{a}{b} + Y_t\right)d_t + \sqrt{\frac{2c}{b}Y_t}\mathrm{d}W_t,$$

for $a, b, c > 0$ and where $(W_t)_{t\geq 0}$ denotes the standard Brownian motion. This model can also be expressed through a stationary Gamma-Poisson process (see, for example Mena & Walker 2009), where its invariant and transition densities are

$$\begin{aligned}\pi(y_t; a, b) &= \mathrm{Ga}(y_t; a, b)\\ p(y_0, y_t; a, b, c) &= \sum_{x=0}^{\infty} \mathrm{Ga}(y_t; x + a, \phi_t + b)\mathrm{Poi}(x; y_0\phi_t)\\ &= \frac{\exp\left\{-\left(\phi_t(y_t + y_0) + by_t\right)\right\}}{(\phi_t + b)^{-(a+1)/2}\phi_t^{(a-1)/2}}\left\{\frac{y_t}{y_0}\right\}^{\frac{a-1}{2}} I_{a-1}\left(2\sqrt{y_t y_0\phi_t(\phi_t + b)}\right),\end{aligned}$$

for $\phi_t = b/(e^{ct} - 1)$. Here, $\mathrm{Ga}(y; a, b)$ denotes the gamma density with mean a/b, $\mathrm{Poi}(x)$ denotes the Poisson density with parameter x, and $I_\nu(x)$ is the modified Bessel function of the first kind with index ν, that is

$$I_\nu(x) = \sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m + \nu + 1)}\left(\frac{x}{2}\right)^{2m+\nu}.$$

The autocorrelation is controlled by the parameter ϕ_t.

Under this setting, the likelihood function (17) is given by

$$\prod_{j=1}^{n_j} \frac{b_j^{a_j}(y_{t_{j,1}} y_{t_{j,n_j}})}{\Gamma(a_j)} \exp\left\{-b_j \sum_{l=1}^{n_j} y_{j,t_l} - \sum_{l=1}^{n_j-1} \phi'_{j,l}(y_{j,t_{l+1}} + y_{j,t_l})\right\} \times \prod_{l=1}^{n_j-1} \frac{(\phi'_{j,l} + b_j)^{(a_j+1)/2}}{\phi'^{(a_j-1)/2}_{j,l}} I_{a_j-1}\left(2\sqrt{y_{j,t_{l+1}} y_{j,t_l}\phi'_{j,l}(\phi'_{j,l} + b_j)}\right),$$

where the driving parameters are $x_j = (a_j, b_j, c_j)$, $1 \leq k \leq n$, and where

$$\phi'_{j,l} = \frac{b_j}{e^{c_j(t_{l+1} - t_l)} - 1}.$$

Wright-Fisher model. Another stochastic process is the one known as the Wright-Fisher model, an appealing option when modeling $(0, 1)$-valued data. This can be seen

as the solution to

$$\mathrm{d}Y_t = \frac{1}{2}\left(a(1-Y_t) + bY_t\right)\mathrm{d}t + \sqrt{Y_t(1-Y_t)}\mathrm{d}W_t,$$

for $a, b > 0$ and where $(W_t)_{t\geq 0}$ denotes the standard Brownian motion. In this case, we have that its invariant and transition densities are given by

$$\pi(y_t; a, b) = \mathrm{Be}(y_t; a, b)$$
$$p(y_0, y_t; a, b) = \sum_{m=0}^{\infty} e^{-m(m+a+b-1)t/2}\omega_m(y_0)\omega_m(y_t)\pi(y_t; a, b),$$

where

$$\omega_m(y) = c_m(a)_{m\uparrow}\sum_{j=0}^{m}\binom{m}{j}\frac{(m+a+b-1)_{j\uparrow}}{(a)_{j\uparrow}}(-y)^j,$$

with

$$c_m = \sqrt{\frac{a+b+2m-1}{a+b+m-1}\frac{(a+b)_{m\uparrow}}{(a)_{m\uparrow}(b)_{m\uparrow}}\frac{1}{m!}},$$

and $(x)_{n\uparrow}$ the Pochhammer symbol (Griffiths & Spanò 2010).

2.2 Inferences

Even when the number of segmentations $\mathcal{S}_{[n]}$ is much smaller than the number of partitions $\mathcal{P}_{[n]}$, simulating from the posterior distribution of Υ_n becomes unfeasible for bigger datasets. The need of an MCMC algorithm is evident. At this point, we focus on the posterior distribution of Υ_n given by

$$\mathbf{P}(\Upsilon_n|y) \propto \binom{n}{n_1, \ldots, n_k}\frac{1}{k!}\Pi_k^{(n)}(n_1, \ldots, n_k)\prod_{j=1}^{k}\int_{\mathbb{X}}\pi(y_{j,1}; x)\prod_{l=1}^{n_j-1}p(y_{j,l}, y_{j,l+1}; x)\nu_0(\mathrm{d}x),$$

where $\Pi_k^{(n)}$ is an EPPF and (π, p, ν_0) are defined according to the selected stochastic process. The case where the driving parameters are not integrated out will be discussed in further sections. On the other hand, due to the nature of the support of Υ_n, estimates like mean, variance or quantiles are not straightforward to obtain or they might have no a clear interpretation. Then, it is also necessary to find an adequate estimate.

2.2.1 MCMC algorithm

The identification of multiple change points might bring several computational issues. For instance, Green (1995) makes use of reversible jump MCMC to draw posterior inferences on the number of change points. Concretely, he focuses on change points among data modeled through a non-homogeneous Poisson process with a step function as intensity and where the number of steps, namely the number of regimes, is determined via a birth and death mechanism. Therefore, this approach requires the construction of trans-dimensional samplers which, depending on the choice of the model, do not always lead to a simple way to disentangle the prior and posterior mass assigned to a given change-point structure.

However, an appealing MCMC algorithm which can be adapted for change-point detection is the one used in Fuentes-García et al. (2010*b*). Concretely, their algorithm belongs to the class of split-merge algorithms and updates the variables k and $(n_j)_{j=1}^k$ by using Metropolis-Hastings steps. Note that unlike other split-merge MCMC algorithms on the space of partitions, the *split* and *merge* steps make complete sense here as the time-ordering restriction leaves no other sensible moves. More specifically, two possible choices are available at each iteration: a split, which creates a new regime, or a merge, which combines two consecutive existing regimes into a single one. After that, a random pair of adjacent regimes is updated by proposing new sizes in order to speed the sampler up. In next paragraphs, this complete procedure is explained in detail.

Let $p(k, \psi_k) := p(n_1, \dots, n_k | y)$, where ψ_k is a segmentation of the data y having k blocks, and each block has size n_j, $j = 1, \dots, k$. Then, at each iteration, the MCMC algorithm updates values for k and ψ_k as follows. The support for k is $\{1, \dots, n\}$; when updating it, the chain will change the segmentation from one with $k^{(i)}$ blocks to another with $k^{(i+1)}$, involving perhaps a change in dimension. To avoid this latter issue, a series of latent variables, ψ_j for $j = 1, \dots, n$ and $j \neq k$, is introduced (Godsill 2001). The resulting density is written as

$$p(k, \psi) = p(k, \psi_k) \prod_{j=1}^{k-1} p(\psi_j | \psi_{j+1}) \prod_{j=k+1}^{n} p(\psi_j | \psi_{j-1}). \tag{18}$$

Updating k is, then, done via a Metropolis-Hastings step with target distribution $p(k|\psi)$. Let $0 < q < 1$ and let

$$\check{p}(k|r) = q\, \mathbf{1}(k = r + 1) + (1 - q)\, \mathbf{1}(k = r - 1) \tag{19a}$$

be the proposal distribution of k, whenever $1 < r < n$, and

$$\check{p}(2|1) = \check{p}(n-1|n) = 1, \tag{19b}$$

otherwise. Thus, the acceptance probability to update k is given by

$$\alpha = \min\left\{\frac{\check{p}(k|k')}{\check{p}(k'|k)}\frac{p(k'|\psi)}{p(k|\psi)}, 1\right\},$$

with k' simulated from the proposal (19).

For a given k, there are only two possible values for updating it: $k+1$ and $k-1$; namely split and merge moves, respectively. These moves are related to the conditional distributions in (18), since $p(\psi_{k+1}|\psi_k)$ represents a split move and $p(\psi_{k-1}|\psi_k)$ a merge. So it is necessary to be able to simulate from them.

When a split is performed, one block with size greater than one is selected and then split into two. These two choices are made uniformly, so

$$p(\psi_{k+1}|\psi_k) = \frac{1}{n_{g,k}(n_s - 1)}, \tag{20}$$

where $n_{g,k}$ is the number of blocks of size greater than one and n_s the size of the selected block. On the other hand—when a merge is performed—two adjacent blocks are selected to be merged. Again, they are chosen uniformly, so

$$p(\psi_{k-1}|\psi_k) = \frac{1}{k-1}. \tag{21}$$

Therefore, updating k is done as follows. When a split move is proposed, i.e. $k' = k+1$, Equations (18–21) simplify α to

$$\alpha = \min\left\{\frac{1-q}{q}\frac{p(k+1,\psi_{k+1})}{p(k,\psi_k)}\frac{n_{g,k}(n_s-1)}{k}, 1\right\}, \tag{22a}$$

whenever $1 < k < n$, and

$$\alpha = \min\left\{(1-q)\,(n-1)\frac{p(2,\psi_2)}{p(1,\psi_1)}, 1\right\}, \tag{22b}$$

if $k = 1$. Otherwise, a merge move is proposed, i.e. $k' = k-1$. Then two adjacent blocks

are selected, say s and $s+1$, and Equations (18–21) simplify α to

$$\alpha = \min\left\{\frac{q}{1-q}\frac{p(k-1,\psi_{k-1})}{p(k,\psi_k)}\frac{k-1}{n_{g,k-1}(n_s+n_{s+1}-1)},1\right\}, \tag{23a}$$

whenever $1 < k < n$, and

$$\alpha = \min\left\{q\,(n-1)\frac{p(n-1,\psi_{n-1})}{p(n,\psi_n)},1\right\}, \tag{23b}$$

when $k = n$.

A second step—called *shuffle*—is included in the algorithm in order to improve it. This step takes two adjacent blocks, say s and $s+1$, and updates their sizes uniformly splitting the combined block into two, of sizes n_s^* and n_{s+1}^*, with $n_s^*, n_{s+1}^* \geq 1$. In this case, the acceptance probability is given by

$$\alpha = \min\left\{\frac{p(k,\psi_k^*)}{p(k,\psi_k)}\frac{(n_s^*+n_{s+1}^*-1)}{(n_s+n_{s+1}-1)},1\right\} = \min\left\{\frac{p(k,\psi_k^*)}{p(k,\psi_k)},1\right\}. \tag{24}$$

After these steps, additional variables can be updated, e.g. those related to the prior distribution of Υ_n or to the likelihood function. Figure 2 summarizes this MCMC scheme.

2.2.2 Point estimates

As explained at the beginning of this chapter, two estimates are of interest in change-point analysis: the location of change points and their number. The former is a more delicate issue, but the latter has already studied in Section 1. Therefore, the random variable C_n—modeling the number of change points—is defined as

$$C_n := B_n - 1,$$

where B_n models the number of blocks under an $\mathcal{S}_{[n]}$-valued approach.

On the other hand—due to the the particular structure of the set $\mathcal{S}_{[n]}$—providing an estimate of the location of change points is not straightforward. As mentioned before, estimates of the random segmentation Υ_n like the mean might be difficult to obtain and others like the variance or quantiles could not have an obvious interpretation. Among this kind of statistics, the mode may be the most useful and easiest to interpret. However, in the literature, there are some attempts to provide meaningful $\mathcal{P}_{[n]}$- or $\mathcal{S}_{[n]}$-valued estimates. In particular, two estimates are explained below: the first one is related with the

read $y_1, \dots, y_n$ and hyper-parameters
initiate k and $(n_1, \dots, n_k)$
repeat
  with probability $q\,\mathbb{I}(1 < k < n) + \mathbb{I}(k = 1)$ ▷ split
    choose j uniformly from $\{j : 1 \le j \le k, n_j > 1\}$
    choose l uniformly from $\{1 \dots, n_j - 1\}$
    with probability α in (22)
      $(n_1, \dots, n_{k+1}) \leftarrow (n_1, \dots, n_{j-1}, l, n_j - l, n_{j+1}, \dots, n_k)$
      $k \leftarrow k + 1$
    end with
  otherwise ▷ merge
    choose j uniformly from $\{1, \dots, k - 1\}$
    with probability α in (23)
      $(n_1, \dots, n_{k-1}) \leftarrow (n_1, \dots, n_{j-1}, n_j + n_{j+1}, n_{j+2}, \dots, n_k)$
      $k \leftarrow k - 1$
    end with
  end with
  if $k > 1$ **then** ▷ shuffle
    choose i uniformly from $\{1, \dots, k - 1\}$
    choose j uniformly from $\{1, \dots, n_i + n_{i+1} - 1\}$
    with probability α in (24)
      $n_{i+1} \leftarrow n_i + n_{i+1} - j$
      $n_i \leftarrow j$
    end with
  end if
  [simulate values for additional parameters]
until converge and have the required number of samples
write the sampled values of k and $(n_1, \dots, n_k)$ [and additional parameters]

FIGURE 2: Split-merge MCMC algorithm used for change point detection.

idea of a 'mean partition', whereas the second one makes use of the marginal probabilities that each instant is a change point.

Least-squares clustering. The least-squares clustering (Dahl 2006) emulates a 'mean partition'. It is defined as the partition ρ^* minimizing the sum of squared deviations of the association matrix $\delta(\rho)$ from $\hat{\rho}$.

Let ρ be a partition, then the association matrix $\delta(\rho)$ is such that its (i, j) element, $\delta_{i,j}(\rho)$, is the indicator of whether observations y_i and y_j, $i, j = 1, \dots, n$, belong to the same block in ρ. Given a sample $P = (\rho^{(1)}, \dots, \rho^{(m)})$ of m partitions, $\hat{\rho}$ is defined as the element-wise mean of the association matrices $\delta(\rho^{(1)}), \dots, \delta(\rho^{(m)})$. Thus, the least-squares clustering ρ^* is such that

$$\rho^* = \arg\min_{\rho \in P} \sum_{i,j} \left(\delta_{i,j}(\rho) - \hat{\rho}_{i,j}\right)^2.$$

Even though this estimate is defined for $\mathcal{P}_{[n]}$-valued partitions, it can be also used with segmentations.

Marginal probability to be change point. Another method used by Loschi & Cruz (2005) to obtain a posterior estimate of Υ_n is based on the posterior probability that each time t_i, $i = 2, \dots, n$, is a change point. It has already stated that every segmentation with k blocks induces a unique sequence of change points $\langle t_1, \dots, t_{k-1} \rangle$. So, define T_i as the event where the ith instant t_i is a change point for $i = 2, \dots, n$, then

$$\zeta_i := \mathbb{P}(T_i | y) = \sum_{\langle j_1, \dots, j_{k-1} \rangle \in S} \mathbb{P}(\Upsilon_n = \langle j_1, \dots, j_l = t_i, \dots, j_{k-1} \rangle | y), \tag{25}$$

with S the sampled segmentations. Using this, Loschi and Cruz's estimate, denoted by $\Upsilon_n^\star$, is given by all the times $(t_{j_1}, t_{j_2}, \dots, t_{j_r})$ whose probabilities ζ_{j_i}, $i = 1, \dots, r$, are greater than certain threshold; hence $\Upsilon_n^\star = \langle t_{j_1}, \dots, t_{j_r} \rangle$. While this approach might be appealing in monitoring contexts, it is not clear how to define a threshold for change point determination. Having said this, within the context at issue, this estimate is useful to illustrate the role prior specifications.

2.3 Sensitivity analysis

Once Model (14) has been fully specified, it will be tested under different scenarios. There are two important aspects to be studied: (i) the role of the parameters related to the prior distribution of Υ_n and the likelihood function, and (ii) the model's ability to detect different structural changes. These two aspects are closely related, so there might happen that the ability to detecting certain type of changes depends on the prior specification.

The prior distribution for Υ_n is the one induced by the two-parameter Poisson-Dirichlet process with $0 \leq \sigma < 1$ and $\theta > -\sigma$, i.e.

$$p'(n_1, \dots, n_k) = \frac{n!}{k!} \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1\uparrow}} \prod_{j=1}^{k} \frac{(1 - \sigma)_{n_j - 1\uparrow}}{n_j!}$$

(Equation 13c). The election of these parameter ranges corresponds to the case of always having new change points as the sample increases (see De Blasi et al. (2015) for more details). Regarding the likelihood function, we consider the marginal likelihood induced by the Ornstein-Uhlenbeck process where the observations were recorded at equally-spaced times, using the simpler notation $t_i := i$ for $i = 1, \dots, n$. The parameters (μ_j, λ_j) were in-

tegrated out using a Normal-Gamma prior $\mathrm{N}(\mu_j; 0, (c\lambda_j)^{-1})\mathrm{Ga}(\lambda_j; a, b)$ and using a single correlation among regimes: $\phi_j = \phi$. Therefore, the likelihood function is written as

$$\mathbb{P}(y|\Upsilon_n) = \prod_{j=1}^{k} \frac{(2b(1-\phi^2))^a \Gamma(n_j/2+a)}{\pi^{n_j/2}\Gamma(a)} \left(\frac{c(1+\phi)(1-\phi^2)}{c+n_j-\phi(n_j-c-2)}\right)^{1/2} \times \left\{ \mathbf{y}_j' \mathbf{S}_j \mathbf{y}_j - \frac{(1-\phi)\left(\sum_{i=1}^{n_j} y_{j,i} - \phi \sum_{i=2}^{n_j-1} y_{j,i}\right)^2}{c+n_j-\phi(n_j-c-2)} + 2b(1-\phi^2) \right\}^{-(n_j/2+a)}, \quad (26)$$

where $\mathbf{y}_j = (y_{j,i} : i = 1 \ldots, n_j)$ and $\mathbf{S}_j \in \mathbb{R}^{n_j \times n_j}$ is given by

$$\mathbf{S}_j = \begin{pmatrix} 1 & -\phi & 0 & \ldots & 0 \\ -\phi & 1+\phi^2 & -\phi & \ldots & 0 \\ 0 & -\phi & 1+\phi^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

The structure of the simulated datasets involves changes in mean, in variance, in correlation, as well as a combination of all them. First, to avoid errors introduced by MCMC algorithms, small datasets will be used and—in second place—bigger datasets will be used in order to study the performance of the proposed MCMC method. For the small dataset, data are assumed to be independent and normally distributed—corresponding to the case $\phi = 0$—since dependence is hard to reproduce with a few observations. For the bigger datasets, however, there is not such an issue.

Consider a small dataset where 15 observations are simulated as follows:

$$y_i \sim \mathrm{N}(0, 0.5) \quad i = 1, \ldots, 6; \qquad y_i \sim \mathrm{N}(2, 0.5) \quad i = 7, \ldots, 15$$

(see Figure 3). The prior specification for the likelihood function was fixed to $a = b = 1$, $c = 0.1$ and $\phi = 0$. Regarding the prior parameters for Υ_n, particular interest is in parameter σ; it was set to 0.0, 0.1, 0.3, 0.6 and 0.9 and then the corresponding values for θ were found such that the prior expected value for the number of change points, C_n, matches 1, 5 and 11. The results are shown in Table 2.

The most illustrative results in this example are those when the prior guess at the number of change points is far from the number that generated the data. As σ increases, the probability assigned to the posterior mode of Υ_n increases, and, at the same time, the

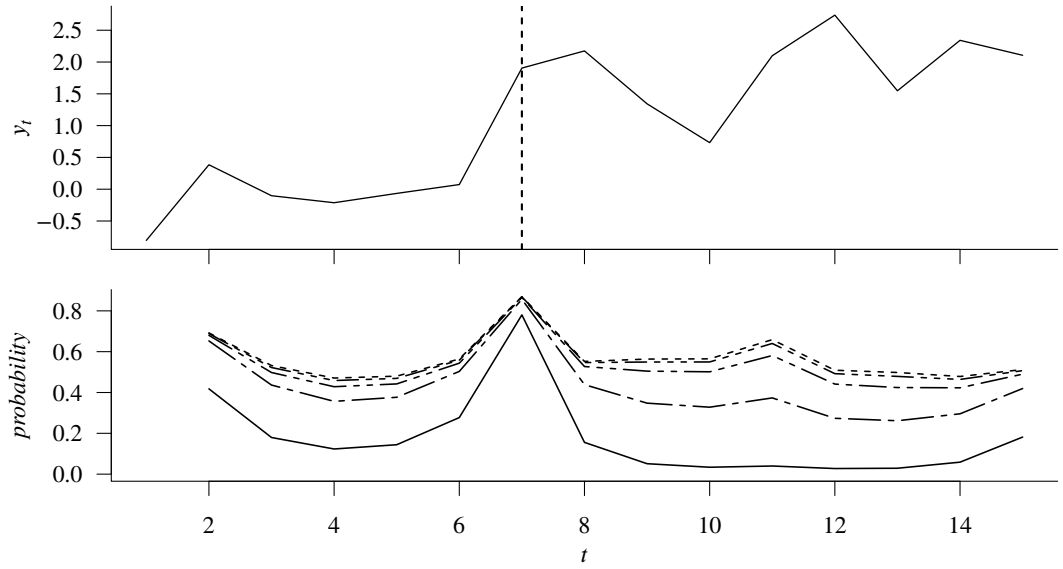


FIGURE 3: (top) Small simulated dataset with change point at ⟨7⟩, indicated with a dashed line. (bottom) Posterior probabilities that each time is a change point assuming $\mathbf{E}[C_n] = 11$ *a priori* and taking different values of σ: ----· 0.0, ---– 0.1, --— 0.3, -— 0.6 and —— 0.9. Points are connected by straight lines for visual simplification.

$\mathbf{E}[C_n]$	σ	θ	$\tilde{\Upsilon}_n$	*prob.*	$\tilde{C}_n$	*prob.*
1	0.0	0.356	⟨7⟩	0.6985	1	0.7986
	0.1	0.194	⟨7⟩	0.6788	1	0.7776
	0.3	−0.114	⟨7⟩	0.6301	1	0.7254
	0.6	−0.531	⟨7⟩	0.5026	1	0.5868
	0.9	−0.890	⟨⟩	0.8229	0	0.8229
5	0.0	3.201	⟨7⟩	0.1457	2	0.3293
	0.1	2.626	⟨7⟩	0.1749	2	0.3413
	0.3	1.527	⟨7⟩	0.2507	2	0.3508
	0.6	0.097	⟨7⟩	0.3923	1	0.4581
	0.9	−0.822	⟨⟩	0.3564	1	0.3630
11	0.0	25.683	⟨2, 3, 4, 5, 6, 7⟩	0.0017	8	0.2168
	0.1	22.670	⟨2, 3, 4, 5, 6, 7⟩	0.0025	8	0.2103
	0.3	16.672	⟨2, 3, 4, 5, 6, 7⟩	0.0052	7	0.1951
	0.6	7.832	⟨2, 3, 4, 5, 6, 7⟩	0.0176	6	0.1799
	0.9	0.087	⟨7⟩	0.2512	1	0.3179

TABLE 2: Posterior results for the small dataset. The first three columns show the prior parameters for Υ_n. Columns 4 and 5 show the posterior mode, $\tilde{\Upsilon}_n$, together with its probability. The last two columns show the mode for the number of change points, $\tilde{C}_n$, and its corresponding probability.

posterior mode of C_n shifts towards smaller values. Also observe that the posterior mode of Υ_n contains fewer groups as σ increases. In a similar way, the posterior probabilities τ_i, $i = 2, \dots, 15$ (Equation 25)—shown in Figure (3) for the case $\mathbf{E}[C_n] = 11$—decrease on all times, except the correct one, as σ increases.

As a second part, the performance of the MCMC algorithm of Section 2.2.1 is studied using three simulated datasets. Concretely, the first dataset has changes in mean, the second has changes in variance and the last one has changes in mean, variance and correlation. Each of them consist of 150 observations and they are described next.

1. The first dataset has two changes in mean. This is sequentially simulated from different Ornstein-Uhlenbeck processes as follows:

$$\begin{aligned} y_i &\sim \mathrm{OU}(0, 2, 0.4) \quad i = 1, \dots, 50 \\ y_i &\sim \mathrm{OU}(5, 2, 0.4) \quad i = 51, \dots, 85 \\ y_i &\sim \mathrm{OU}(2, 2, 0.4) \quad i = 86, \dots, 150. \end{aligned}$$

   So its change points are $\langle 51, 86 \rangle$ (Figure 4a).

2. The second dataset has two changes in variance. Like the previous dataset, it is sequentially simulated as follows:

$$\begin{aligned} y_i &\sim \mathrm{OU}(0, 1.5, 0.7) \quad i = 1, \dots, 50 \\ y_i &\sim \mathrm{OU}(0, 0.2, 0.7) \quad i = 51, \dots, 120 \\ y_i &\sim \mathrm{OU}(0, 0.8, 0.7) \quad i = 121, \dots, 150. \end{aligned}$$

   Its change points are $\langle 51, 121 \rangle$ (Figure 4b).

3. The third dataset has two change points and it is sequentially simulated as follows:

$$\begin{aligned} y_i &\sim \mathrm{OU}(0, 0.5, 0.1) \quad i = 1, \dots, 50 \\ y_i &\sim \mathrm{OU}(-1, 2, 0.7) \quad i = 51, \dots, 100 \\ y_i &\sim \mathrm{OU}(0, 1, 0.2) \quad i = 101, \dots, 150. \end{aligned}$$

   Then, its change points are $\langle 51, 101 \rangle$ (Figure 4c). This is a more complex dataset since has changes in mean, variance and also in correlation. Even when the model uses a single parameter, ϕ, to control the correlation in the data, this dataset is included to test how well the model works. In particular, the number of change points

is expected to be different from the above, as, in principle, more of these might be needed to represent the above data under similar intra-regime dependency scenarios.

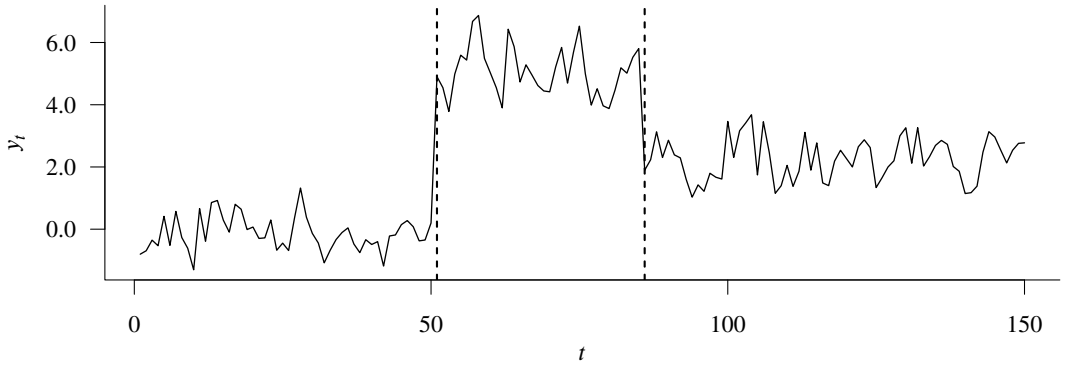
Various simulations were done using different specifications, each of them consisted of 20,000 iterations after 10,000 of burn-in with the likelihood's prior parameters set as $a = b = 1$ and $c = 0.1$. The independent and dependent cases were considered by letting $\phi = 0$ and $\phi \sim \mathrm{U}(0, 1)$, respectively. Regarding the prior parameters for Υ_n, like in the small dataset, parameter σ was set to 0.1, 0.3, 0.6 and 0.9 and their corresponding values for θ were found such that the prior expected value for C_n matches 2, 49 and 99. Additionally, some specifications were considered where $\theta|\sigma \sim \mathrm{Ga}(1, 1, -\sigma)$ with $\sigma \sim \mathrm{Be}(1, 1)$ or fixed taking the aforementioned values; see Appendix A for details about the posterior distributions of these parameters and their inclusion in the MCMC algorithm.

The results are shown in Tables 4, 5 and 6. From these, changes in mean are correctly detected for almost all configurations (Table 4). Both—the location and the number of the change points—were correctly identified. The most notable difference related to the performance of ϕ is that the independent case assigns lower probabilities than those corresponding to the dependent case. Regarding the parameter σ, it can be seen that under a complete misspecification of the prior, i.e. $\mathbf{E}[C_n] = 49$ and $\sigma = 0.1$, neither the nor the number of change points were correctly identified. However, as σ increases the results become more favorable. Indeed, by inspecting more extreme cases, e.g. $\mathbf{E}[C_n] = 99$, only for $\sigma = 0.9$, the correct number of change points are recovered.

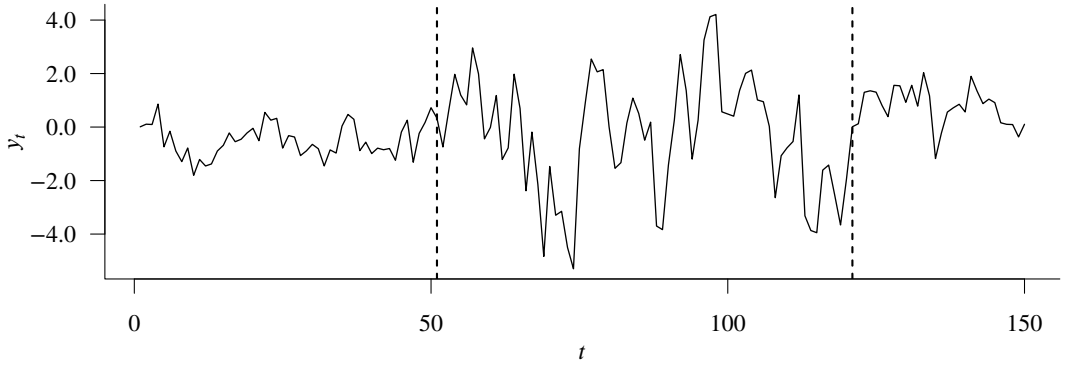
The second example, reported in Table 5, shows the improvement under a more robust model, i.e. allowing for the dependency by setting $\phi \sim \mathrm{U}(0, 1)$. In such scenario, most change points were correctly detected. Indeed, the small variations might be due to the randomness inherent to the simulation. In the independent case, however, even when the prior guess at the number of change points was close enough to the correct one, the model was unable to correctly identify them. This clearly favors the use of a model able to capture serial dependence. On the other hand, parameter σ shows a similar performance as in the previous dataset. However, in this case the cost of placing a misspecified prior, e.g. $\mathbf{E}[C_n] = 49$, tends to be higher.

For the third dataset, which has change points in mean, variance and correlation, the results are favored under the dependent case. Having similar sensitivity conclusions for the parameter σ. See Table 6.

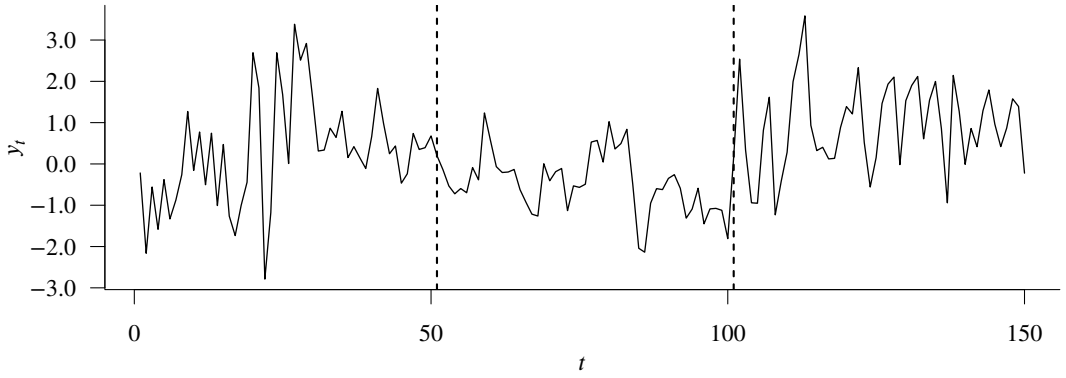
Throughout this analysis, the modal partition was used as indicative of change points, however, as mentioned before, the least-squares clustering could be also employed. In-



(A) Changes in mean at times $\langle 51, 86 \rangle$.



(B) Changes in variance at times $\langle 51, 121 \rangle$.



(C) Changes in mean, variance and correlation at times $\langle 51, 101 \rangle$.

FIGURE 4: Simulated datasets; change points are indicated with dashed lines.

deed, evaluating these two statistics, the resulting change points coincided in almost all scenarios; for the sake of completeness, Table 7 shows both estimates for some cases.

2.3.1 Comparison with a product partition model

In order to have a better picture of our proposed model, its results were compared with the ones obtained by using an existing model for change-point detection. Concretely, a PPM-based approach with Yao's cohesion function, independent inter- and intra-regimes and Gaussian likelihood is used. This method—developed by Loschi & Cruz (2005)—has been widely used in change-point detection in financial data; see, for example Zantedeschi et al. (2011). Within this approach, the prior distribution of Υ_n is given by

$$\mathbf{P}(\Upsilon_n = \{\tau_1, \ldots, \tau_k\}) = p^{k-1}(1-p)^{n-k},$$

for some $0 < p < 1$, whereas the distribution of the number of blocks is

$$\mathbf{P}(B_n = k) = \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}, \qquad k = 1, \ldots, n$$

which corresponds to a binomial distribution with parameter p. It is worth saying that this distribution over segmentations—according to Barry & Hartigan (1992)—belongs to the second approach we studied in Section 1.2, therefore, it is also a segmentation model. Furthermore—for the Gaussian case—the likelihood function can be written as (26) with $\phi = 0$. Sampling from the corresponding posterior distribution is done in Loschi & Cruz (2005) by using an extension of the Gibbs sampling scheme proposed by Barry & Hartigan (1993); see also Loschi et al. (2003). In our view, this is one of the most competitive models for change point problems available in the literature, apart of having certain features similar to our approach.

The simulations for this model were also done using different scenarios, taking $20,000$ iterations after a burn-in of $10,000$. Regime parameters and cohesion function's hyper-parameters, (α, β), were taken as it is described in their paper. Furthermore, hyper-parameters (α, β) were fixed such that the expected number of change points matches the same values as above: 2, 49 and 99.

The results are shown in Table 8. Their model performs well for the dataset with changes in mean; hyper-parameters do not affect the results. For the dataset with changes in variances, an extra change point at $t = 4$ is detected in all scenarios; additionally, there are some scenarios including more change points. However, for the third dataset, their

method could not detect any change point. This behavior might be due to the fact that their prior distribution $\mathbf{P}(\Upsilon_n)$ is function of n and k only, that is, giving the same weight to all compositions $(n_1, \ldots, n_k)$ for a given k. So, it could be said that the proposed model has an advantage over theirs. The distribution of Υ_n based on EPPFs provides more flexibility to detect change points. In particular, the parameter σ plays a key role since it enhances the detection under more uncertain scenarios. Moreover, this distribution may be also used as an alternative product distribution within PPMs restricted to $\mathcal{S}_{[n]}$.

2.4 Application to real data

As mentioned before, one of the areas where change point models are of great interest is financial econometrics. For example, they can help to establish financial markers for global financial crisis (Allen et al. 2013), to detect the *de facto* exchange rate regime in operation of a particular central bank (Reinhart & Rogoff 2004, Bubula & Ötker-Robe 2002), to measure business cycles (Harding & Pagan 2008), etc. In order to illustrate our proposal, we analyze the exchange rate between US Dollars and Mexican Pesos during the period of January 2007–December 2012 (Figure 5), available at the website `www.federalreserve.gov`. Estimations are based on 20,000 iterations after a burn-in of 10,000. Likelihood's parameters were set as before—i.e. $a = b = 1$ and $c = 0.1$—and, for ϕ, the dependent ($\phi \sim \mathrm{U}(0, 1)$) and independent ($\phi = 0$) cases were considered. From the analysis made in the previous section, parameter σ was fixed to 0.9 to allow a strong reinforcement and let $\theta \sim \mathrm{Ga}(1, 1, -0.9)$. For simplicity during the comparison, the dependent and independent cases are called Model A and Model B, respectively.

For Model A, the most probable change points, occurring with probability 0.0245, are detected at: *September 10, 2008*, *May 4, 2009*, *July 16, 2010*, *August 4, 2011* and *August 3, 2012*. For Model B, they occur with probability 0.116 and are at: *April 1, 2008*, *October 7, 2008*, *January 14, 2009*, *April 2, 2009*, *November 24, 2009*, *January 3, 2011* and *September 8, 2011*. Figure 5 indicates all these change points for each model and Figure 6 shows posterior probabilities τ_i for $i = 2, \ldots, 1566$ (Equation 25). Figure 7 exhibits the posterior distributions of θ, for both models, and ϕ, for Model A. Additionally, Table 3 displays the posterior distribution of the number of change points.

These results share change points around relevant events with Mexico and USA: 1) the 2008 financial crisis with a sharp downward in September, 2) the flu pandemic suffered by Mexico in the period of March–May 2009, where pharmaceutical industry tried to reactivate global economy, and 3) the US debt-ceiling crisis in 2011. The rest of the change points detected by Model A are also closed to other relevant events: 4) the so called *cur-*

rency war in 2010 as a consequence of the 2008 financial crisis, and 5) the Mexican presidential election.

On the other hand, Loschi and Cruz's proposal was tested under a similar scenario. This model will be called Model C. Its results are also shown in Figures 5 and 6 and in Table 3. In addition, the modal change points were detected, with probability 0.08465, at: *September 9, 2008*, *March 11, 2009*, *March 18, 2011* and *August 25, 2011*.

§ 3 Estimation of regime parameters

After detecting the locations of the change points, there might be interest in additionally providing some estimates regarding the regimes, for example, their governing parameters. However, this is not a straightforward task. Assuming a model like (14) but with the regime parameters, say x, included, the joint posterior of $(\Upsilon_n, x^{(k)})$ is written as

$$\mathbf{P}(\Upsilon_n, x^{(k)}|y) \propto \mathbf{P}(\Upsilon_n)\mathbf{P}(x^{(k)}|\Upsilon_n)\mathbf{P}(y|\Upsilon_n, x^{(k)}).$$

Note that the dimension of parameter x and the number of blocks in Υ_n is the same, say k; this is indicated with the superscript on x. However—if we only focus on the posterior distribution of the regime parameters—we have

$$\mathbf{P}(x|y) \propto \sum_{\Upsilon_n \in \mathcal{S}_{[n]}} \mathbf{P}(\Upsilon_n, x^{(k)}|y) \propto \sum_{k=1}^{n} \sum_{\Upsilon_n \in \mathcal{S}^k_{[n]}} \mathbf{P}(\Upsilon_n)\mathbf{P}(x^{(k)}|\Upsilon_n)\mathbf{P}(y|\Upsilon_n, x^{(k)}),$$

which indicates that the parameter x has different dimensions. This explains the need of introducing methodologies like reversible jump MCMC or transdimensional samplers. Moreover—even when the dimension k is fixed, except for the trivial cases $k = 1$ or $k = n$—it is not possible to provide a point estimate of the regime parameters.

One approach to overcome this issue and provide estimates of regime parameters has been proposed by Barry & Hartigan (1992, 1993). Let $[i, j] := \{i, i + 1, \dots, j\}$ for any $0 < i \leq j$. Under their PPM approach, they compute a 'marginal' posterior distribution for parameter x_j as follows

$$\mathbf{P}(x_j|y) = \sum_{i=1}^{j} \sum_{r=j}^{n} \mathbf{P}(x_j|y, [i, j])r([i, j]|y), \qquad j = 1, \dots, n \tag{27}$$

where $\mathbf{P}(x_j|y, [i, j])$ corresponds to the posterior distribution of x_j given the data y and given that the block $[i, j]$ is part of the segmentation, and $r([i, j]|y)$—known as the pos-

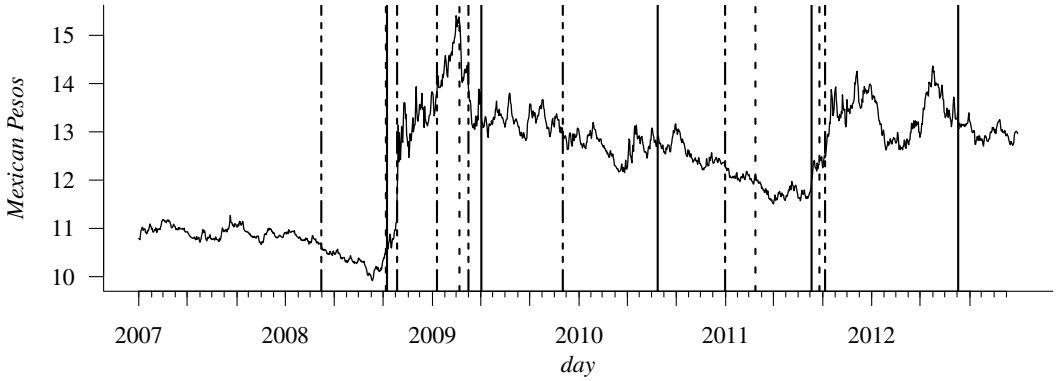


FIGURE 5: Price of one US Dollar in Pesos. Change points are indicated with dashed lines for the different models: A) —— proposal with $\phi \sim U(0,1)$, B) —-- proposal with $\phi = 0$, and C) - - - Loschi and Cruz's approach.

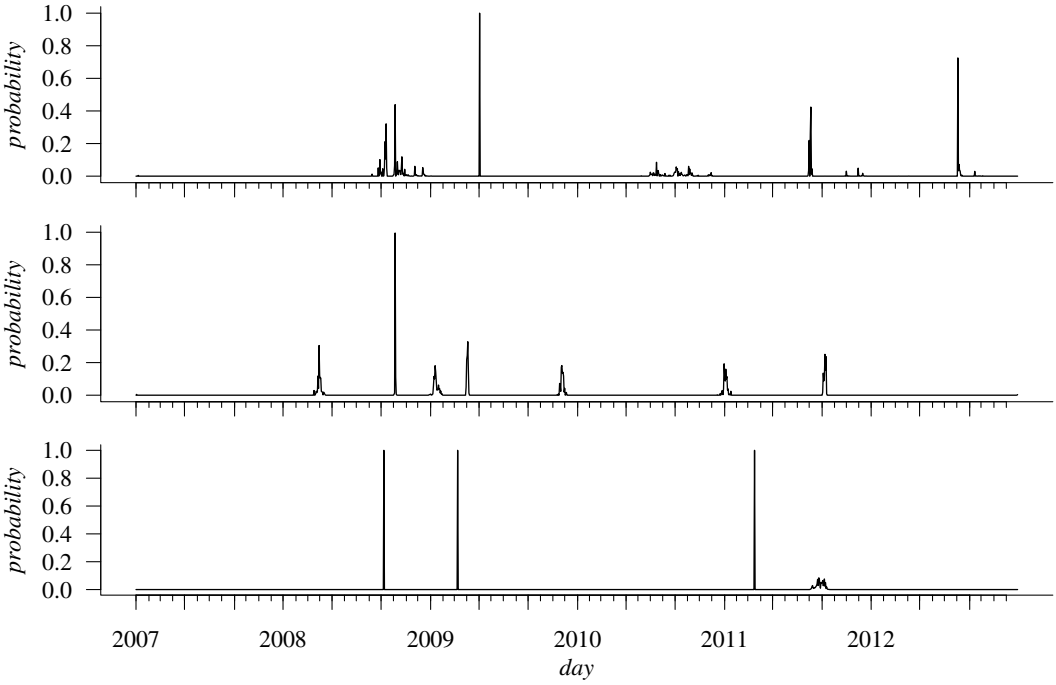
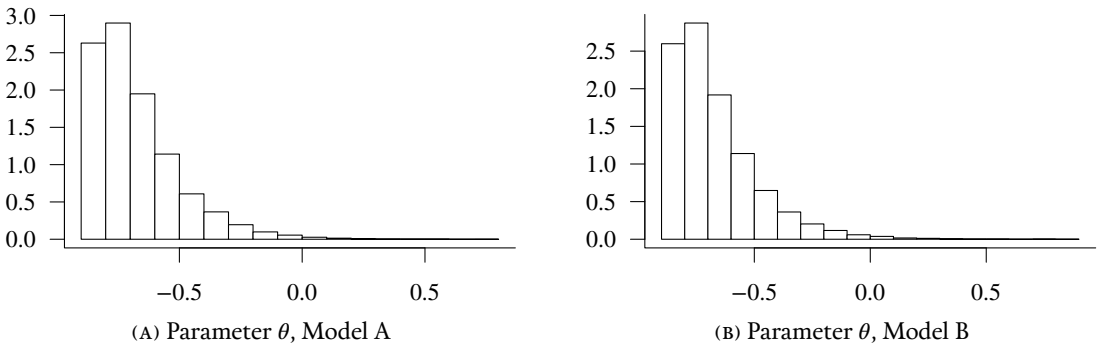


FIGURE 6: Posterior probabilities that each time is a change point for the different models: (top) proposal with $\phi \sim U(0,1)$, (middle) proposal with $\phi = 0$, and (bottom) Loschi and Cruz's approach.

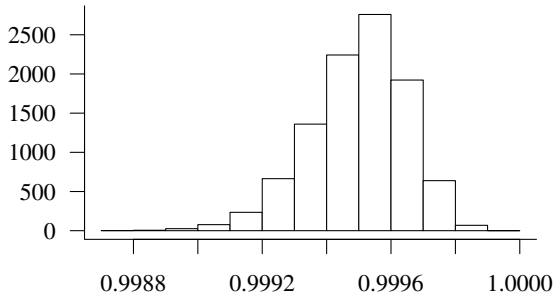
	Number of change points					
Model	4	5	6	7	8	9
proposal with $\phi \sim U(0,1)$	—	0.23265	0.23015	0.49245	0.04365	0.00110
proposal with $\phi = 0$	—	—	—	0.83290	0.15695	0.01015
Loschi and Cruz's	0.00025	0.99975	—	—	—	—

TABLE 3: Posterior probabilities for the number of change points, C_n, for each model.



(A) Parameter θ, Model A

(B) Parameter θ, Model B



(C) Parameter ϕ, Model A

FIGURE 7: Posterior distributions for the proposed model with $\phi \sim U(0,1)$ (Model A) and $\phi = 0$ (Model B).

terior relevance of the block $[i, j]$—is defined as

$$r([i,j]|y) = \sum_{\tau \in \mathcal{S}_{[n]}} \mathbf{P}(\Upsilon_n = \tau|y)\mathbf{1}([i,j] \in \tau),$$

namely, the posterior distribution of the random segmentation restricted to the cases where the block $[i, j]$ appears in the segmentations. Hence, a point estimate of x_j is obtained by taking the expectation of (27).

This estimation approach provides a point estimate for every time point t_j which allows to have a picture about the dynamics of the mean behavior of the parameters. However, there is not yet a clear way to provide a regime estimate.

From this discussion, we can conclude that it is necessary to conditioning to a certain segmentation in order to provide any regime estimate. Using the model of Section 2, this can be done after an estimate for the change-point locations is provided, since observations within each regime are modeled by a specific stochastic process, and in addition, these processes are independent among regimes. Therefore, it is only necessary to be able to obtain the posterior distribution of the driving parameters for each induced regime.

Although this procedure can be done as a second step—i.e. after performing the simulation for obtaining the posterior segmentation—it is also possible to do it in the same simulation based on the MCMC algorithm provided in Section 2.2.1.

Focusing on the second option for estimating regime parameters, we can provide some insights with regard to how it can be done. The MCMC algorithm is summarized in Figure 2. In this, we update the segmentation through the variables k and ψ_k. If we want to infer about regime parameters $x = (x_1, \dots, x_k)$, we need to sample from the posterior distribution $p(k, \psi_k, x|y)$. Therefore, we require samples from the full conditional distributions $p(k, \psi_k|x, y)$ and $p(x|k, \psi_k, y)$. Sampling from the second distribution is easier than from the second one, it only depends on how each regime is modeled. The first distribution, however, presents some details.

In order to sample from $p(x|k, \psi_k, y)$, consider the likelihood function (15). Then—if we assumed $x_j \sim \nu_0$—the corresponding full conditional distribution is given by

$$p(x_j|x_{-j}, k, \psi_k, y) \propto \nu_0(x_j) \prod_{i=1}^{n_j} F(y_{t_{j,i}}; x_j), \qquad j = 1, \dots, k,$$

where x_{-j} denotes the vector x without the jth entry.

For the other distribution, $p(k, \psi_k|x, y)$, we can devise the following. As already explained, sampling the random segmentation is done by choosing from a split or a merge move and, later, performing a shuffle; some of this steps are not affected by the inclusion of the regime parameters. When performing a split, regime parameters are all known, even for the proposed new group. Suppose group s is selected to be split, then parameter x_s should be used for both regimes; therefore Equation (22) can be easily adjusted. Similarly for the case of a shuffle—since the number of groups does not change—all regime parameters are also known. The merge move might not be straightforward to perform tough. If groups s and $s+1$ are chosen to be merged, the main difficulty resides in whether which parameter—x_s or x_{s+1}—should be used as the parameter for the proposed regime or it should be necessary to perform some change of dimension or other computation to obtain such a parameter.

§ 4 One step predictions

When taking about prediction under the context of change-point detection, it can be understood in different ways. We might be interested in forecasting the next change point, which is related to the distribution of the regimes' lengths, or blocks' sizes, or—

given a time position t_{n+1}—we might want to provide a point prediction of $Y_{t_{n+1}}$.

If N_j denotes the random variable modeling the length of the jth regime, any $S_{[n]}$-valued distribution discussed in Section 1 define a joint distribution for the random vector $(N_1, \dots, N_k)$. Moreover, predicting the next change point corresponds to give an estimate of N_{k+1} based on the posterior predictive distribution $\mathbf{P}(N_{k+1}|y, n_1, \dots, n_k)$. On the other hand, predicting the value of a new observation, $Y_{t_{n+1}}$, can be obtained from our proposed model, since

$$\mathbf{P}(Y_{t_{n+1}}|y) = \sum_{\Upsilon_n \in S_{[n]}} \mathbf{P}(Y_{t_{n+1}}|y, \Upsilon_n)\mathbf{P}(\Upsilon_n|y).$$

Then, under quadratic loss, the point prediction of $Y_{t_{n+1}}$ is given by the expectation with respect to this distribution. Although both kinds of prediction could be useful under a specific context, only the second form will be studied next.

Under the context of curve fitting, Wade et al. (2014) provide some developments in prediction which can be used in our context. Following their proposal, the time indices $t_1, \dots, t_n$ play the role of covariates, so the data $(y_{t_1}, \dots, y_{t_n})$ can be thought as a set of pairs (y_i, t_i) for $i = 1, \dots, n$. Therefore, a point prediction of $Y_{t_{n+1}}$ is defined as the point prediction of Y_{n+1} at t_{n+1} by

$$\hat{y}(t_{n+1}) = \mathbf{E}[Y_{n+1}|t_{n+1}, y, t],$$

where $y = (y_1, \dots, y_n)$ and $t = (t_1, \dots, t_n)$. Note that the main role of the covariates t_j is to make clear the location where the prediction is required: (i) outside the observed range of time, i.e. $t_{n+1} < t_1$ or $t_{n+1} > t_n$; or (ii) between two point times already observed, $t_i < t_{n+1} < t_{i+1}$ for some $i \leq n$. The posterior predictive distribution is, then, given by

$$\begin{aligned}\mathbf{P}(Y_{n+1}|t_{n+1}, y, t) &= \sum_{\Upsilon_n \in S_{[n]}} \mathbf{P}(Y_{n+1}|t_{n+1}, y, t, \Upsilon_n)\mathbf{P}(\Upsilon_n|t_{n+1}, y, t) \\ &= \sum_{\Upsilon_n \in S_{[n]}} \sum_{s_{n+1} \in S_n} \mathbf{P}(Y_{n+1}|t_{n+1}, y, t, \Upsilon_n, s_{n+1})\mathbf{P}(s_{n+1}|\Upsilon_n, t_{n+1}, y, t)\mathbf{P}(\Upsilon_n|t_{n+1}, y, t), \quad (28)\end{aligned}$$

where the set S_n is defined according to Υ_n and t_{n+1}; details are provided next.

The auxiliary variable s_{n+1} appears since it necessary to perform a kind of data augmentation in the random segmentation Υ_n in order to consider all the possible groupings of $n+1$ items. Thus, the random segmentation Υ_{n+1} can be defined in terms of the random vector (Υ_n, s_{n+1}); any possible grouping induced by (Υ_n, s_{n+1})—properly relabeled—will

have associated a unique element of S_{n+1}. From this, the computation of the distribution (28) is done as follows.

1. Since s_{n+1} and y are conditionally independent given Υ_n, t_{n+1}, t, the second term in the sum is rewritten as

$$\mathbf{P}(s_{n+1}|\Upsilon_n, t_{n+1}, y, t) = \frac{\mathbf{P}(\Upsilon_n, s_{n+1}|t, t_{n+1})}{\mathbf{P}(\Upsilon_n|t, t_{n+1})} = \frac{\mathbf{P}(\Upsilon_{n+1}|t, t_{n+1})}{\mathbf{P}(\Upsilon_n|t, t_{n+1})}.$$

2. The third term, on the other hand, is rewritten as

$$\mathbf{P}(\Upsilon_n|t_{n+1}, y, t) = \frac{\mathbf{P}(\Upsilon_n|t, t_{n+1})}{\mathbf{P}(\Upsilon_n|t)} \frac{\mathbf{P}(y|t)}{\mathbf{P}(y|t, t_{n+1})} \mathbf{P}(\Upsilon_n|y, t).$$

Therefore, the distribution (28) can be re-expressed as

$$\mathbf{P}(Y_{n+1}|t_{n+1}, y, t) = \sum_{\Upsilon_n \in S_{[n]}} \sum_{s_{n+1} \in S_n} \mathbf{P}(Y_{n+1}|t_{n+1}, y, t, \Upsilon_{n+1}) \frac{\mathbf{P}(\Upsilon_{n+1}|t, t_{n+1})}{\mathbf{P}(\Upsilon_n|t)} \frac{\mathbf{P}(y|t)}{\mathbf{P}(y|t, t_{n+1})} \mathbf{P}(\Upsilon_n|y, t) \quad (29)$$

Assuming the prior distribution for Υ_n has the form (12) and for the specific cases already explained, this predictive distribution simplifies as follows. Consider first the case $t_{n+1} > t_n$. It means that the new observation, Y_{n+1}, either belongs to the last regime or starts a new one, so the point prediction $\hat{y}(t_{n+1})$ is given by

$$\hat{y}(t_{n+1}) = \sum_{k=1}^{n} \sum_{(n_1,\dots,n_k)\in\Delta_{n,k}} \left\{ \frac{n+1}{n_k+1} \frac{\Pi_k^{(n+1)}(n_1, \dots, n_k+1)}{\Pi_k^{(n)}(n_1, \dots, n_k)} \mathbf{E}(Y_{n+1}|t_{n+1}, y, t, (n_1, \dots, n_k+1)) \right.$$
$$\left. + \frac{n+1}{k+1} \frac{\Pi_k^{(n+1)}(n_1, \dots, n_k, 1)}{\Pi_k^{(n)}(n_1, \dots, n_k)} \mathbf{E}(Y_{n+1}|t_{n+1}, y, t, (n_1, \dots, n_k, 1)) \right\} \frac{\mathbf{P}(y|t)}{\mathbf{P}(y|t, t_{n+1})} \mathbf{P}(n_1, \dots, n_k|y, t).$$

Similarly for the case $t_{n+1} < t_1$, the point prediction $\hat{y}(t_{n+1})$ is

$$\hat{y}(t_{n+1}) = \sum_{k=1}^{n} \sum_{(n_1,\dots,n_k)\in\Delta_{n,k}} \left\{ \frac{n+1}{n_1+1} \frac{\Pi_k^{(n+1)}(n_1+1, \dots, n_k)}{\Pi_k^{(n)}(n_1, \dots, n_k)} \mathbf{E}(Y_{n+1}|t_{n+1}, y, t, (n_1+1, \dots, n_k)) \right.$$
$$\left. + \frac{n+1}{k+1} \frac{\Pi_k^{(n+1)}(1, n_1, \dots, n_k)}{\Pi_k^{(n)}(n_1, \dots, n_k)} \mathbf{E}(Y_{n+1}|t_{n+1}, y, t, (1, n_1, \dots, n_k)) \right\} \frac{\mathbf{P}(y|t)}{\mathbf{P}(y|t, t_{n+1})} \mathbf{P}(n_1, \dots, n_k|y, t),$$

where $\Pi_k^{(n)}$ is the underlying EPPF of the prior for Υ_n.

The last case—where the new time point t_{n+1} is located between two already observed points t_i and t_{i+1}, for some $i \leq n$—can be obtained similarly, but the resulting set S_n varies as follows. Given a segmentation of $[n]$, there are four possible cases. The time points t_i and t_{i+1} belong to the same regime; then, the point t_{n+1} must belong to such regime. Otherwise, these points, t_i and t_{i+1}, belong to different (adjacent) regimes; then, the point t_{n+1} can belong to the regime of t_i or to the regime of t_{i+1} or it can form a new regime, between the to existing ones. Putting together all these cases, the point prediction $\hat{y}(t_{n+1})$ is given by

$$\hat{y}(t_{n+1}) = \sum_{k=1}^{n} \sum_{\{\tau_1,\dots,\tau_k\}\in S_{[n]}^k} \hat{y}(t_{n+1};\{\tau_1,\dots,\tau_k\}) \frac{\mathbf{P}(y|t)}{\mathbf{P}(y|t,t_{n+1})} \mathbf{P}(n_1,\dots,n_k|y,t).$$

where, for $j = 1,\dots,k$, $n_j = \#\tau_j$ as usual, and

$$\hat{y}(t_{n+1};\{\tau_1,\dots,\tau_k\}) = \frac{n+1}{n_j+1} \frac{\Pi_k^{(n+1)}(n_1,\dots,n_j+1,\dots,n_k)}{\Pi_k^{(n)}(n_1,\dots,n_k)} \mathbf{E}(Y_{n+1}|t_{n+1},y,t,(n_1,\dots,n_j+1,\dots,n_k)),$$

if $i, i+1 \in \tau_j$, and—for the case $i \in \tau_j$ and $i+1 \in \tau_{j+1}$—

$$\begin{aligned}\hat{y}(t_{n+1};\{\tau_1,\dots,\tau_k\}) = {} & \frac{n+1}{n_j+1} \frac{\Pi_k^{(n+1)}(n_1,\dots,n_j+1,\dots,n_k)}{\Pi_k^{(n)}(n_1,\dots,n_k)} \mathbf{E}(Y_{n+1}|t_{n+1},y,t,(n_1,\dots,n_j+1,\dots,n_k)) + \\ & \frac{n+1}{n_{j+1}+1} \frac{\Pi_k^{(n+1)}(n_1,\dots,n_{j+1}+1,\dots,n_k)}{\Pi_k^{(n)}(n_1,\dots,n_k)} \mathbf{E}(Y_{n+1}|t_{n+1},y,t,(n_1,\dots,n_{j+1}+1,\dots,n_k)) + \\ & \frac{n+1}{k+1} \frac{\Pi_{k+1}^{(n+1)}(n_1,\dots,n_j,1,n_{j+1},\dots,n_k)}{\Pi_k^{(n)}(n_1,\dots,n_k)} \mathbf{E}(Y_{n+1}|t_{n+1},y,t,(n_1,\dots,n_j,1,n_{j+1},\dots,n_k)).\end{aligned}$$

§ II.A Simulation results for the retrospective model

In this appendix, the simulation results of the models of Section 2.1. Tables 4–6 show the MCMC results of our proposed model presented in Section 2.2.1 for the synthetic datasets explained in Section 2.3. Table 7 displays a comparison between the posterior mode and the least-squares clustering estimate for the same datasets. Finally, Table 8 contains the respective results of Loschi and Cruz's method discussed in Section 2.3.1.

ϕ	$\mathbb{E}[C_n]$	σ	θ	$\tilde{Y}_n$	*prob.*	$\tilde{C}_n$	*prob.*
0.0	2	0.1	0.1897	$\langle 51, 86\rangle$	0.7880	2	0.7966
		0.3	−0.1634	$\langle 51, 86\rangle$	0.7379	2	0.7454
		0.6	−0.5709	$\langle 51, 86\rangle$	0.6757	2	0.6757
		0.9	−0.8979	$\langle 51, 86\rangle$	0.6130	2	0.6142
	49	0.1	21.4127	$\langle 2, 50, 51, 86, 87, 149, 150\rangle$	0.0008	15	0.1195
		0.3	13.0593	$\langle 51, 86, 87\rangle$	0.0035	9	0.1267
		0.6	3.0021	$\langle 51, 86\rangle$	0.1969	3	0.2842
		0.9	−0.7974	$\langle 51, 86\rangle$	0.6069	2	0.6069
	99	0.1	113.4390	$\langle \dots \rangle_{40}$	0.0003	43	0.0700
		0.3	80.8346	$\langle \dots \rangle_{37}$	0.0004	35	0.0891
		0.6	34.2148	$\langle \dots \rangle_{10}$	0.0008	18	0.0866
		0.9	0.4399	$\langle 51, 86\rangle$	0.4056	2	0.4127
		0.1	*r.v.*	$\langle 51, 86\rangle$	0.7234	2	0.7235
		0.3	*r.v.*	$\langle 51, 86\rangle$	0.6558	2	0.6603
		0.6	*r.v.*	$\langle 51, 86\rangle$	0.5905	2	0.5927
		0.9	*r.v.*	$\langle 51, 86\rangle$	0.5354	2	0.5359
		r.v.	*r.v.*	$\langle 51, 86\rangle$	0.7030	2	0.7030
r.v.	2	0.1	0.1897	$\langle 51, 86\rangle$	0.8134	2	0.8134
		0.3	−0.1634	$\langle 51, 86\rangle$	0.7908	2	0.7908
		0.6	−0.5709	$\langle 51, 86\rangle$	0.7579	2	0.7579
		0.9	−0.8979	$\langle 51, 86\rangle$	0.7631	2	0.7634
	49	0.1	21.4127	$\langle \dots \rangle_{11}$	0.0004	13	0.1416
		0.3	13.0593	$\langle 51, 86\rangle$	0.0071	8	0.1436
		0.6	3.0021	$\langle 51, 86\rangle$	0.3486	2	0.3486
		0.9	−0.7974	$\langle 51, 86\rangle$	0.7531	2	0.7534
	99	0.1	113.4390	$\langle \dots \rangle_{32}$	0.0004	36	0.0946
		0.3	80.8346	$\langle \dots \rangle_{24}$	0.0005	28	0.0893
		0.6	34.2148	$\langle \dots \rangle_{18}$	0.0008	14	0.1205
		0.9	0.4399	$\langle 51, 86\rangle$	0.6336	2	0.6339
		0.1	*r.v.*	$\langle 51, 86\rangle$	0.7414	2	0.7414
		0.3	*r.v.*	$\langle 51, 86\rangle$	0.7046	2	0.7046
		0.6	*r.v.*	$\langle 51, 86\rangle$	0.7139	2	0.7139
		0.9	*r.v.*	$\langle 51, 86\rangle$	0.7632	2	0.7632
		r.v.	*r.v.*	$\langle 51, 86\rangle$	0.7162	2	0.7162

TABLE 4: Posterior results for the simulated dataset with changes in mean (Figure 4a). The first four columns show the prior specifications; the label *r.v.* means that a prior distribution was assigned. The last four columns show the modal change points, $\tilde{Y}_n$, and the modal number of change points, $\tilde{C}_n$, together with their corresponding probabilities. The modal change points indicated as $\langle \dots \rangle_k$ denote that there are k of them.

ϕ	$\mathbb{E}[C_n]$	σ	θ	$\tilde{\Upsilon}_n$	*prob.*	$\tilde{C}_n$	*prob.*
0.0	2	0.1	0.1897	$\langle 50, 66, 75, 113, 121\rangle$	0.0094	6	0.2164
		0.3	−0.1634	$\langle 53, 66, 76, 113, 121\rangle$	0.0022	9	0.0927
		0.6	−0.5709	$\langle 54, 55, 111, 112, 113, 121\rangle$	0.0014	19	0.0602
		0.9	−0.8979	$\langle \dots \rangle_{10}$	0.0018	23	0.0376
	49	0.1	21.4127	$\langle \dots \rangle_{21}$	0.0005	33	0.1058
		0.3	13.0593	$\langle \dots \rangle_{29}$	0.0006	31	0.1034
		0.6	3.0021	$\langle \dots \rangle_{26}$	0.0007	26	0.0887
		0.9	−0.7974	$\langle 53, 54, 112, 113, 121\rangle$	0.0034	10	0.0418
	99	0.1	113.4390	$\langle \dots \rangle_{67}$	0.0004	65	0.0818
		0.3	80.8346	$\langle \dots \rangle_{58}$	0.0004	59	0.0827
		0.6	34.2148	$\langle \dots \rangle_{55}$	0.0005	51	0.0767
		0.9	0.4399	$\langle \dots \rangle_{15}$	0.0019	37	0.0583
		0.1	*r.v.*	$\langle 49, 66, 75, 97, 99, 108, 121\rangle$	0.0013	11	0.1011
		0.3	*r.v.*	$\langle \dots \rangle_{13}$	0.0008	15	0.1054
		0.6	*r.v.*	$\langle \dots \rangle_{24}$	0.0008	24	0.0814
		0.9	*r.v.*	$\langle \dots \rangle_{36}$	0.0014	33	0.0486
		r.v.	*r.v.*	$\langle \dots \rangle_{11}$	0.0010	17	0.0746
r.v.	2	0.1	0.1897	$\langle 53, 121\rangle$	0.0559	2	0.4422
		0.3	−0.1634	$\langle 53, 121\rangle$	0.0452	2	0.4258
		0.6	−0.5709	$\langle 53, 121\rangle$	0.0434	2	0.4341
		0.9	−0.8979	$\langle 53, 121\rangle$	0.0486	2	0.5111
	49	0.1	21.4127	$\langle \dots \rangle_{23}$	0.0006	27	0.1013
		0.3	13.0593	$\langle \dots \rangle_{21}$	0.0006	21	0.0891
		0.6	3.0021	$\langle 53, 123\rangle$	0.0086	5	0.1640
		0.9	−0.7974	$\langle 53, 121\rangle$	0.0413	2	0.4763
	99	0.1	113.4390	$\langle \dots \rangle_{67}$	0.0005	66	0.0784
		0.3	80.8346	$\langle \dots \rangle_{71}$	0.0005	60	0.0721
		0.6	34.2148	$\langle \dots \rangle_{39}$	0.0008	49	0.0578
		0.9	0.4399	$\langle 53, 121\rangle$	0.0290	2	0.3403
		0.1	*r.v.*	$\langle 53, 121\rangle$	0.0437	2	0.3218
		0.3	*r.v.*	$\langle 53, 121\rangle$	0.0340	2	0.2949
		0.6	*r.v.*	$\langle 53, 121\rangle$	0.0263	2	0.3730
		0.9	*r.v.*	$\langle 53, 121\rangle$	0.0379	2	0.4616
		r.v.	*r.v.*	$\langle 53, 121\rangle$	0.0405	2	0.3407

TABLE 5: Posterior results for the simulated dataset with changes in variance (Figure 4b). The first four columns show the prior specifications; the label *r.v.* means that a prior distribution was assigned. The last four columns show the modal change points, $\tilde{\Upsilon}_n$, and the modal number of change points, $\tilde{C}_n$, together with their corresponding probabilities. The modal change points indicated as $\langle \dots \rangle_k$ denote that there are k of them.

ϕ	$\mathbb{E}[C_n]$	σ	θ	$\tilde{\Upsilon}_n$	*prob.*	$\tilde{C}_n$	*prob.*
0.0	2	0.1	0.1897	$\langle 2, 51, 103\rangle$	0.0168	5	0.3322
		0.3	−0.1634	$\langle 2, 51, 82, 91, 101\rangle$	0.0107	5	0.1684
		0.6	−0.5709	$\langle 2, 51, 103\rangle$	0.0291	3	0.1588
		0.9	−0.8979	$\langle\rangle$	0.9071	0	0.9071
	49	0.1	21.4127	$\langle\dots\rangle_{32}$	0.0004	29	0.0929
		0.3	13.0593	$\langle\dots\rangle_{19}$	0.0006	24	0.0817
		0.6	3.0021	$\langle 2, 3, 50, 51, 103, 150\rangle$	0.0021	14	0.1111
		0.9	−0.7974	$\langle 2\rangle$	0.5550	1	0.5682
	99	0.1	113.4390	$\langle\dots\rangle_{71}$	0.0004	64	0.0756
		0.3	80.8346	$\langle\dots\rangle_{60}$	0.0004	62	0.0731
		0.6	34.2148	$\langle\dots\rangle_{42}$	0.0005	44	0.0798
		0.9	0.4399	$\langle 2\rangle$	0.0943	7	0.1109
		0.1	*r.v.*	$\langle 2, 51, 103\rangle$	0.0075	6	0.1779
		0.3	*r.v.*	$\langle 2, 51, 82, 91, 101\rangle$	0.0088	6	0.1837
		0.6	*r.v.*	$\langle 2\rangle$	0.0107	8	0.1267
		0.9	*r.v.*	$\langle 2\rangle$	0.1454	1	0.1477
		r.v.	*r.v.*	$\langle 2, 51, 103\rangle$	0.0129	6	0.1807
r.v.	2	0.1	0.1897	$\langle 51, 108\rangle$	0.0846	2	0.4178
		0.3	−0.1634	$\langle 51, 108\rangle$	0.0450	2	0.3212
		0.6	−0.5709	$\langle\rangle$	0.3802	0	0.3802
		0.9	−0.8979	$\langle\rangle$	0.9946	0	0.9946
	49	0.1	21.4127	$\langle\dots\rangle_{20}$	0.0008	24	0.0909
		0.3	13.0593	$\langle\dots\rangle_{10}$	0.0006	18	0.0998
		0.6	3.0021	$\langle 51, 108, 118, 120\rangle$	0.0056	7	0.1532
		0.9	−0.7974	$\langle\rangle$	0.5752	0	0.5752
	99	0.1	113.4390	$\langle\dots\rangle_{67}$	0.0004	64	0.0811
		0.3	80.8346	$\langle\dots\rangle_{55}$	0.0004	62	0.0649
		0.6	34.2148	$\langle\dots\rangle_{23}$	0.0004	36	0.0612
		0.9	0.4399	$\langle 2\rangle$	0.2874	1	0.3665
		0.1	*r.v.*	$\langle 51, 108\rangle$	0.0464	2	0.2493
		0.3	*r.v.*	$\langle 51, 108\rangle$	0.0357	2	0.2212
		0.6	*r.v.*	$\langle 51, 108\rangle$	0.0345	3	0.2274
		0.9	*r.v.*	$\langle\rangle$	0.5786	0	0.5786
		r.v.	*r.v.*	$\langle 51, 103\rangle$	0.0254	2	0.2599

TABLE 6: Posterior results for the simulated dataset with changes in mean, variance and correlation (Figure 4c). The first four columns show the prior specifications; the label *r.v.* means that a prior distribution was assigned. The last four columns show the modal change points, $\tilde{\Upsilon}_n$, and the modal number of change points, $\tilde{C}_n$, together with their corresponding probabilities. The modal change points indicated as $\langle\dots\rangle_k$ denote that there are k of them.

Dataset	$\mathbb{E}[C_n]$	σ	θ	ϕ	*mode*	*co-cluster*
A	49	0.3	13.0593	0.0	$\langle 51, 86, 87\rangle$	$\langle 4, 30, 50, 51, 86, 87\rangle$
				r.v.	$\langle 51, 86\rangle$	$\langle 5, 11, 15, 51, 86\rangle$
B	99	0.9	0.4399	0.0	$\langle 50, 66, \dots, 113, 121\rangle_{15}$	$\langle 50, 66, \dots, 113, 121\rangle_{30}$
				r.v.	$\langle 53, 121\rangle$	$\langle 52, 122\rangle$
C	99	0.9	0.4399	0.0	$\langle 2\rangle$	$\langle 2, 51, 103\rangle$
				r.v.	$\langle 2\rangle$	$\langle 2, 3\rangle$

TABLE 7: Posterior results for different datasets under distinct scenarios. Datasets used have changes in: A) mean, B) variance, and C) mean, variance and correlation (Figure 4). Columns 2 to 5 display the prior specifications; label *r.v.* means that a random variable was assigned. Column 6 displays the posterior mode and Column 7 displays the co-cluster; the subindex in the partition indicates its length.

$\mathbb{E}[C_n]$	α	β	$\tilde{\Upsilon}_n$	*prob.*	$\tilde{C}_n$	*prob.*
2	1.00	73.50	$\langle 51, 86\rangle$	0.9900	2	0.9901
	0.01	1.00	$\langle 51, 86\rangle$	0.9924	2	0.9925
49	1.00	2.04	$\langle 51, 86\rangle$	0.9848	2	0.9849
	0.49	1.00	$\langle 51, 86\rangle$	0.9891	2	0.9892
99	1.00	0.51	$\langle 51, 86\rangle$	0.9847	2	0.9848
	1.98	1.00	$\langle 51, 86\rangle$	0.9773	2	0.9774
75	1	1	$\langle 51, 86\rangle$	0.9847	2	0.9848
3	1	50	$\langle 51, 86\rangle$	0.9900	2	0.9891
14	5	50	$\langle 51, 86\rangle$	0.9660	2	0.9662
146	50	1	$\langle 51, 86\rangle$	0.6242	2	0.6244
136	50	5	$\langle 51, 86\rangle$	0.6347	2	0.6349

(A) Results for the dataset with changes in mean (Figure 4a).

$\mathbb{E}[C_n]$	α	β	$\tilde{\Upsilon}_n$	*prob.*	$\tilde{C}_n$	*prob.*
2	1.00	73.50	$\langle 4, 53, 121\rangle$	0.0456	3	0.3674
	0.01	1.00	$\langle 4, 53, 121\rangle$	0.0463	4	0.3379
49	1.00	2.04	$\langle 4, 53, 121\rangle$	0.0280	4	0.3494
	0.49	1.00	$\langle 4, 53, 121\rangle$	0.0365	4	0.3519
99	1.00	0.51	$\langle 4, 53, 121\rangle$	0.0275	4	0.3471
	1.98	1.00	$\langle 4, 53, 121, 123, 146\rangle$	0.0335	4	0.3332
75	1	1	$\langle 4, 53, 121, 123, 146\rangle$	0.0275	4	0.3474
3	1	50	$\langle 4, 53, 121\rangle$	0.0362	4	0.3595
14	5	50	$\langle 4, 50, 121, 123, 146\rangle$	0.0257	5	0.3485
146	50	1	$\langle 4, 53, 121, 123, 146, 149, 150\rangle$	0.0143	7	0.2914
136	50	5	$\langle 4, 53, 121, 123, 146, 149, 150\rangle$	0.0152	7	0.2970

(B) Results for the dataset with changes in variance (Figure 4b).

$\mathbb{E}[C_n]$	α	β	$\tilde{\Upsilon}_n$	*prob.*	$\tilde{C}_n$	*prob.*
2	1.00	73.50	$\langle\rangle$	1.0000	0	1.0000
	0.01	1.00	$\langle\rangle$	1.0000	0	1.0000
49	1.00	2.04	$\langle\rangle$	1.0000	0	1.0000
	0.49	1.00	$\langle\rangle$	1.0000	0	1.0000
99	1.00	0.51	$\langle\rangle$	1.0000	0	1.0000
	1.98	1.00	$\langle\rangle$	0.9993	0	0.9993
75	1	1	$\langle\rangle$	1.0000	0	1.0000
3	1	50	$\langle\rangle$	1.0000	0	1.0000
14	5	50	$\langle\rangle$	0.9971	0	0.9971
146	50	1	$\langle\rangle$	0.6214	0	0.6214
136	50	5	$\langle\rangle$	0.6050	0	0.6050

(C) Results for the dataset with changes in mean, variance and correlation (Figure 4c).

TABLE 8: Posterior results for the simulated datasets using Loschi and Cruz's method. For each table, the first three columns show the prior specifications and the last four columns show the modal change points, $\tilde{\Upsilon}_n$, and the modal number of change points, $\tilde{C}_n$, together with their corresponding probabilities.

CHAPTER III

DECREASING-WEIGHT MIXTURE MODELS

In this chapter, we explore the relationship between clustering and mixture models, focusing on a particular class of mixture densities generated by discrete random probability measures (RPMs). Among the great diversity of clustering methodologies, mixture models emerge as a natural extension of exchangeability to tackle this problem. As depicted in the Clustering Scenario 1.1, random partitions can be used for clustering, therefore, we extend their usage to the case when observations are not generated from discrete distributions.

In Section 1, an introduction to mixture models is given. Besides of motivating their usage, we explain some of the specific problems where these models are applied as well as some of their main drawbacks. It is worth saying that, in this chapter, we will concentrate on mixture densities with an infinite number of components. Thus, in this section, we present some examples, starting with mixture densities whose mixing distribution corresponds to the Dirichlet process. Afterwards, we introduce a recent construction of discrete RPMs—the geometric process—which diminishes some of the weaknesses of other models, when it is used under this context.

Due to the novelty of the geometric-weight approach—in Section 2—we present some extensions to this process, preserving the same restriction in the mixing weights. Specifically, these extensions are based on different power series distributions as well as Lagrangian distributions. We also implement a Gibbs sampler in order to test their performance in density estimation.

Finally, in Section 3, we study how it is possible to infer about the clustering structure based on mixture models. In this section, we also explain our current work in this line: the computation of the exchangeable partition probability function (EPPF) for discrete

RPMs with decreasing weights.

§1 INFINITE-COMPONENT MIXTURE MODELS

Mixture models are of interest in many contexts due to their flexibility to model heterogeneity. In these, observations are assumed independent and identically distributed according to a mixture density of the form

$$f(y) = \sum_{j=1}^{M} w_j \kappa_j(y; x_j), \tag{1}$$

where $(w_1, \dots, w_M)$ are non-negative weights summing up to one, and κ_j is a kernel density function, for each j, with parameter x_j. Early models assume a finite number of components, but it can also be infinite. Moreover, the kernel density is commonly simplified as $\kappa_j(\cdot; x_j) = \kappa(\cdot; x_j)$, i.e. the same kernel is assumed for all components and only its parameter x_j varies.

The study of this topic goes back to Pearson (1894), and there are also several extensive expositions about it in Titterington et al. (1985), McLachlan & Peel (2000), Marin et al. (2005) and Frühwirth-Schnatter (2006) to mention just a few. One of the interpretations of mixture models—which makes them attractive in several and different fields—was noted by Feller (1943). This assumes a population formed by M categories, also called groups or subpopulations, each one with proportion w_j, $j = 1, \dots, M$. The term category refers to the fact that the feature of interest—represented by parameter x_j—is homogeneous within groups but heterogeneous across them. Therefore, in practice, one of the main interests has been trying to describe the latent subpopulations in terms of their relative proportion, w_j, and their characterizing property x_j.

However, it has been noticed that such an interpretation hardly can be achieved due to different reasons. Mixture models of the form (1) are known to be *model-based*, which means that inferences might change drastically according the kernel κ we choose. As a consequence, there are many different approaches dealing with this issue; a general assumption has been to restrict the form of the kernel density to be unimodal. Another characteristic of mixture models—which, in general, represents a problem for inferences—is known as *non-identifiability*, and it appears when there are different combinations of components—each one formed by a weight and its corresponding kernel parameter, (w_j, x_j)—resulting in the same density for a single observation. Furthermore, the lack of identifiability also causes convergence problems in Monte Carlo Markov chain

(MCMC) methods—known as the *label-switching* problem—where the labels associating observations and components change between iterations.

Under a Bayesian nonparametric framework, when $M = \infty$, the first approach to define mixture models was given by Lo (1984) and it makes use of the Dirichlet process in order to obtain expressions for the mixing weights and locations. But it was not until Escobar & West (1995) proposed a computational implementation that the interest in infinite-component mixture models increased. Among the current uses of these mixture models, density estimation and clustering are the most predominant. Unfortunately, we can find the same undesirable features of the finite-component case: (i) they are model-based, (ii) we have seldom a non-identifiable model, and, for this reason (iii) label-switching appears in MCMC methods. Hence, during this chapter, we will focus on this class of mixture models in order to study these issues and look for a way to deal with them.

1.1 Dirichlet process mixture models and related constructions

The model proposed by Lo (1984) can be written by means of a discrete RPM $\tilde{p}$ as mixing distribution. Recalling, any discrete RPM $\tilde{p}$ can be expressed as

$$\tilde{p}(\cdot) = \sum_{j=1}^{\infty} w_j \delta_{x_j}(\cdot), \tag{2}$$

where w_j is a $(0,1)$-valued random variable for all $j \geq 1$, and the sequence $(w_j)_{j\geq 1}$ is such that $\sum_{j\geq 1} w_j = 1$ a.s.; $(x_j)_{j\geq 1}$ is a sequence of i.i.d. $\mathbb{X}$-valued random variables—independent of $(w_j)_{j\geq 1}$—from a non-atomic distribution ν_0. We already studied these measures in Section 1.1. Hence, the resulting mixture density is such that

$$f(y) = \int_{\mathbb{X}} \kappa(y;x)\tilde{p}(\mathrm{d}x) = \sum_{j=1}^{\infty} w_j \kappa(y;x_j), \tag{3}$$

for a kernel density function κ with finite-dimensional parameter x_j. When the distribution of $\tilde{p}$ coincides with the Dirichlet process, the model proposed by Lo—better known as the *Dirichlet process mixture model*—is obtained.

An alternative method to construct mixture models consists on using the so-called *stick-breaking representation* to define the distribution of the mixing weights, $(w_j)_{j\geq 1}$. The locations $(x_j)_{j\geq 1}$ are defined as before. From the *stick-breaking* representation (Halmos

1944), mixing weights are defined as

$$w_1 = v_1, \qquad w_j = v_j \prod_{l<j}(1 - v_l) \quad j \geq 2, \tag{4}$$

for $(v_j)_{j\geq 1}$ a sequence of independent $(0, 1)$-valued random variables. The only restriction of this construction is that $\sum_{i\geq 1} \log \mathbf{E}(1 - v_i) = -\infty$; see, for example, Ghosal & van der Vaart (2015) for more details. Several well known mixture models can be defined following this representation. For example, Sethuraman (1994) finds that the Dirichlet process mixture model corresponds to the case $v_j \overset{\text{i.i.d.}}{\sim} \text{Be}(1, c)$, for all $j \geq 1$. It is also possible to recover the two-parameter Poisson-Dirichlet process by letting $v_j \overset{\text{ind.}}{\sim} \text{Be}(1 - \sigma, \theta + j\sigma)$, for some $0 < \sigma < 1$ and $\theta > -\sigma$ (cf. Example 1.5). Further examples and a more extensive discussion about this representation under the mixture-model framework can be found in Ishwaran & James (2001). It is worth mentioning that—in some cases—it is not possible to use the RPM-based and the *stick-breaking*-based approaches interchangeably. In other words, not all mixture models (3) built through (2) can be expressed using directly the representation (4) and vice versa. One of these cases is given by the normalized generalized Gamma process (Example 1.4), where it is necessary to break the assumption of independence in (4) and make use of latent variables; see Lau & Cripps (2015) and Favaro et al. (2015) for details.

1.1.1 Density estimation

As mentioned before, one of the main uses of mixture models is density estimation. Under a Bayesian nonparametric framework, we can find a great number of methodologies. These, in general, are based on one of the two aforementioned constructions and can be classified according to whether—in order to implement an MCMC method— (i) the underlying RPM is marginalized out, or (ii) it is used to sample from. Following the first approach, Escobar & West (1995) were the firsts in implementing a sampling scheme for the Dirichlet case through the very well known *Pólya urn scheme*. We also found some alternatives for this idea in Neal (2000) and more general schemes presented by Ishwaran & James (2001), Lijoi et al. (2007*b*) and Favaro & Teh (2013). Regarding the second approach, also Ishwaran & James (2001) propose a method to truncate the random measure based on a certain deterministic distance. Later on, Walker (2007) and Kalli et al. (2011) propose also a truncation of the random measure but it is done randomly; their proposal is known as the *slice sampler*. More recently, Favaro & Teh (2013) propose an MCMC al-

gorithm for a particular family of random distributions $\tilde{p}$ based on this sampler scheme.

For the sake of completeness, we summarize two of these methods: (i) the Pólya urn scheme, and (ii) the slice sampler.

Pólya urn scheme. Writing the mixture model (3) in a hierarchical form, we have

$$\begin{aligned} y_i|x_i &\overset{\text{ind.}}{\sim} \kappa(y_i; x_i), \qquad i = 1, \dots, n \\ x_i|\tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim Q \end{aligned} \tag{5}$$

where Q is the distribution of $\tilde{p}$. Under the Pólya urn scheme, the random distribution $\tilde{p}$ is integrated out, leading the following model

$$\begin{aligned} y_i|x_i &\overset{\text{ind.}}{\sim} \kappa(y_i; x_i), \qquad i = 1, \dots, n \\ x_1, \dots, x_n &\sim \pi(x_1, \dots, x_n), \end{aligned}$$

where π is a joint distribution. The key of this algorithm resides in the fact that $\tilde{p}$ needs to have an explicit prediction rule (cf. Equation 1.13). Thus, due to the exchangeability of the sequence $x = \{x_1, \dots, x_n\}$, the full conditional distribution for each x_j is given by

$$p(x_i|x_{-i}, y) = q^*_{i,0} p(x_i|y_i) + \sum_{j=1}^{k_{i,n-1}} q^*_{i,j} \delta_{x^*_{i,j}}(x_i), \qquad i = 1, \dots, n,$$

with $x_{-i} := \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}$, and where $x^*_{-i} = \{x^*_{i,1}, \dots, x^*_{i,k_{i,n-1}}\}$ are the $k_{i,n-1}$ distinct values of the vector x_{-i}. The weights $(q_{i,j})_{j=0}^{k_{i,n-1}}$ are determined by the random distribution of $\tilde{p}$; for example, if Q coincides with the two-parameter Poisson-Dirichlet process, we have

$$\begin{aligned} q_{i,0} &\propto (\theta + k_{i,n-1}\sigma) \int_{\mathbb{X}} p(y_i|x) g_0(x) \mathrm{d}x \\ q^*_{i,j} &\propto (n^*_{i,j} - \sigma) p(y_i|x^*_{i,j}), \qquad j \geq 1 \end{aligned}$$

where g_0 is the density function of the base measure ν_0, and $(n^*_{i,1}, \dots, n^*_{i,k_{i,n-1}})$ are the frequencies of x^*_{-i}. An additional step was introduced by MacEachern (1994) in order to speed up the convergence. This consists on introducing a membership vector $(d_1, \dots, d_n)$ to identify which component the ith observation is associated to—that is, $d_i = j$ if and

only if $x_i = x_j^*$, $i = 1, \dots, n$, where $x^* = \{x_1^*, \dots, x_{k^*}^*\}$ are the k^* different values of x—and then, updating each parameter x_j^* as follows

$$p(x_j^* | \cdots) \propto g_0(x_j^*) \prod_{d_i = j} \kappa(y_i; x_j^*), \qquad j = 1, \dots, k^*. \tag{6}$$

Under this sampling scheme, the estimated density is given by

$$\tilde{f}(y) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{k^{*(t)}} w_j^{(t)} \kappa(y; x_j^{*(t)}), \tag{7}$$

where the mixing weights are given by $w_j^{(t)} = n_j^{*(t)}/n$ and T is the sample size. For the case of the Dirichlet process, Escobar & West (1995) have provided a proof about its convergence.

Slice sampler. Under the context of mixture modeling, Walker (2007) introduced a MCMC sampling scheme which, in turn, was generalized by Kalli et al. (2011). This second approach starts with the mixture density (3) and introduces latent variables u and d as well as a sequence of nonnegative numbers $(\xi_j)_{j \geq 1}$ such that (3) is augmented as

$$f(y, u, d) = \mathbf{1}(u < w_d) w_d \xi_d^{-1} \kappa(y; x_d^*).$$

Therefore—given a sample $y_1, \dots, y_n$—the MCMC draws samples from the parameters $(x_j^*)_{j=1}^m$, $(w_j)_{j=1}^m$, $(u_i)_{i=1}^n$ and $(d_i)_{i=1}^n$—where $m = \max\{d_i : i = 1, \dots, n\}$—as follows.

1. For the kernel parameters, the full conditional distribution is the same as Equation (6) in the Pólya urn scheme.

2. Updating the mixing weights (w_j) varies according to their prior distribution. If the *stick-breaking* representation is used, updating (w_j) is equivalent to updating the random variables (v_j). Assuming $v_j \overset{\text{ind.}}{\sim} \text{Be}(a_j, b_j)$, the full conditional distribution for each v_j is also a Beta distribution with parameters (a_j', b_j') such that

$$a_j' = a_j + \sum_{i=1}^{n} \mathbf{1}(d_i = j) \qquad \text{and} \qquad b_j' = b_j + \sum_{i=1}^{n} \mathbf{1}(d_i > j).$$

3. For u_i, $i = 1, \dots, n$, we have

$$p(u_i| \cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i}).$$

4. Indicator variables d_i, $i = 1, \dots, n$, are updated as follows

$$\mathbf{P}(d_i = k| \cdots) \propto \mathbf{1}(k : \xi_k > u_i) w_k \xi_k^{-1} \kappa(y_i; x_k^*).$$

In this case, the estimated density is

$$\hat{f}(y) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \kappa(y; x_j^{(t)}), \tag{8}$$

1.2 Geometric mixture models

The Dirichlet process mixture model constitutes the most representative infinite-component mixture model, and many other models have been conceived based on it. As a consequence, many of the resulting models are, in general, more challenging to be applied; for example, the required sampling schemes are often more elaborated. This fact is mainly related to their weight structure.

With respect to this, Fuentes-García et al. (2010*a*) propose a novel mixture model with a particular structure in its mixing weights. In particular, they explain that a desirable property of a mixture model is that subpopulations are represented by the mixture components—similar to Feller's interpretation—but, additionally, the first component should represent the largest subpopulation, the second component should represent the second largest subpopulation and so on (cf. Mena 2013, Mena & Walker 2015). Hence, they propose a mixture model with a simple but remarkable structure in its weights: they are decreasingly ordered, i.e. $w_j > w_{j+1}$ a.s. for all $j \geq 1$. Note that stick-breaking based weights do not satisfy this constrain.

A derivation of their proposed model is the following. Taking an infinite mixture density (3), first Walker (2007) introduces a latent variable u such that the joint density has the following form

$$f(y, u) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j) \kappa(y; x_j).$$

Thus, given u, the mixture density has a finite number of components since it can be

written as

$$f(y|u) = \frac{1}{|J_u|} \sum_{j \in J_u} \kappa(y; x_j),$$

where $J_u = \{j : u < w_j\}$ is a finite set. Notice that—due to the structure of the weights generated by the stick-breaking scheme or other similar methodologies—the random set J_u usually contains a sequence of non-consecutive integers. Thus, Fuentes-García et al. (2010*a*) consider the set $J_u = \{1, \dots, r\}$, which is a set formed by the first r integers, for a random $r \geq 1$ a.s. The conditional mixture density is, then, written as

$$f(y|r) = \frac{1}{r} \sum_{i=1}^{r} \kappa(y; x_i).$$

Marginalizing over r, an infinite mixture model is recovered, i.e.

$$f(y) = \sum_{j=1}^{\infty} \frac{1}{j} \sum_{i=1}^{j} \kappa(y; x_i) \pi(j; \lambda),$$

with $\pi(\cdot; \lambda)$ the prior distribution of r, however, unlike stick-breaking based approaches, this new mixture density has decreasing weights since they are given by

$$w_j = \sum_{r=j}^{\infty} \frac{\pi(r; \lambda)}{r}, \qquad j = 1, 2, \dots, \tag{9}$$

where, clearly, $w_j > w_{j+1}$ for all $j = 1, 2, \dots$. The specification is complete after assigning a prior distribution to parameter λ.

With this structure, Fuentes-García et al. (2010*a*) take $\pi(r; \lambda)$ the negative binomial with parameters $(2, \lambda)$, for $0 < \lambda < 1$, so the weights are simplified to

$$w_j = \lambda(1 - \lambda)^{j-1}, \qquad j = 1, 2, \dots. \tag{10}$$

The underlying process generating such structure is called the *geometric process* and the resulting RPM is known as the *geometric RPM*.

Although the weights generated by the geometric process look quite different to the ones obtained with the Dirichlet process, there is an interesting relation between both of them. Consider the stick-breaking representation (4) for the Dirichlet process with

parameter c, and denote by w'_j its jth weight. Then, taking expectations, we have

$$\mathbf{E}(w'_j) = \mathbf{E}\left(v_j \prod_{l<j}(1 - v_l) \right) = \frac{1}{1+c} \left(\frac{c}{1+c} \right)^{j-1},$$

where $v_j \overset{\text{i.i.d.}}{\sim} \text{Be}(1, c)$ for $c > 0$. If we set $\lambda = 1/(1+c)$, the weights (10) are recovered. The randomization of the parameter λ—for the geometric process—plays a similar role to the randomization of the total mass parameter, c, for the Dirichlet process. In the latter approach—this extra step is performed in order to remove the bias imposed over the number of groups (cf. Escobar & West 1995), whereas—for the geometric process mixture model—making random the parameter λ allows it to have *full support*, i.e. the resulting random density f is dense in the same space $\mathbb{X}$ where the base measure ν_0 is defined. Bissiri & Ongaro (2014) present some theoretical results about this topic, and also explain—in their Section 3.3—why geometric-process mixture models have full support.

As a side note, the mixing weights obtained by the two-parameter Poisson-Dirichlet process are also ordered in mean, even though its stick-breaking representation generates unordered weights a.s. In this case, we have that

$$\mathbf{E}(w_j) = \frac{(1-\sigma)\prod_{i=1}^{j-1}(\theta + i\sigma)}{\sigma^j (\frac{\theta+1}{\sigma})_{j\uparrow}}, \qquad j = 1, 2, \ldots,$$

for some $0 \leq \sigma < 1$ and $\theta > -\sigma$, and therefore

$$\frac{\mathbf{E}(w_j)}{\mathbf{E}(w_{j+1})} = 1 + \frac{1}{\theta + j\sigma} > 1$$

for all the admissible values of σ and θ.

§ 2 Decreasing-weight mixture models

The geometric process has become an appealing option for different problems. Besides of density estimation (Fuentes-García et al. 2010*a*), it has been applied in regression (Fuentes-García et al. 2009), dependent models (Mena et al. 2011), classification (Gutiérrez et al. 2014), and others. In the next sections, two families of sequences of decreasing weights are studied. Both are based on the different constructions, already explained, of this process.

2.1 PARTIAL-SUM BASED WEIGHTS

From the derivation of the geometric process in the previous section, we can devise two approaches to construct sequences of decreasing weights. In the first approach, we require a discrete probability distribution, $\pi(r;\lambda)$—supported on $\mathbb{N}$—and compute the partial sum in Equation (9). The second approach consists on selecting a random variable X—taking values also in $\mathbb{N}$—with an L-shape distribution, i.e. a probability distribution satisfying

$$\mathbf{P}(X = x) > \mathbf{P}(X = x + 1), \qquad x = 1, 2, \dots,$$

and, then, define the jth weight as $w_j := \mathbf{P}(X = j)$. This method comes from Equation (10) where X follows a geometric distribution. With regard to the first approach—in this section—we work with the family of *power series distributions*.

DEFINITION 1. A distribution F is a *power series distribution* if its probability density function can be written as

$$F(k) = \mathbf{P}(X = k) = \frac{a_k(\varphi)\theta^k}{f(\theta,\varphi)}, \quad k = 1, 2, \dots$$

for $\theta > 0$ and where the coefficients $a_1(\varphi), \dots, a_k(\varphi), \dots$ are all positive and depend only on k and φ. Using these coefficients, the function f is defined as

$$f(\theta,\varphi) = \sum_{k=1}^{\infty} a_k(\varphi)\theta^k,$$

and it is positive, finite and differentiable for all the admissible values of θ and φ.

This definition has been introduced by Patil (1964) as a two-parameter version of the power series distribution generally credited to Noack (1950), which depends only on θ. Furthermore, notice that the value zero has been excluded from the support in order apply Equation (9). However, it is straightforward to write any well known distribution in terms of this definition.

Using this class of probability distributions, we can apply (9) to get

$$w_j = \sum_{r=j}^{\infty} \frac{a_r(\phi)\theta^r}{r f(\theta,\phi)}, \qquad j = 1, 2, \dots. \tag{11}$$

It is necessary to specify the form of the sequence $(a_r(\phi))_{r\geq 1}$ and their corresponding function $f(\theta,\phi)$—or equivalently—the power series distribution. We study, in the rest

of this section, three examples of power series distributions: negative binomial, Poisson and logarithmic.

EXAMPLE 1 (Negative binomial based case). Let $\theta = 1 - \lambda$, $f(\lambda, s) = (1 - \lambda)/\lambda^s$ and

$$a_k(s) = \binom{k + s - 2}{k - 1}, \qquad k = 1, 2, \dots,$$

for some $0 < \lambda < 1$ and $r > 0$. Then, a random variable X has a *negative binomial distribution* with parameters (s, λ), denoted by $\mathrm{NB}(s, \lambda)$, if its probability distribution can be written as

$$\mathbf{P}(X = x) = \binom{x + s - 2}{x - 1} \lambda^s (1 - \lambda)^{x-1}, \quad x = 1, 2, \dots. \tag{12}$$

Furthermore, $\mathbf{E}(X) = 1 + \frac{(1-\lambda)s}{\lambda}$, and $\mathrm{Var}(X) = \frac{(1-\lambda)s}{\lambda^2}$.

Therefore, if $r \sim \mathrm{NB}(s, \lambda)$, the weights $(w_j)_{j \geq 1}$ in Equation (11) are given by

$$w_j = \frac{1}{j} \binom{j + s - 2}{j - 1} \lambda^s (1 - \lambda)^{j-1} {}_2F_1(j + s - 1, 1, j + 1; 1 - \lambda), \qquad j = 1, 2, \dots, \tag{13}$$

where

$${}_2F_1(a, b, c; z) = \sum_{k=0}^{\infty} \frac{(a)_{k\uparrow} (b)_{k\uparrow}}{(c)_{k\uparrow}} \frac{z^k}{k!},$$

is the Gaussian hypergeometric function, and $(x)_{n\uparrow}$ denotes the Pochhammer symbol.

EXAMPLE 2 (Poisson based case). The Poisson distribution with parameter λ, denoted by $\mathrm{Poi}(\lambda)$, can be expressed as a power series distribution by letting $f(\lambda) = \lambda e^{\lambda}$, for $\lambda > 0$, and taking the coefficients $(a_k)_{k \geq 1}$ from the series expansion of f, i.e.

$$a_k = \frac{1}{(k - 1)!}, \qquad k = 1, 2, \dots.$$

Thus, if $X \sim \mathrm{Poi}(\lambda)$, its probability distribution is such that

$$\mathbf{P}(X = x) = \frac{\lambda^{x-1} e^{-\lambda}}{(x - 1)!}, \qquad x = 1, 2, \dots. \tag{14}$$

Moreover, $\mathbf{E}(X) = 1 + \lambda$, and $\mathrm{Var}(X) = \lambda$.

Letting $r \sim \mathrm{Poi}(\lambda)$ in Equation (11), the corresponding weights are

$$w_j = \frac{\Gamma(j) - \Gamma(j, \lambda)}{\lambda \Gamma(j)}, \qquad j = 1, 2, \dots, \tag{15}$$

where $\Gamma(a, z)$ is the (upper) incomplete Gamma function, which has the following repre-

sentation

$$\Gamma(a, z) = \frac{\Gamma(a)}{e^z} \sum_{k=0}^{a-1} \frac{z^k}{k!},$$

for any positive integer a.

EXAMPLE 3 (Logarithmic based case). The logarithmic distribution is a power series distribution obtained from the Taylor expansion at zero of the function $f(\lambda) = -\log(1-\lambda)$, for $0 < \lambda < 1$. In this case, the coefficients are given by $a_k = 1/k$, $k = 1, 2, \ldots$. Then, a random variable X having logarithmic distribution with parameter λ, denoted by $\text{Log}(\lambda)$, is such that

$$\mathbf{P}(X = x) = \frac{1}{-\log(1-\lambda)} \frac{\lambda^x}{x}, \qquad x = 1, 2, \ldots .$$

In this case,

$$\mathbf{E}(X) = \frac{1}{-\log(1-\lambda)} \frac{\lambda}{(1-\lambda)} \quad \text{and} \quad \text{Var}(X) = \frac{-\lambda(\lambda + \log(1-\lambda)}{(1-\lambda)^2 \log^2(1-\lambda)}.$$

Using this distribution for r in Equation (11), the weights $(w_j)_{j\geq 1}$ are such that

$$w_j = \frac{\lambda^j \Phi(\lambda, 2, j)}{(-\log(1-\lambda))}, \qquad j = 1, 2, \ldots, \tag{16}$$

where

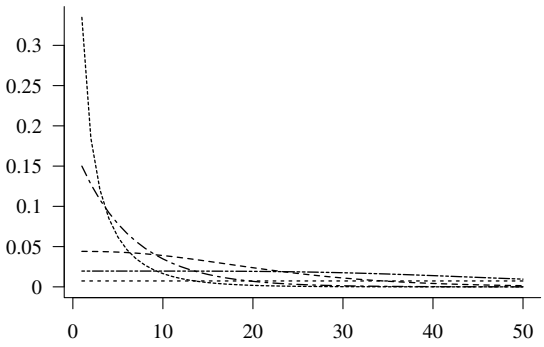
$$\Phi(z, s, a) = \sum_{k=0}^{\infty} \frac{z^k}{(k+a)^s}$$

is known as the Lerch transcendent.

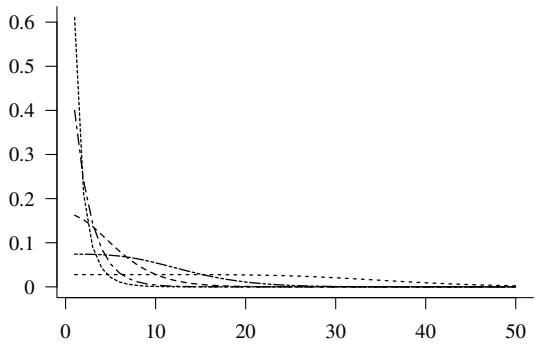
The decreasing rate of these cases varies according to the value of their parameters. Note that the geometric process is obtained from the negative binomial (NB) case when $s = 2$, so we can say that the NB case is a generalization of the geometric. In Figure 1, different examples for these three cases are displayed.

2.2 L-SHAPE DISTRIBUTION BASED WEIGHTS

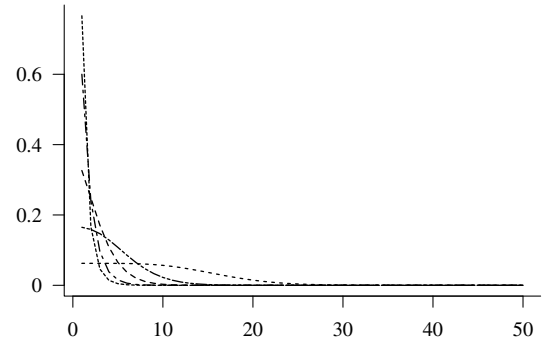
The second approach explained above to obtain sequences of decreasing weights requires an L-shape probability distribution taking values in $\mathbb{N}$ and with mode at 1, that is, a probability function such that $\mathbf{P}(X = x) > \mathbf{P}(X = x + 1)$ for all $x \geq 1$. For the geometric process (10), a geometric distribution shifted at one is used, but there are other distributions satisfying these constrains, for example: the logarithmic and Zeta distributions as well as others belonging to the Lagrangian family. The logarithmic distribution has



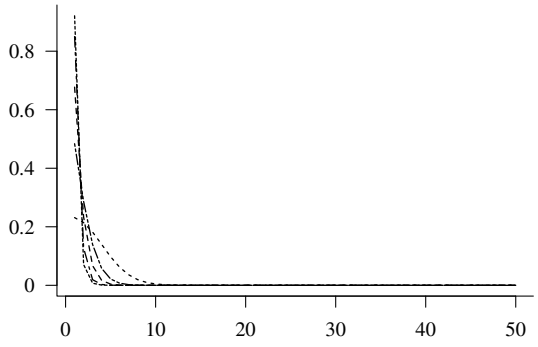
(A) NB case for $\lambda = 0.15$ and different values of s: ----- 1, —-- 2, --- 5, ---· 10 and ---- 25.



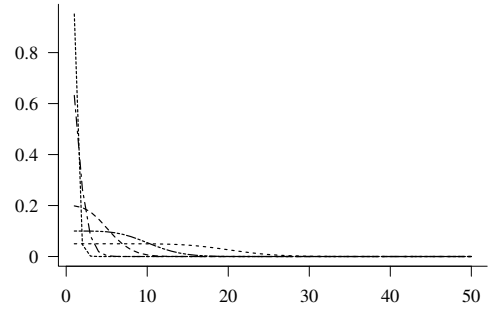
(B) NB case for $\lambda = 0.4$ and different values of s: ----- 1, —-- 2, --- 5, ---· 10 and ---- 25.



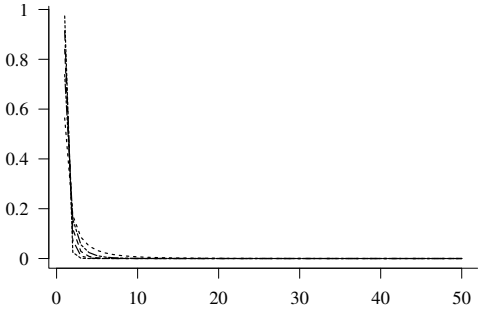
(C) NB case for $\lambda = 0.6$ and different values of s: ----- 1, —-- 2, --- 5, ---· 10 and ---- 25.



(D) NB case for $\lambda = 0.85$ and different values of s: ----- 1, —-- 2, --- 5, ---· 10 and ---- 25.



(E) Poisson base weights for different values of λ: ----- 0.1, —-- 1, --- 5, ---· 10 and ---- 20.



(F) Logarithmic based weights for different values of λ: ----- 0.1, —-- 0.3, --- 0.5, ---· 0.7 and ---- 0.9.

FIGURE 1: Decreasing weights for the different constructions based on power series distributions. Points are connected by straight lines for visual simplification.

already been studied, and the rest of them are explained next.

EXAMPLE 4 (Zeta distribution). A random variable X has a *Zeta distribution* with parameter $s > 1$, if its probability distribution is given by

$$\mathbf{P}(X = k) = \frac{k^{-s}}{\zeta(s)}, \qquad k = 1, 2, \dots,$$

where $\zeta(s)$ is the Riemann zeta function defined as

$$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}.$$

Lagrangian distributions are another wide family of discrete probability distributions whose early examples include the Otter's multiplicative process (Otter 1949) and the Borel distribution (Haight & Breuer 1960). A comprehensive study of this methodology can be found in Consul & Famoye (2006). Based on their classification, we focus on three examples belonging to the class of *basic Lagrangian distributions*.

DEFINITION 2. Let $g(z)$ be a successively differentiable function such that $g(1) = 1$ and $g(0) \neq 0$. Then a *basic Lagrangian distribution* is defined as

$$\mathbf{P}(X = x) = \frac{1}{x!}\left[\frac{\mathrm{d}^{x-1}}{\mathrm{d}z^{x-1}}\big(g(z)\big)^x\right]_{z=0}, \qquad x = 1, 2, \dots.$$

EXAMPLE 5 (Borel distribution). Let $g(z) = \exp(\lambda(z-1))$ for $0 < \lambda < 1$. A random variable X has a *Borel distribution* with parameter λ if

$$\mathbf{P}(X = x) = \frac{(x\lambda)^{x-1}}{x!} e^{-x\lambda}, \qquad x = 1, 2, \dots.$$

EXAMPLE 6 (Consul distribution). Let $g(z) = (1 - \theta + \theta z)^m$ for $0 < \theta < 1$ and $m \geq 1$. A random variable X has a *Consul distribution* with parameters (θ, m) if

$$\mathbf{P}(X = x) = \frac{1}{x}\binom{mx}{x-1}\theta^{x-1}(1-\theta)^{(m-1)x+1}, \qquad x = 1, 2, \dots.$$

EXAMPLE 7 (Geeta distribution). Let $g(z) = (1-\theta)^{m-1}(1-\theta z)^{1-m}$ where θ and m are such that $0 < \theta < 1$ and $1 < m < \theta^{-1}$. A random variable X has a *Geeta distribution* with parameters (θ, m) if

$$\mathbf{P}(X = x) = \frac{1}{mx-1}\binom{mx-1}{x}\theta^{x-1}(1-\theta)^{(m-1)x}, \qquad x = 1, 2, \dots.$$

These three Lagrangian distributions have mode at $x = 1$ for all values of their respective parameters (see Consul & Famoye 2006, for a proof of each case).

2.3 Performance in density estimation

In this section, we test the two extensions of the geometric process studied above. First, we explain the Gibbs sampler for each case, and then, we Illustrate their performance with simulated datasets under the context of density estimation.

2.3.1 Partial-sum approach

Considering the partial-sum approach, we have the following mixture model

$$f(y_i|r_i) = \frac{1}{r_i}\sum_{j=1}^{r_i} \kappa(y_i; x_j), \tag{17}$$

for $i = 1, \dots, n$. In order to derive a Gibbs sampler scheme, we introduce a latent variable d_i associating each observation y_i to a specific component, so $f(y_i|r_i, d_i, x) = \kappa(y_i; x_{d_i})$. Thus, written in hierarchical form, the mixture model is written as

$$\begin{aligned}
y_i|x, r_i, d_i &\overset{\text{ind.}}{\sim} \kappa(y_i; x_{d_i}), \qquad i = 1, \dots, n \\
d_i|r_i &\overset{\text{ind.}}{\sim} \mathrm{U}(1, r_i) \\
r_i|\lambda &\overset{\text{i.i.d.}}{\sim} \pi(\lambda) \\
x_j &\overset{\text{i.i.d.}}{\sim} \nu_0(\theta) \\
\lambda &\sim \phi(\omega)
\end{aligned}$$

where $j = 1, \dots, \max(r_1, \dots, r_n)$, ν_0 is a non-atomic distribution and θ and ω are finite dimensional parameters. Distributions π and ϕ are chosen according to one of the provided examples. Therefore, the full conditional distributions are the following:

1. Assuming ν_0 has density function g_0, the full conditional for the kernel parameter is

$$p(x_j|\cdots) \propto g_0(x_j; \theta) \prod_{d_i = j} \kappa(y_i; x_j),$$

for $j = 1, \dots, \max(r_1, \dots, r_n)$.

2. For the membership variables, we have

$$p(d_i|\cdots) \propto \kappa(y_i; x_{d_i})\mathbf{1}(1 \leq d_i \leq r_i),$$

for $i = 1, \ldots, n$.

3. In the case of the number of components, the full conditional is given by

$$p(r_i|\cdots) \propto \frac{\pi(r_i;\eta)}{r_i}\mathbf{1}(r_i \geq d_i).$$

4. Finally, parameter λ is updated by

$$p(\lambda|\cdots) \propto \phi(\lambda;\omega)\prod_{i=1}^{n}\pi(r_i;\lambda).$$

For the last two distributions, we use Examples 1–3 for π and, then, ϕ becomes a beta or a gamma distribution. The corresponding full conditional distributions are the following.

Negative binomial based case. Assuming $r_i \sim \text{NB}(s,\lambda)$ for $i = 1, \ldots, n$, and $\lambda \sim \text{Be}(\alpha,\beta)$, their conditional distributions are

$$p(r_i|\cdots) \propto \binom{r_i + s - 2}{r_i}(1-\lambda)^{r_i}\mathbf{1}(r_i \geq d_i),$$

that is, a negative binomial distribution with parameters $(s-1,\lambda)$ truncated at $r_i \geq d_i$, and

$$p(\lambda|\cdots) \propto \lambda^{sn+\alpha-1}(1-\lambda)^{\sum_{i=1}^{n} r_i+\beta-n-1},$$

which is a beta distribution with parameters $(sn+\alpha, \sum_{i=1}^{n} r_i + \beta - n)$. A simulation procedure from a truncated negative binomial distribution is presented in Appendix B.2.

Poisson based case. If it is assumed $r_i \sim \text{Poi}(\lambda)$ for all $i = 1, \ldots, n$, and $\lambda \sim \text{Ga}(\alpha,\beta)$, their corresponding conditional distributions are

$$p(r_i|\cdots) \propto \frac{\lambda^{r_i}}{r_i!}\mathbf{1}(r_i \geq d_i),$$

which is a Poisson distribution with parameter λ truncated at $r_i \geq d_i$, and

$$p(\lambda| \cdots) \propto \lambda^{\sum_{i=1}^n r_i + \alpha - n - 1} e^{-(\beta+n)\lambda},$$

a gamma distribution with parameters $(\sum_{i=1}^n r_i + \alpha - n, \beta + n)$. In Appendix B.1, a procedure is also described to simulate from a truncated Poisson distribution.

Logarithmic based case. The third option is when $r_i \sim \text{Log}(\lambda)$, $i = 1, \ldots, n$. Then, the conditional distribution, for all i, is

$$p(r_i| \cdots) \propto \frac{\lambda^{r_i}}{r_i^2} \mathbf{1}(r_i \geq d_i),$$

and setting $\lambda \sim \text{Be}(\alpha, \beta)$, its conditional distribution is

$$p(\lambda| \cdots) \propto \frac{\lambda^{\sum_{i=1}^n r_i + \alpha - 1}(1-\lambda)^{\beta-1}}{(-\log(1-\lambda))^n}.$$

Although neither of these distributions has a known form, it is possible to simulate from them by introducing latent variables. For the distribution of r_i, let $v \sim \text{Ga}(2, r_i)$, then

$$p(r_i, v| \cdots) \propto \lambda^{r_i} v e^{-r_i v} \mathbf{1}(r_i \geq d_i).$$

Note that $0 < \lambda e^{-v} < 1$, therefore

$$p(r_i|v, \cdots) \propto (\lambda e^{-v})^{r_i} \mathbf{1}(r_i \geq d_i)$$

corresponds to a geometric distribution with parameter $1 - \lambda e^{-v}$ truncated at $r_i \geq d_i$, which is straightforward to simulate. Similarly for parameter λ, let $z \sim \text{Ga}(n, \gamma)$ for $\gamma = -\log(1-\lambda)$. Then

$$p(\lambda, z| \cdots) \propto \lambda^{\alpha'-1}(1-\lambda)^{\beta-1} z^{n-1} e^{-\gamma z},$$

for $\alpha' = \sum_{i=1}^n r_i + \alpha$, thus

$$p(\lambda|z, \cdots) \propto \lambda^{\alpha'-1}(1-\lambda)^{\beta+z-1}.$$

which it is a beta distribution with parameters $(\sum_{i=1}^n r_i + \alpha, z + \beta)$.

For all these cases, the Monte Carlo density estimator is given by

$$\hat{f}(y) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i^{(t)}}\sum_{j=1}^{r_i^{(t)}}\kappa(y; x_j^{(t)}), \tag{18}$$

where T is the MCMC sampling size and the superscript (t) denotes the values sampled at iteration t. Equivalently, this estimator can be written as

$$\hat{f}(y) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{R^{(t)}} w_j^{(t)}\kappa(y; x_j^{(t)}), \tag{19}$$

with $R^{(t)} = \max_i(r_i^{(t)})$ and where the weights are given by

$$w_j^{(t)} = \sum_{i=1}^{n}\frac{\mathbf{1}(r_i^{(t)} \geq j)}{nr_i^{(t)}} = \sum_{r_i^{(t)}\geq j}\frac{1}{nr_i^{(t)}}. \tag{20}$$

We can also estimate these weights by substituting the sampled values of λ in the corresponding equation: (13), (15) or (16). Under this second estimation, it is important to pay attention to the sum of the mixing weights since it has to be one; if it is not the case, for the missing proportion, its kernel parameter can be sampled from the prior.

2.3.2 *L*-shape approach

Using the approach of Section 2.2, we obtain a mixture density of the form

$$f(y) = \sum_{j=1}^{\infty} w_j\kappa(y; x_j). \tag{21}$$

Inferences can be made through the slice sampler (Walker 2007, Kalli et al. 2011). Following the more general approach of Kalli et al. (2011), a fixed sequence $\xi = (\xi_1, \xi_2, \dots)$ and two latent variables (d, u) and are introduced such that $f(y|d, u, \xi, x) = w_d\kappa(y; x_d)/\xi_d\mathbf{1}(u < \xi_d)$. Thus, a Gibbs sampler can be implemented, where the full conditional distributions are given by

1. Like in the previous model, the full conditional for the kernel parameter is

$$p(x_j|\cdots) \propto g_0(x_j;\theta)\prod_{d_i=j}\kappa(y_i; x_j),$$

where g_0 is the density function of ν_0.

2. For the membership variables d_i, we have

$$p(d_i|\cdots) \propto w_{d_i}\kappa(y_i; x_{d_i})/\xi_{d_i}\mathbf{1}(d_i : \xi_{d_i} > u_i),$$

for $i = 1, \ldots, n$.

3. In the case of the second latent variable, u_i,

$$p(u_i|\cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i}),$$

$i = 1, \ldots, n$.

4. Finally, the sequence of weights is updated as follows

$$p(w_j|\cdots) \propto p(w_j)\prod_{i=1}^{n} w_{d_i}\mathbf{1}(d_i = j).$$

As pointed out by the authors, the size of the sequences $(x_1, x_2, \ldots)$ and $(w_1, w_2, \ldots)$ is determined by the maximum value of the membership variables $(d_1, \ldots, d_n)$.

Regarding the updating of the sequence of weights, the last step depends on how each weight is defined. Consider for example, the Dirichlet process. Its sequence of weights can be built using the stick-breaking representation (4) with $v_j \overset{\text{i.i.d.}}{\sim} \text{Be}(1, c)$, for $c \geq 0$. Therefore, the full conditional for each v_j is a Beta distribution with parameters $(1 + \sum_{i=1}^{n}\mathbf{1}(d_i = j), c + \sum_{i=1}^{n}\mathbf{1}(d_i > j))$. On the other hand—when working with the geometric process—there is only one random variable, $\lambda \in (0, 1)$, defining the whole sequence of weights. If we assume $\lambda \sim \text{Be}(a, b)$, its full conditional is also a beta distribution but with parameters $(a + n, b + \sum_{i=1}^{n} d_i - n)$.

The Monte Carlo density estimator for this second model is given by

$$\hat{f}(y) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{r^{(s)}}\sum_{j=1}^{r^{(s)}}\kappa(y; x_j^{(s)}),$$

where $r^{(s)}$ is the size of the sequence of sampled kernel parameters $x_j^{(s)}$. We can use also the estimator (19) together with the corresponding construction for the mixing weights. The study of this second class of decreasing weight mixture models will be pursued elsewhere.

2.3.3 Simulation examples

Along this section, we explore the performance of the models under study using the partial-sum approach for the mixing weights. We use a Gaussian kernel with unknown mean and variance, so $x_j = (m_j, \tau_j)$ and $\kappa(y; x_j) = N(y; \mu_j, 1/\tau_j)$. A normal–gamma prior with parameters (m, c, a, b) is used for the parameters (μ_j, τ_j), that is

$$p(\mu_j, \tau_j) = \mathrm{N}(\mu_j; m, c/\tau_j)\mathrm{Ga}(\tau_j; a, b). \tag{22}$$

Thus, the posterior distribution is conjugate with parameters (m', c', a', b') given by

$$\begin{aligned} m' &= \frac{cn_j\bar{y}_j + m}{cn_j + 1} & a' &= \frac{n_j}{2} + a \\ c' &= \frac{c}{cn_j + 1} & b' &= \frac{n_j(\bar{y}_j - m)^2}{2(cn_j + 1)} + \frac{S_j}{2} + b, \end{aligned}$$

where $\bar{y}_j = \frac{1}{n_j}\sum_{d_i=j} y_i$, $S_j = \sum_{d_i=j}(y_i - \bar{y}_j)^2$ and $n_j = \sum_{j=1}^{n} \mathbf{1}(d_i = j)$.

Given this data model, we test the different decreasing weight cases under different scenarios. For this purpose we use two datasets. The first dataset is a sample of size 1 000 from the following mixture density

$$\begin{aligned} f(y) = &\tfrac{39}{100}\mathrm{N}(y; 3, 1) + \tfrac{21}{100}\mathrm{N}(y; 5.5, 4) + \tfrac{15}{100}\mathrm{N}(y; 15, 1) + \\ &\tfrac{10}{100}\mathrm{N}(y; 20, 0.25) + \tfrac{15}{100}\mathrm{N}(y; 22, 16). \end{aligned} \tag{23}$$

The second dataset consists also on a sample of size 1 000 from the mixture density

$$f(y) = \sum_{j=1}^{10} \tfrac{1}{10}\mathrm{N}(y; 6j - 33, 1). \tag{24}$$

For both datasets, 50 000 iterations of the MCMC scheme were obtained; the first 30 000 of them were discarded. For all the specifications, we assign vague priors by setting $(m, c, a, b) = (0, 1000, 0.01, 0.01)$ for the kernel hyperparameters, and letting $\lambda \sim \mathrm{Be}(1, 1)$ for the NB case, and $\lambda \sim \mathrm{Ga}(0.01, 0.01)$ for the Poisson case. Additionally, we fix the second parameter in the NB case, s, taking the values 1, 2 (corresponding to the geometric process), 5, 10 and 25.

In Figures 2–7, different results of the simulations are shown; for each dataset, the estimated densities, the means of the posterior weights and the posterior distribution

of the number of components, r, are displayed. Regarding the estimated weights, two estimations were computed: the first one is given by (20), whereas the second one has been obtained by plugging in the sampled values of λ in the corresponding definition of the weights (Equations 13 or 15). We can see that both estimations are quite similar for the two datasets (Figures 3 and 6).

Focusing on the estimated density, for the first dataset—drawn from the 5-component mixture model (Equation 23)—it is displayed in Figure 2. Although there are some variations, the prior specifications for the mixing weights almost do not affect the estimations. However, the most interesting results to be considered are those of the second dataset, drawn from the 10-component mixture model (Equation 24).

In this second case, the estimated densities are shown in Figure 5. It is clear that for some prior specifications, the estimation is not good. Analyzing the posterior weights (Figure 6) and the posterior distribution of the number of components (Figure 7), we can say that the fact that the mixture model (24) has the same weight for each component is reflected on the necessity of having many sampled components accumulated around a single location in order to recover it. Therefore, the flexibility of the mixing weights' structure plays a crucial role. We can see that—for the scenarios where the posterior mixing weights have heavier tails (cf. Figure 6d)—the estimated mixing weights replicate better those of the data generating process. This gets closer to the desirable interpretation of the estimated components of the mixture as being those replicating the true mixing proportions.

For the NB case, this effect is achieved for big values of s; actually, it can be appreciated that as it increases, the density estimator adjusts better to the original function. Furthermore, the number of components, r_i, helps also to understand this idea. Heavier-tail mixing weights allows for more components to be captured, which help us to concentrate on its posterior distribution around the true value in examples like the one at issue, i.e. with many components (cf. Figure 7). Conversely, for the case $s = 1$, where more prior mass is accumulated in fewer components, the sampler has to simulate more components in order to attempt recovering each original mixture component, which still doing it fail.

These observations clearly ponder with the flexibility of the model to adapt their tail behavior to the data structure. For example, in the Poisson case, where the bigger the λ the heavier the tails (cf. Figure 1e), placing a prior on λ allows the model to adapt to the required tail behavior. Indeed—while, *a priori*, its weights' structure was similar to that of the NB case with $s = 1$—we observe that a posteriori it replicates the required tail behavior, being comparable with that of the NB case with $s = 25$. If we let $p = 1 - \frac{\lambda}{s+\lambda}$, for

some $\lambda, s > 0$, then $0 < p < 1$. Hence, setting (p, s) the parameters of the NB probability function (12), we have

$$\binom{x+s-2}{x-1} p^s (1-p)^{x-1} = \frac{\lambda^{x-1}}{(x-1)!} \frac{\Gamma(x+s-1)}{(s+\lambda)^{x-1}\Gamma(s)} \left(1+\frac{\lambda}{s}\right)^{-s}.$$

Taking its limit as $s \to \infty$, the middle term converges to one, whereas for the last term, we have $\left(1+\frac{\lambda}{s}\right)^{-s} \to e^{-\lambda}$. Thus, the probability function (14) is recovered. This tells us that when randomizing the parameters we can achieve similar adaptation in the tails with either model, NB or Poisson.

As a side note, the results for the Logarithmic based case (not shown) resulted worse than the ones for the NB case with $s = 1$. Therefore, we concentrated our discussion only on the other two cases.

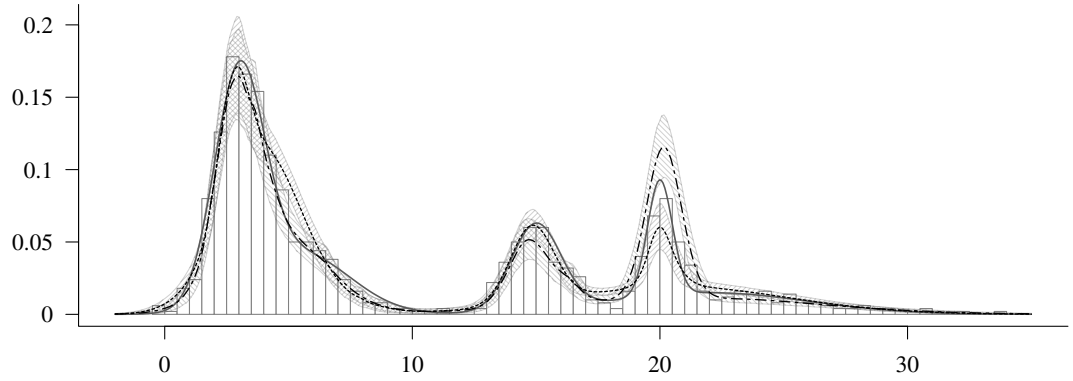
§ 3 Clustering via mixture models

The mixture models we have studied along this chapter, as explained at the beginning of Section 1.2, are obtained by fixing a kernel function κ and mixing over its parameter, where the mixing distribution is a discrete RPM. This latter fact implies the existence of a random partition which models the clustering structure of the data, but—unlike the explanation of Chapter 1—these clusters are formed by observations modeled by the same kernel parameters. We can find some mixture models making use of random partitions, for example, Neal (2000), Ishwaran & James (2001) and Lijoi et al. (2007*b*), which are mainly based on Pólya urn schemes. However, as we will see, they also appear in the models studied previously.

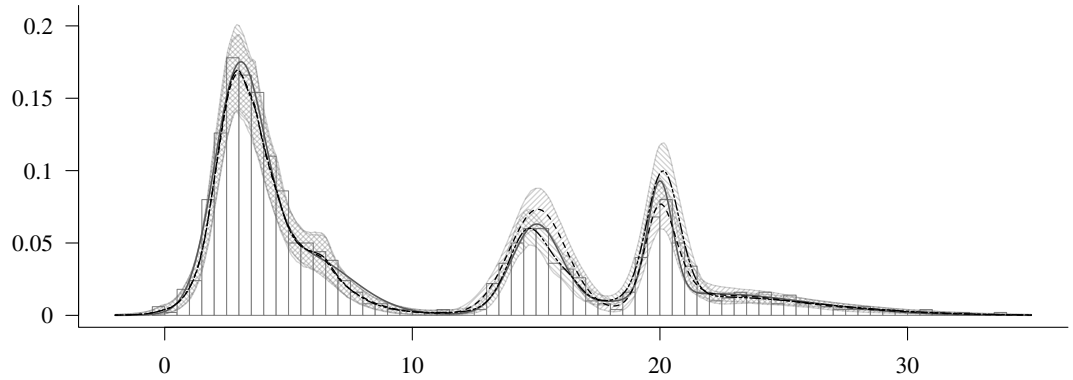
Assume we have a sample $(y_1, \dots, y_n)$ of size n from the mixture model (3) with a discrete RPM $\tilde{p}$ as mixing distribution. Written in a hierarchical form—following (5)—we have

$$\begin{aligned} y_i | x_i &\overset{\text{ind.}}{\sim} \kappa(y_i; x_i), \quad i = 1, \dots n \\ x_i | \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p} \\ \tilde{p} &\sim Q. \end{aligned}$$

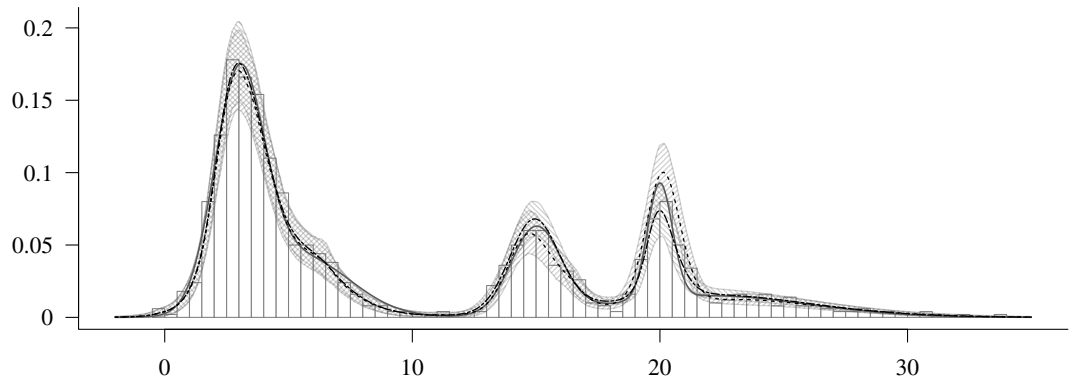
As we can see, due to the discreteness nature of $\tilde{p}$, any sample $(x_1, \dots, x_n)$ from $\tilde{p}$, of size n, contains only $k^* \leq n$ different values, say $(x_1^*, \dots, x_{k^*}^*)$, cf. Section 1.1.1.1. In other words, a random partition π^* is induced. We can encode this random partition through the usage



(A) Estimated density for the NB case with $s =$ ····· 1, and −·− 2.

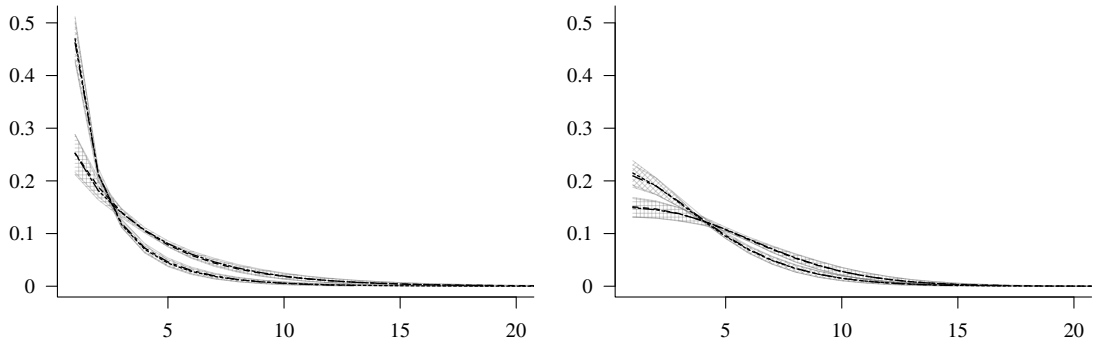


(B) Estimated density for the NB case with $s =$ - - - 5, and ·-·- 10.



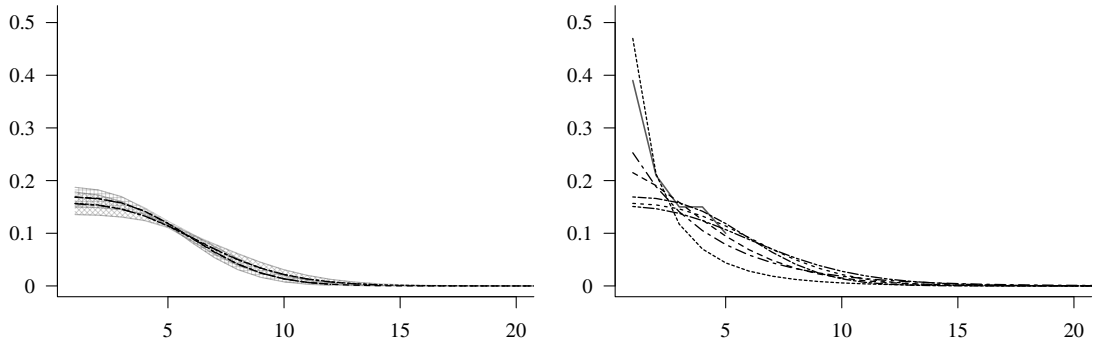
(C) Estimated density for the NB case with ···· $s = 25$, and ·-·- the Poisson case.

FIGURE 2: Estimated density of Model (23)—using estimator (19)—for each prior specification; 95% highest posterior density intervals are included (shaded areas). The data and the real density function are shown in gray.



(A) NB case. Estimations were computed using Equation (13), for $s =$ ····· 1 and --- 2, and Equation (20), $s =$ –·– 1 and ·–· 2.

(B) NB case. Estimations were computed using Equation (13), for $s =$ ····· 5 and --- 10, and Equation (20), $s =$ –·– 5 and ·–· 10.



(C) NB and Poisson cases. Estimations were computed using ····· Equation (13) with $s = 25$, --- Equation (15), and Equation (20), for –·– $s = 25$ and for ·–· the Poisson case.

(D) Posterior mean using Equations (13) and (15): NB case with $s =$ ····· 1, –·– 2, --- 5, ·–· 10 and ···· 25 and the ·–· Poisson case. True mixing weights are displayed in gray.

FIGURE 3: Posterior mean of the mixing weights for each prior specification with their 95% highest posterior density interval. Points are connected by straight lines for visual simplification.

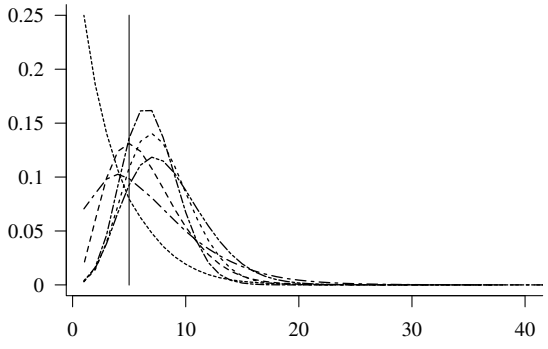
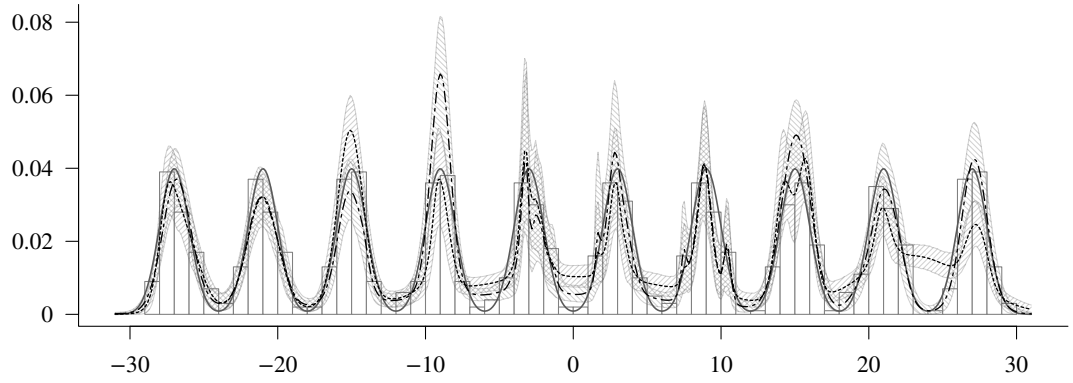
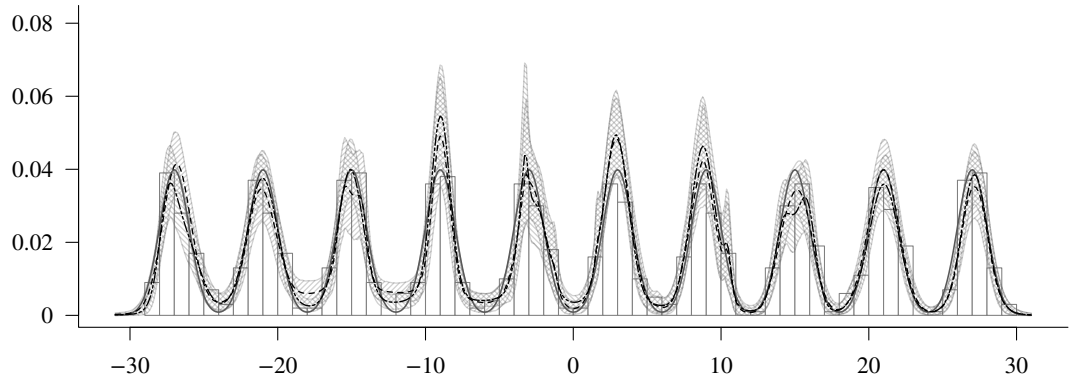


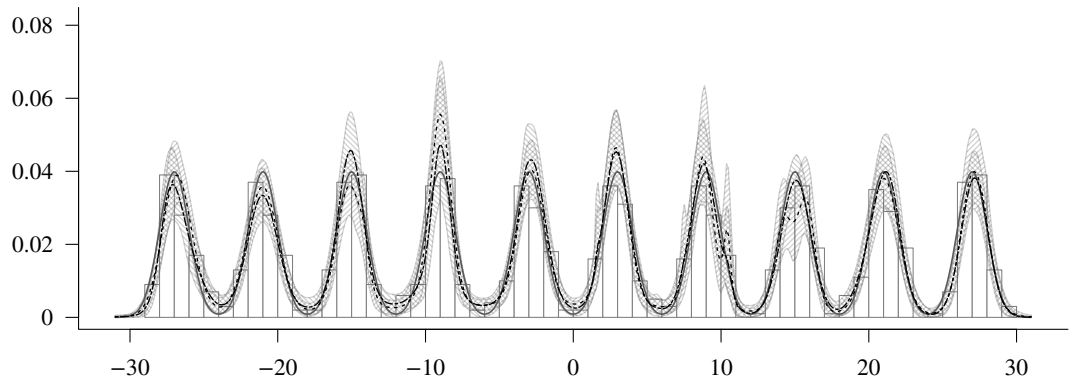
FIGURE 4: Posterior distribution of the number of components, r: NB case with $s =$ ····· 1, –·– 2, --- 5, ·–· 10 and ···· 25 and the ·–· Poisson case. Points are connected by straight lines for visual simplification. The vertical line indicates the true number of components, $r = 5$.



(A) Estimated density for the NB case with $s =$ ····· 1, and −·− 2.

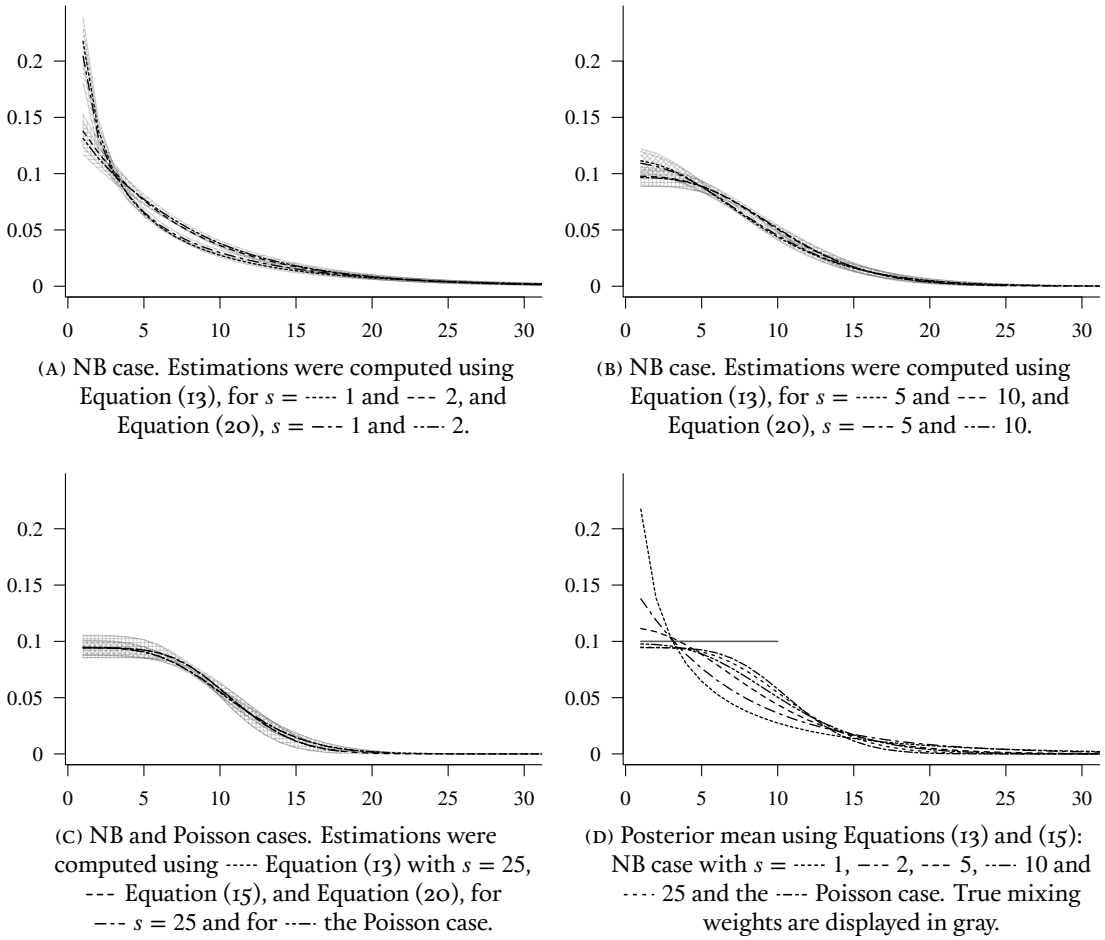


(B) Estimated density for the NB case with $s =$ − − − 5, and ·−· 10.



(C) Estimated density for the NB case with ····· $s = 25$, and ·−·· the Poisson case.

FIGURE 5: Estimated density of Model (24)—using estimator (19)—for each prior specification; 95% highest posterior density intervals are included (shaded areas). The data and the real density function are shown in gray.



(A) NB case. Estimations were computed using Equation (13), for s = ····· 1 and --- 2, and Equation (20), s = –·– 1 and ·–· 2.

(B) NB case. Estimations were computed using Equation (13), for s = ····· 5 and --- 10, and Equation (20), s = –·– 5 and ·–· 10.

(C) NB and Poisson cases. Estimations were computed using ····· Equation (13) with s = 25, --- Equation (15), and Equation (20), for –·– s = 25 and for ·–· the Poisson case.

(D) Posterior mean using Equations (13) and (15): NB case with s = ····· 1, –·– 2, --- 5, ·–· 10 and ···· 25 and the ·–· Poisson case. True mixing weights are displayed in gray.

FIGURE 6: Posterior mean of the mixing weights for each prior specification with their 95% highest posterior density interval. Points are connected by straight lines for visual simplification.

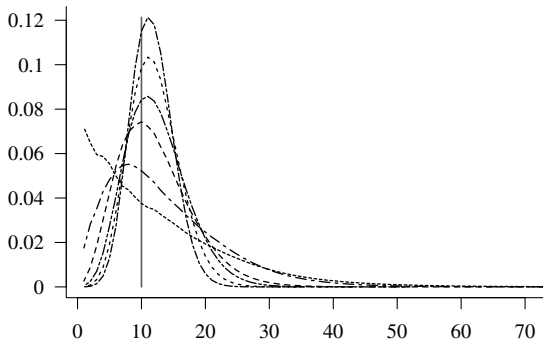


FIGURE 7: Posterior distribution of the number of components, r: NB case with s = ····· 1, –·– 2, --- 5, ·–· 10 and ···· 25 and the ·–· Poisson case. Points are connected by straight lines for visual simplification. The vertical line indicates the true number of components, $r = 10$.

of *membership sets* $\{\pi_1, \dots, \pi_k\}$ defined as

$$\pi_j = \{i : x_i = x_j^*,\ i = 1, \dots, n\}, \qquad j = 1, \dots, k^*. \tag{25}$$

Notice that each unique value x_j^* is associated to at least one observation, that is $\#\pi_j > 0$ for all $j = 1, \dots, k^*$. Thus, $\{\pi_1, \dots, \pi_{k^*}\}$ induces a partition of $\{1, \dots, n\}$, and then the clusters—at the level of the data—are formed by sequences of observations

$$\left\{y_{i_{1,1}}, \dots, y_{i_{1,n_1}}\right\}, \left\{y_{i_{2,1}}, \dots, y_{i_{2,n_2}}\right\}, \dots, \left\{y_{i_{k^*,1}}, \dots, y_{i_{k^*,n_{k^*}}}\right\},$$

such that $i_{j,l} \in \pi_j$ for $j = 1, \dots, k^*$.

This technique has been extensively employed by many authors (see, for example MacEachern 1994, Bush & MacEachern 1996, Griffin & Holmes 2010), most of the cases to improve sampling schemes—as we explained previously in this chapter—even though, as we can see, it is immediate to obtain samples from the posterior distribution of the random partition and of the number of groups.

In addition to this implicit usage of random partitions, we can state a hierarchical mixture model where they appear explicitly. In this case, we have

$$\begin{aligned} y_i | x_i, \pi^* &\overset{\text{ind.}}{\sim} \kappa(y_i; x_i) \qquad i = 1, \dots, n \\ x_j^* | \pi = \pi^* &\overset{\text{i.i.d.}}{\sim} \nu_0(x_j^*) \qquad j = 1, \dots, k^* \\ \pi &\sim P, \end{aligned} \tag{26}$$

with $\pi^* = \{\pi_1, \dots, \pi_{k^*}\} \in \mathcal{P}_{[n]}$ for $k^* \leq n$ and P is some prior distribution taking values in $\mathcal{P}_{[n]}$, for example, an EPPF (Section 1.1.1). The joint posterior distribution is given by

$$\begin{aligned} p(x^*, \pi^* | y) &\propto p(\pi^*) p(x^* | \pi^*) p(y | x^*, \pi^*) \\ &\propto p(\pi^*) \prod_{j=1}^{k^*} g_0(x_j^*) \prod_{i \in \pi_j} \kappa(y_i; x_j^*), \end{aligned}$$

where g_0 is the density function of ν_0. Hence, if we are interested only in the posterior distribution of the random partition, we integrate out parameters x_j^*, so

$$p(\pi^* | y) \propto p(\pi^*) \prod_{j=1}^{k^*} \int_{\mathbb{X}} \prod_{i \in \pi_j} \kappa(y_i; x) g_0(\mathrm{d}x).$$

On the other hand, random partitions can be also used to provide an estimate of the density f. Let $n_j^* = \#\pi_j$ for $j = 1, \dots, k^*$. Therefore, we have the Monte Carlo estimate

$$\tilde{f}(y) = \frac{1}{S} \sum_{s=1}^{S} \sum_{j=1}^{k^{*(s)}} w_j^{(s)} \kappa(y; x_j^{*(s)}), \tag{27}$$

where the mixing weights are given by $w_j^{(s)} = n_j^{*(s)}/n$. Notice that this estimation is actually the same obtained under a Pólya urn scheme, given in Equation (7). Moreover, it can be used instead of the Monte Carlo estimates (8) and (18) for the approaches studied before.

As a side note, in sampling schemes like the ones studied in the previous section, there are some sampled values x_j^* not being associated to any observation y_i, $i = 1, \dots, n$, so we have $n_j = 0$ for some values of j. However, this does not modify the estimator $\tilde{f}$ in (27).

3.1 Numerical illustration

The performance of random partitions in mixture modeling is illustrated by considering a small simulated dataset of size 10. We focus on the posterior distribution of the random partition. For this purpose, we test two mixture models: those with underlying mixing distribution given by the geometric and Dirichlet RPMs; samples were drawn using, in both cases, the slice sampler (Section 1.1.1). We simulated datasets of size 10 from the mixture model

$$f(y) = 0.5\mathrm{N}(y; -3, 1/6) + 0.2\mathrm{N}(y; 0, 1/6) + 0.3\mathrm{N}(y; 3, 1/6),$$

and obtained 5 000 iterations from the MCMC after discarding 3 000. All this procedure was done 1 000 times. Regarding the hyperparameters, the Gaussian kernel (22) was used with base-measure parameters $(m, c, a, b) = (0, 100, 0.1, 0.1)$, and for the RPM parameters, we proceeded as follows. In the case of the geometric process, a beta prior distribution with parameters (α, β) was assigned to λ, thus, such parameters were fixed such that the prior expected number of groups, k^*, matches 2, 5 and 9 for a given variance. Afterwards, for the Dirichlet process, its total mass parameter was given by $\theta = (1 - \lambda)/\lambda$, where $\lambda = \alpha/(\alpha + \beta)$ for each pair (α, β) found for the geometric process. These parameters are presented in Table 1. The expressions for the expected value and for the variance, in the

$\mathbf{E}(k^*)$	α	β	$\mathrm{Var}(k^*)$	θ
2	0.680	0.119	0.25	0.175
5	1.881	3.093	0.99	1.644
9	0.110	1.980	0.50	18.000

TABLE 1: Hyperparameters for the 10-data example. Columns 2–4 correspond to the geometric case, whereas the last one is for the Dirichlet case.

geometric case, are the following:

$$\mathbf{E}(k^*) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \sum_{r=1}^{n} \binom{n}{r} \frac{(-1)^{r-1}\lambda^{\alpha+r-1}(1-\lambda)^{\beta-1}}{1-(1-\lambda)^r} \mathrm{d}\lambda,$$

and

$$\mathrm{Var}(k_n^*) = \mathbf{E}(k_{2n}^*) - \mathbf{E}(k_n^*) + \frac{2\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 I(\lambda, n)\lambda^{\alpha-1}(1-\lambda)^{\beta-1}\mathrm{d}\lambda,$$

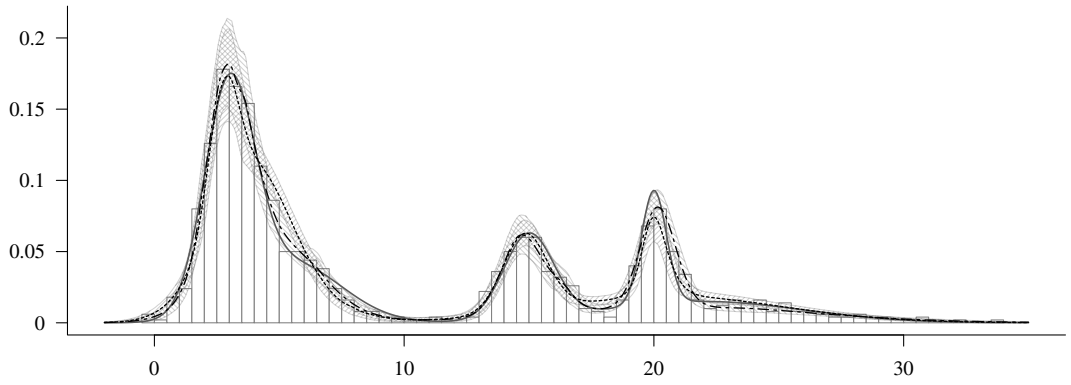
with

$$I(\lambda, n) = \sum_{r=1}^{n} \binom{n}{r} \frac{(-1)^r \lambda^r (1-\lambda)^r}{1-(1-\lambda)^r} \left(\sum_{s=0}^{r-1} \binom{r}{s} \frac{1}{(1-\lambda)^s - (1-\lambda)^r} - \sum_{s=0}^{n} \binom{n}{s} \frac{(-1)^s \lambda^s}{1-(1-\lambda)^{r+s}} \right);$$

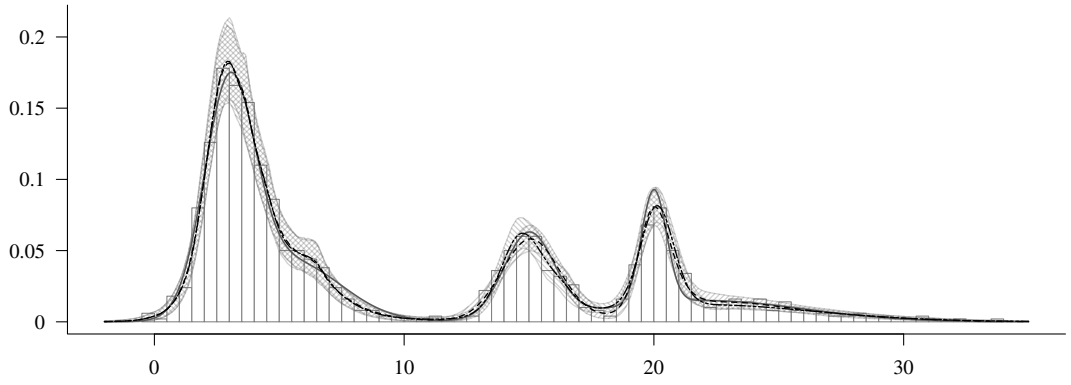
in this case, the subscript m in k_m indicates the sample size.

In Table 2, the results of these simulations are presented. From them, we can have a picture about the effect of the prior specification for each model. In the scenario $\mathbf{E}(k^*) = 2$, we assigned a very small variance, and as a consequence, both models detected the correct partition in less cases. Under the second scenario, $\mathbf{E}(k^*) = 5$, both models performed well, since there was more variability. Finally, for the misspecified scenario $\mathbf{E}(k^*) = 9$, the geometric was better. However, it is worth emphasizing that, in all these results, the geometric process recovered the true partition with higher probabilities than the Dirichlet case. The posterior mean of the number of components, k^*, also displayed in Table 2, helps to support this fact.

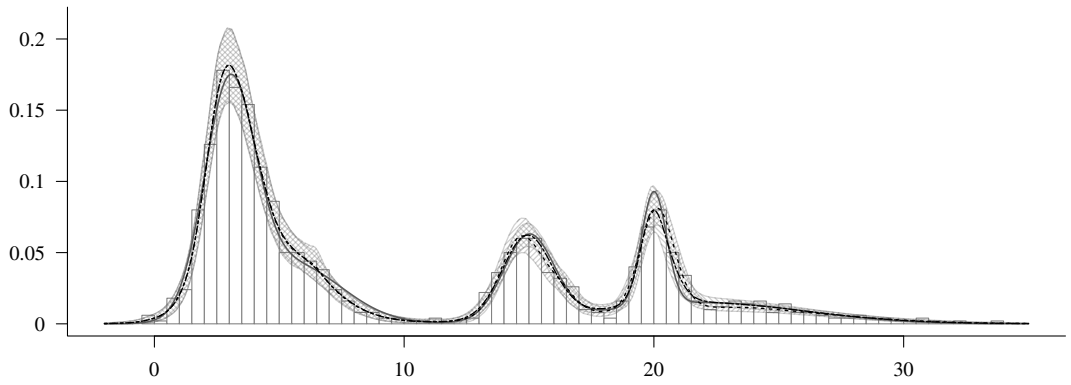
Finally, and for the sake of completeness, we compute the estimated density (27) for the simulated datasets of the previous section. Figure 8 shows this estimator $\tilde{f}$ for the data model (23), and Figure 9 shows it for the data model (24). In both cases, we can see an improvement in the estimation. Complementing the discussion of the number of groups, Figure 10 displays its posterior distribution. As it is expected—if we compare these distributions with 3d and 6d—the distribution of k^* is usually concentrated on bigger values than its corresponding distribution of the number of components, r.



(A) Estimated density for the NB case with s = ····· 1, and −·· 2.

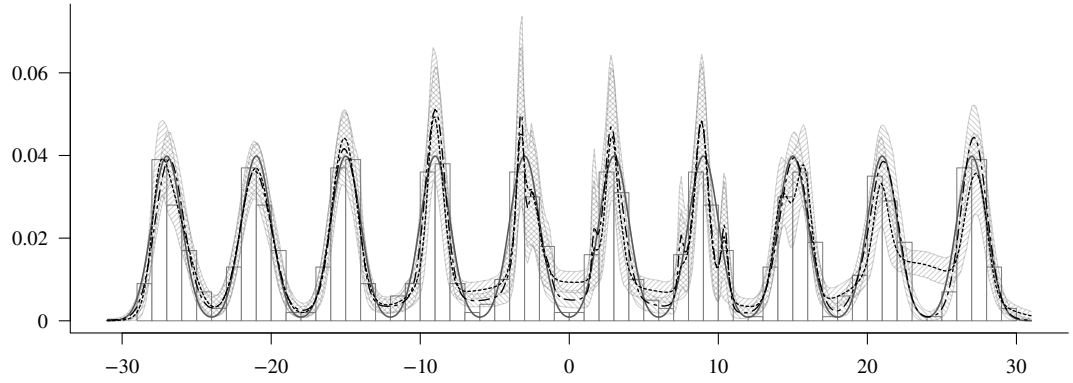


(B) Estimated density for the NB case with s = ··· 5, and ··–· 10.

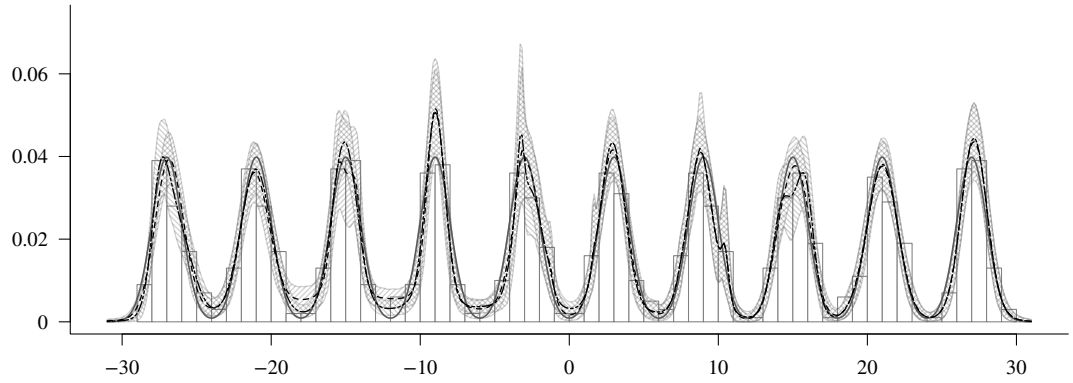


(C) Estimated density for the NB case with ···· s = 25, and ·–·· the Poisson case.

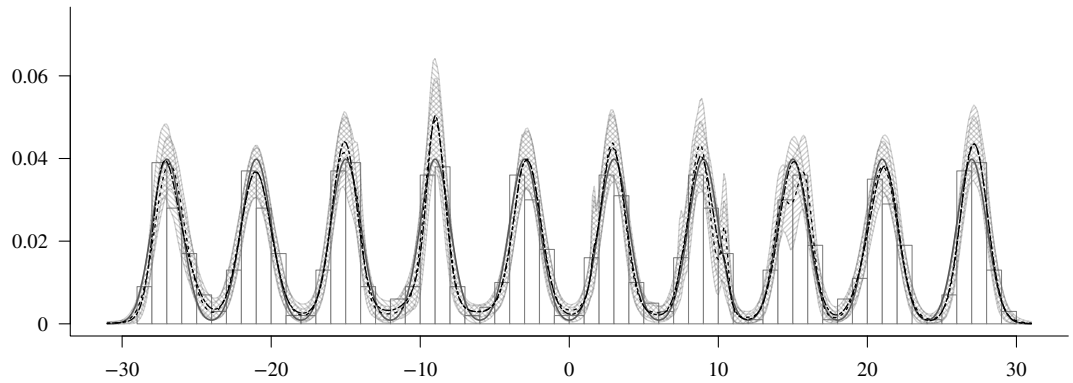
FIGURE 8: Estimated density of Model (24)—using estimator (27)—for each prior specification; 95% highest posterior density intervals are included (shaded areas). The data and the real density function are shown in gray.



(A) Estimated density for the NB case with $s =$ ····· 1, and −·− 2.



(B) Estimated density for the NB case with $s =$ − − − 5, and ··−·· 10.

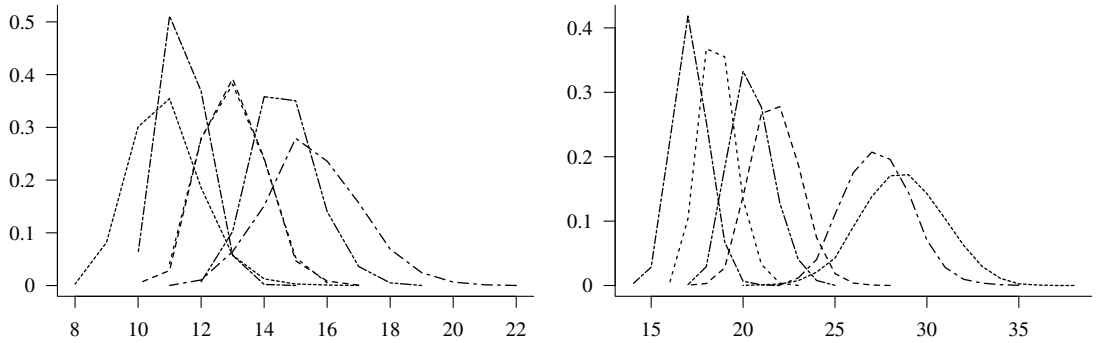


(C) Estimated density for the NB case with ····· $s = 25$, and ·−·− the Poisson case.

FIGURE 9: Estimated density of Model (24)—using estimator (27)—for each prior specification; 95% highest posterior density intervals are included (shaded areas). The data and the real density function are shown in gray.

	$\mathbf{E}(k^*) = 2$			$\mathbf{E}(k^*) = 5$			$\mathbf{E}(k^*) = 9$		
Process	π	$p(\pi)$	$\bar{k^*}$	π	$p(\pi)$	$\bar{k^*}$	π	$p(\pi)$	$\bar{k^*}$
Geometric	(10)	0.865	1.714	(5, 2, 3)	0.865	3.033	(5, 2, 3)	0.870	3.107
	(5, 2, 3)	0.122							
Dirichlet	(10)	0.519	1.815	(5, 2, 3)	0.825	2.906	(5, 1, 1, 3)	0.130	5.359
	(5, 2, 3)	0.061					(5, 2, 3)	0.104	

TABLE 2: Monte Carlo results for the 10-data example under the different prior specifications. For each process, the first row displays the modal partition, its probability and the posterior mean of the number of components. In the cases were this modal partition is not the true one, the second row displays the probability of the true partition. For simplicity, the partitions are represented by their blocks sizes, so, for example, (5, 2, 3) corresponds to the data partition $\{\{y_1, y_2, y_3, y_4, y_5\}, \{y_6, y_7\}, \{y_8, y_9, y_{10}\}\}$.



(A) 5-component mixture case (B) 10-component mixture case

FIGURE 10: Posterior distribution of the number of components, k^*: NB case with $s =$ ····· 1, –·· 2, – – – 5, –·– 10 and ···· 25 and the –·–· Poisson case. Points are connected by straight lines for visual simplification. The vertical lines indicate the true number of components.

3.2 EPPFs DERIVED FROM DECREASING-WEIGHT DISCRETE RPMs

As we see, mixture models are a natural tool to deal with the problem of clustering. The latent RPM which serves as mixing distribution induces a random partition grouping observations with the aid of the information provided by component parameters. This has been noticed by several authors and it has been well established for some examples of RPMs. However, the case of discrete RPMs with decreasing weights is so new that there are much work to be done. In the best of our knowledge, we can only mention the work of Mena & Walker (2012) on this direction.

From the definition of EPPF given in Equation 1.4, we know that—given a discrete RPM like in (2)—the EPPF of such random distribution is given by the expression

$$\sum_{j_1 \neq \cdots \neq j_k} w_{j_1}^{n_1} w_{j_2}^{n_2} \cdots w_{j_k}^{n_k},$$

where—for a given $n \geq 1$ and a fixed $k \leq n$—the sum is over all the k-tuples of distinct positive integers $(j_1, \ldots, j_k)$. Besides of the possible difficulty in evaluating such expression, the approaches we present in this Chapter require to perform an integral before computing the sum. Hence, obtaining the EPPFs for decreasing-weights RPMs forms part of our ongoing research.

CHAPTER IV

ALLOCATION AND LABEL-SWITCHING IN MIXTURE MODELS

We have discussed in the previous chapter how random partitions are inherent objects to mixture models, either they appear implicitly or explicitly. Along this chapter, we continue the study of mixture models and their relationship with other combinatorial structures. In particular, we focus on the combinatorial structures related to the Clustering Scenarios 1.3 and 1.4 presented in Chapter 1.

We start with the random structure derived from the *allocation variables*, introduced in the previous chapter under the name of membership variables. Although these variables allow us to perform classification analyses and are also related to occupancy problems, they are widely used in the context of mixture modeling. We elaborate on all these ideas in Section 1.

As devised also in Chapter 1 (Section 1.3), allocation variables can be encoded—after a suitable modification—through another combinatorial class called *ordered set partitions*. We explain—in Section 2—the importance of this structure in mixture models, since they help us to understand why non-identifiability and *label-switching* appear. Furthermore, we provide some insights about how to make a mixture model *interpretable*, term that goes beyond identifiability, since we aim to add a meaning to each mixture component with respect to the whole population.

§1 Allocation variables

Along this chapter, the mixture densities we consider will have a finite number of components, unless otherwise is indicated. Briefly recalling the motivation of mixture models, given a set of categories, we assume an observation y is modeled by some distribution

with density $\kappa_d(y; x_d)$, for x_d some finite-dimensional parameter. The variable d indicates the category or subpopulation to where y belongs. However, in almost all cases, it is not possible to record this indicator variable d, so it is integrated out which yields the mixture density

$$f(y) = \sum_{j=1}^{m} w_j \kappa_j(y; x_j), \tag{1}$$

for some positive m and where $w_j = \mathbf{P}(d = j)$. We call variable d an *allocation variable*.

Allocation variables appear in different probabilistic and statistic problems. The most common example in Probability theory is the problem known as *occupancy problem*, where they are used to obtain the probability of having non-empty urns after throwing an arbitrary number of balls into them (see, for example, Charalambides 2005, Kolchin et al. 1978, Gnedin et al. 2007, Collet et al. 2013). Under a mixture modeling framework, we have extensive expositions by, for example, Titterington et al. (1985), Richardson & Green (1997), Marin et al. (2005) and Frühwirth-Schnatter (2006), and there are also some methodologies dealing with their posterior distribution, such as Nobile & Fearnside (2007), Papaspiliopoulos & Roberts (2008) and Mena & Walker (2015). These variables are used to improve sampling schemes and to encode the underlying random partition, as explained in the previous chapter.

Before studying some of the applications of allocation variables, let us introduce some notation and definitions. An *allocation variable* is defined as a discrete random variable d_j taking values in the set $[k] = \{1, \dots, k\}$, for some fixed k. An *allocation vector* $d = (d_1, \dots, d_n)$ is a random vector whose entries d_i, $i = 1, \dots, n$, are $[k]$-valued allocation variables.

It will be useful—in further sections—to identify the set where an allocation vector takes values. Let k be a fixed positive number and n be the dimension of the allocation vector. If we denote by $\mathcal{A}_n^{(k)}$ the set of all the possible values that an allocation vector can take, it is clear that $\#\mathcal{A}_n^{(k)} = k^n$; indeed, this set can be defined as

$$\mathcal{A}_n^{(k)} = \left\{(d_1, \dots, d_n) : d_i = 1, \dots, k, \, i = 1, \dots, n\right\}. \tag{2}$$

In Section 1.3.1, we studied a combinatorial class called *words*, denoted by $\mathcal{W}^{\mathcal{L}}$. There, it has been said that the counting sequence of such a class is k^n, $n = 0, 1, \dots$, where k is the number of available letters in the alphabet $\mathcal{L}$. Without loss of generality we can take $\mathcal{L} = [k]$. Since two combinatorial classes are isomorphic if and only if their counting sequences are the same, we have that the class of words $\mathcal{W}^{[k]}$ and the class $\{\mathcal{A}_0^{(k)}, \mathcal{A}_1^{(k)}, \dots\}$

are isomorphic. Therefore, we can use any of these classes when working with allocation vectors.

In subsequent paragraphs, a brief account of some applications of allocation vectors are explained. They are focused on occupancy and mixture models.

1.1 Allocation variables and occupancy problems

The occupancy problem is a classic problem in Probability Theory—which was studied first by De Moivre (1712)—and can be stated as follows. Given a fixed number m of urns, n balls are independently thrown into the urns; to deposit each ball, an urn is chosen uniformly. Let X_j be the random variable counting the number of balls in the jth urn. The random variables $X_1, X_2, \ldots, X_m$ are known as *occupancy numbers*, and their joint distribution is called an *occupancy model*. Hence, in occupancy problems, it is of interest modeling the number of nonempty urns and the number of urns containing exactly r balls.

Let U be a random variable defined as

$$U = \sum_{j=1}^{m} \mathbf{1}(X_j > 0),$$

and, similarly, define U_r as

$$U_r = \sum_{j=1}^{m} \mathbf{1}(X_j = r).$$

Therefore, U models the number of non empty urns, whereas U_r models the number of urns with exactly r balls. Furthermore, it is clear that they are related, since $U = \sum_{r=1}^{n} U_r$.

Similarly to Chapter I, we can obtain the support of an occupancy model. Since the balls are indistinguishable in this kind of models, we use the symbolic method for unlabeled classes. Let m be a fixed number of urns. Therefore, the class

$$\mathcal{C}^{(m)} = \mathbf{Seq}_m(\mathbf{Seq}(\mathcal{Z})),$$

encodes all the possible values of the joint vector $(X_1, \ldots, X_m)$. The generating function of this class is, then, given by

$$C^{(m)}(z) = \frac{1}{(1-z)^m} = \sum_{j=0}^{\infty} \binom{j+m-1}{m-1} z^j,$$

from where it is clear that its nth coefficient is

$$C_n^{(m)} = [z^n]C^{(m)}(z) = \binom{n+m-1}{m-1}.$$

This number is equal to the quantity given by Feller (1968) and used, for example, in Collet et al. (2013) to define exchangeable occupancy models.

Allocation vectors are related to occupancy models as follows. Suppose there are m urns available to collect n balls. Let d_i be the random variable indicating the urn in which the ith ball has been thrown, $i = 1, \dots, n$. Thus, it is immediate to see that the allocation vector $(d_1, \dots, d_n)$ takes values in $\mathcal{A}_n^{(m)}$. Furthermore, the occupancy numbers $(X_1, \dots, X_m)$ can be obtained from this random vector by setting

$$X_j = \sum_{i=1}^{n} \mathbf{1}(d_i = j), \qquad j = 1, \dots, m.$$

1.2 Usages of allocation variables in mixture modeling

Allocation variables are also important in mixture models. As already explained in the introduction—given a sample $y = (y_1, \dots, y_n)$ from a mixture density (1)—an allocation vector $d = (d_1, \dots, d_n)$ is introduced to indicate to which mixture component each observation belongs, that is $f(y_i, d_i) = \kappa_{d_i}(y_i; x_{d_i})$ for $i = 1, \dots, n$. Through this augmented model, it is possible to infer about more features of the sample y; we elaborate on this next.

In Section III.3, it has been explained how—in a mixture model—the induced random partition is encoded due to the ties occurring when sampling kernel parameters. The same procedure can be done using the allocation vector. Without loss of generality, we can assume $\kappa_j(y; x_j) = \kappa(y; x_j)$. Suppose we have the sequence of kernel parameters $(x_1, \dots, x_m)$ corresponding to a mixture density with m components. Let us assume that if an allocation variable d_j takes the value r, $1 \leq r \leq m$, it means that the observation y_j is associated to the kernel parameter x_r. Let

$$\pi_j = \{i : d_i = j,\ i = 1, \dots, n\}, \tag{3}$$

for $j = 1, \dots, m$. Therefore, the set

$$\pi' = \{\pi_j : \pi_j \neq \varnothing\},$$

FIGURE 1: Relabeling procedure for a sample of size $n = 5$ with $m = 4$ mixture components. (left) Original association given by the allocation vector $d = (2, 4, 1, 4, 4)$. (right) Relabeled association; in this case, we have the allocation vector $d' = (1, 2, 3, 2, 2)$ with component parameters $(x_1^*, x_2^*, x_3^*) = (x_2, x_4, x_1)$.

induces a partition of the dataset y, where each group—or cluster—is given by

$$\{y_d : d \in \pi_j\}, \qquad j = 1, \ldots, \#\pi';$$

in this case we have $k = \#\pi'$ groups.

However, it is important to say that—although the previous associating assumption might seem too obvious—it is often not followed in practice. The sequence of kernel parameters is commonly relabeled following a kind of order-of-appearance listing with respect to the data. To be more precise—if we denote by x_j, $j = 1, \ldots, m$, the original kernel parameters, and by x_j^* the relabeled values—the order-of-appearance relabeling is such that $x_1^* = x_{d_1}$, then x_2^* corresponds to the first value in $(x_{d_2}, \ldots, x_{d_n})$ different from x_1^*, and, following the same procedure, x_3^* is the first value in $(x_{d_3}, \ldots, x_{d_n})$ different from x_1^* and x_2^*, and so on. Note that in this relabeling procedure, it may happen that $x_2^* \neq x_{d_2}$, since, in this case, we would have had $x_{d_1} = x_{d_2}$. Figure 1 illustrates this procedure.

Note that there are several implications of using the previous relabeling procedure. On the one hand, the number of parameters x_j^*, say k^*, is at most equal to m, the number of parameters x_j, i.e. $k^* \leq m$. Consider, for example, the case $x_1 = x_2 = \cdots = x_m$, thus we only have x_1^*, so $k^* = 1$; in the example of Figure 1, we have $m = 4$ but $k^* = 3$. On the other hand, we can obtain the induced random partition by building the membership sets (III.25) or by using (3) with the relabeled allocation vector. In both cases, the resulting random partition will be the same up to a permutation of its membership sets. These observations will be used in further sections, since they are useful to understand non-identifiability and label-switching.

A second usage of allocation variables is to improve computations. Updating kernel parameters using these variables is a standard step in finite mixture models. In nonparametric methods, it was introduced by MacEachern (1994) for the Pólya urn sampler, and,

then, became also a standard step in these methods. Given a sample y from (1) and the allocation vector d, the posterior distribution of parameters $x^* = (x_1^*, \dots, x_{k^*}^*)$ is given by

$$p(x^*|y,d) \propto \prod_{j=1}^{k^*} \prod_{i \in \pi_j} \kappa(y_i; x_j^*)\, p(x_j^*),$$

where $p(x_j^*)$ is the prior distribution of x_j^*, for $j = 1, \dots, k^*$, and $\{\pi_1, \dots, \pi_{k^*}\}$ are the membership sets (3). In this case, the posterior distribution for each parameter is

$$p(x_j^*|y,d) \propto \prod_{i \in \pi_j} \kappa(y_i; x_j^*) p(x_j^*).$$

A third usage of allocation variables—and one of the central topics in this chapter—is their role distributing observations into the different mixture components, called here *the allocation problem*. Given a sample $y = (y_1, \dots, y_n)$ from the mixture density (1) and augmenting it with the allocator vector $d = (d_1, \dots, d_n)$, we can obtain the marginal posterior distribution of d given y, namely

$$p(d\,|y) \propto p(y|d)p(d), \tag{4}$$

where

$$p(y|d) = \prod_{j=1}^{m} \int \prod_{i \in \pi_j} \kappa(y_i; x_j) p(\mathrm{d}x_j),$$

with $\{\pi_1, \dots, \pi_j\}$ the membership sets, and

$$p(d) = \int p(d\,|w) p(\mathrm{d}w)$$

with $w = (w_1, \dots, w_m)$. To obtain $p(d)$, allocation variables are assumed conditionally independent given the mixing weights w, so

$$\mathbf{P}(d_i = j|w) = w_j, \qquad i = 1, \dots, n,$$

and, therefore,

$$p(d\,|w) = \prod_{j=1}^{m} w_j^{n_j},$$

where $n_j = \sum_{i=1}^{n} \mathbf{1}(d_i = j)$. For example, Nobile & Fearnside (2007) assign a Dirichlet

distribution with parameters $(\alpha_1, \dots, \alpha_m)$ to the weights, and obtain

$$p(d) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \prod_{j=1}^{m} \frac{\Gamma(\alpha_j + n_j)}{\Gamma(\alpha_j)}, \tag{5}$$

with $\alpha_0 = \sum_{j=1}^{m} \alpha_j$. Another example for the prior distribution of the allocation vector d is given by Mena & Walker (2015). In their approach, a single variable $\lambda \in (0, 1)$ is used to build the mixing weights (cf. Section III.1.2). Thus—fixing λ—we have

$$p(d) \propto \lambda^D, \tag{6}$$

with

$$D = \sum_{i=1}^{n} d_i.$$

Although both prior distributions, (5) and (6), have been defined for the same problem, their performance is quite different.

According to Nobile & Fearnside (2007), posterior samples of (4) using (5) as prior do not help to avoid non-identifiability, and, as a consequence, the problem of label-switching is still present. As pointed out by Jasra et al. (2005), non-identifiability—or in their words, non-identifiability of the components—occurs when symmetric prior distributions are placed upon the mixing parameters; and in this specific case, the prior (5) is symmetric.

Another important aspect to be considered in this models is the presence of *gaps* in the samples of d. Suppose we have a mixture with $m = 3$ components and want to allocate $n = 5$ observations. The allocation vector $d = (1, 3, 1, 1, 3)$ is an element of $\mathcal{A}_5^{(3)}$, but it has a *gap*, since the value 2 does not appear. This indicates that the second component has no observations associated to it. In order to overcome these issues, the authors need to perform some optimization post-process which involves also a relabeling procedure.

The approach of Mena & Walker (2015) considers the aforementioned drawbacks aiming to work with an identifiable mixture model. Their main concern is to obtain an *interpretable* mixture model. With *interpretability*, we refer to the early idea over which mixture models have arose: each component should represent one subpopulation—resembling the idea of identifiability—but, moreover, the first component should represent the largest subpopulation—where w_1 represents its relative proportion—the second component should represent the second largest subpopulation—with relative proportion w_2—and so on. Note that, in general, mixture models are not interpretable. Hence, even

though the prior (6) is still defined in $\mathcal{A}_n^{(m)}$, it assigns low probabilities to elements of $\mathcal{A}_n^{(m)}$ with gaps; this is cleverly done by the form of the prior. Furthermore, this prior also favors, for example, the element $d = (1, 2, 1, 1, 2)$ over $d = (1, 3, 1, 1, 3)$, or the element $d = (1, 1, 1, 1, 1)$ over $d = (2, 2, 2, 2, 2)$ and $d = (3, 3, 3, 3, 3)$, in this way, no relabeling is required, and the mixture model continues being interpretable.

§ 2 Ordered set partitions

A combinatorial structure closely related to allocation variables is the one known as *ordered set partitions*. In fields like Combinatorics, they constitute an active subject of study (Mansour 2013, Godbole et al. 2014, Ishikawa et al. 2008, Kasraoui & Zeng 2009), and they have been also used in Computing for ranking problems (Truyen et al. 2011) and in DNA sequencing (Crane 2014). However, in the best of our knowledge, there is not literature regarding the usage of ordered set partitions in mixture modeling, except for some sentences in the papers of Green & Richardson (2001), McCullagh & Yang (2008) and Miller & Harrison (2014). In the following sections, we illustrate the appearance of ordered set partitions in this context of mixture modeling.

In Section 1.3.1, it has been explained the relationship between the support of allocation vectors—the class of words, $\mathcal{W}^{[k]}$—and the class of ordered set partitions. Let denote by $\mathcal{O}^{(k)}$ the class of ordered set partitions with exactly k blocks, studied in Section 1.3. Both classes are labeled structures. Consider now the subclass of $\mathcal{W}^{[k]}$ formed only by words where each letter appears at least once. Then, its specification is

$$\mathcal{W}_{\geq 1}^{[k]} = \mathbf{Seq}_k(\mathbf{Set}_{\geq 1}(\mathcal{Z})),$$

from where its EGF is given by

$$W_{\geq 1}^{(k)}(z) = (e^z - 1)^k.$$

Note that this function is actually the same as the EFG of $\mathcal{O}^{(k)}$, given in Equation (1.14). As a consequence, both classes, $\mathcal{O}^{(k)}$ and $\mathcal{W}_{\geq 1}^{[k]}$, are isomorphic, which implies that we can model the same problem using any of them.

2.1 Clustering structure in mixture models and label-switching

In order to have a clear picture of the role of ordered set partitions in mixture models, we start analyzing the model given in Section III.3 which is used to infer about the clustering

structure through a random partition. In this section, we assume the mixing distribution is some discrete RPM—hence, the number of components is potentially infinite—though the results can be also applied for finite mixture densities.

Based on a mixture model—in Equation (III.26)—we define the following hierarchical model

$$\begin{aligned} y_i|x_i, \Pi_n &\overset{\text{ind.}}{\sim} \kappa(y_i; x_i) \qquad i = 1, \dots, n \\ x_j^*|\Pi_n = \pi_{k^*} &\overset{\text{i.i.d.}}{\sim} \nu_0(x_j^*; \phi) \qquad j = 1, \dots, k^* \\ \Pi_n &\sim \boldsymbol{P}, \end{aligned} \tag{7}$$

where Π_n is a $\mathcal{P}_{[n]}$-valued random partition with prior $\boldsymbol{P}$, and $(x_1^*, \dots, x_{k^*}^*)$ are the k^* different values of $(x_1, \dots, x_n)$. We have already explained in the relationship between allocation vectors and random partitions in Section 1.2. When the prior $\boldsymbol{P}$ belongs to some class of EPPFs, its probability function, p, is defined as $p : \Delta_n \to [0, 1]$, where $\Delta_n = \bigcup_{k=1}^n \Delta_{n,k}$ is the set of integer compositions of n (Equation 1.6). So, for example, we have

$$\begin{aligned} \mathbf{P}(\Pi_3 = \{\{1, 2\}, \{3\}\}) = \mathbf{P}(\Pi_3 = \{\{1, 3\}, \{2\}\}) &= p(2, 1) \\ = \mathbf{P}(\Pi_3 = \{\{3\}, \{1, 2\}\}) &= p(1, 2) \end{aligned}$$

due to the symmetry of p with respect to Δ_n. A similar symmetry is obtained *a posteriori* but at a different level. Specifically, the posterior distribution of Π_n—from (7)—is given by

$$p(\Pi_n|y) \propto p(\Pi_n)p(y|\Pi_n), \tag{8}$$

where the likelihood is written as

$$p(y|\Pi_n = \pi) = \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \pi_j} \kappa(y_i; x) g_0(\mathrm{d}x),$$

with $\pi = \{\pi_1, \dots, \pi_k\} \in \mathcal{P}_{[n]}$ and g_0 the density function of ν_0. This posterior distribution is a symmetric function with respect to $\mathcal{P}_{[n]}$, that is $\Pi_n \mapsto p(\Pi_n|y) : \mathcal{P}_{[n]} \to [0, 1]$. Notice that, in this case, we can obtain some point estimate $\tilde{\Pi}_n$, like the mode, but it does not give us any precise allocation of the observations into the clusters. For example, if $n = 4$, suppose we obtain $\tilde{\Pi}_4 = \{\{1\}, \{2\}, \{3, 4\}\}$. Let $\tilde{\Pi}_4' = \{\{1\}, \{3, 4\}, \{2\}\}$, namely the last two

blocks of $\tilde{\Pi}_4$ have been permuted. Then—due to the symmetry—we have

$$p(\tilde{\Pi}_4|y) = p(\tilde{\Pi}'_4|y),$$

and this same probability is assigned to the 3! permutations of the blocks $\{1\}, \{2\}$ and $\{3, 4\}$. Therefore, we cannot assure, for example, that observation y_1 belongs to the first cluster, or to the second or to the third, we only know that y_1 forms one group by itself.

Such a symmetry of the posterior distribution $p(\Pi_n|y)$ is acceptable when we just care about of knowing which observations are grouped together; this is the essential clustering problem. However, if we—additionally—want to infer about cluster governing parameters x^*, it is necessary to follow one of two methodologies: (i) compute $p(x^*|\Pi_n = \pi, y)$, or (ii) work with the joint posterior distribution $p(x^*, \Pi_n|y)$, but it is not possible to use the marginal posterior distribution $p(x^*|y)$; under the model (7), it makes no sense. To have a clear picture of this claim, notice that the joint posterior distribution of (x^*, Π_n) is given by

$$p(x^*, \Pi_n|y) \propto p(\Pi_n)p(x^*|\Pi_n)p(y|x^*, \Pi_n), \tag{9}$$

where, for $\pi = \{\pi_1, \ldots, \pi_k\} \in \mathcal{P}_{[n]}$, the prior for x^* is given by

$$p(x^*|\Pi_n = \pi) = \prod_{j=1}^{k} p(x_j^*),$$

with $k = \#\pi$ and the likelihood function is

$$p(y|x^*, \Pi_n = \pi) = \prod_{j=1}^{k} \prod_{i \in \pi_j} \kappa(y_i; x_j^*).$$

Then, the marginal distribution of x^* is

$$p(x^*|y) \propto \sum_{k=1}^{n} \sum_{\pi \in \mathcal{P}_{[n]}^k} p(\Pi_n = \pi) \prod_{j=1}^{k} g_0(x_j^*) \prod_{i \in \pi_j} \kappa(y_i; x_j^*),$$

where $\mathcal{P}_{[n]}^k$ is the set of all the set partitions with exactly k groups. It is evident that this distribution has a complicated form because the outer sum makes the parameter vary its dimension. Even if this distribution is constrained to a fixed number of blocks—except for the trivial cases $k = 1, n$—it still has an odd form. Note that there are $\#\mathcal{P}_{[n]}^k = S(n, k)$ distinct possible groupings for the data for a given k, and—due to the order-of-

appearance listing assumption we discussed before—we cannot think the sum as a mean over the groupings, since, for example, observation y_1 will never appear in any block apart from the first one.

All this discussion serves us as preamble to understand the origins of label-switching. In most of the sampling schemes for mixture models, there is a step to update component parameters and another to update the (latent) random partition, in terms of either the allocation vector or the random partition itself. Hence, the label-switching problem lies in the fact that component parameters x^* and the random partition Π_n are considered as independent each other. This means that—even for a fixed partition π^*—samplers have to consider all the $k!$ permutations of the blocks, where $k = \#\pi^*$, to associate the k component parameters $(x_1^*, \dots, x_k^*)$ (cf. Rodríguez & Walker 2014), forgetting the fact that such parameters—in that order—where used to obtain the partition. In other words, the parameter x_j^* is linked to only one block: π_j; such a parameter can be considered as the *label* of the block π_j.

Mathematically—by linking parameters with blocks—we mean that the joint posterior distribution (9) should be expressed as

$$p(x^*, \Pi_n = \{\pi_1, \dots, \pi_k\}|y) = p((x_1^*, \pi_1), \dots, (x_k^*, \pi_k)|y). \tag{10}$$

Using this form, we are able to infer about the jth cluster parameter, since the expression

$$p(x_j^*|\pi_j, y) \propto \prod_{i \in \pi_j} \kappa(y_i; x_j^*) p(x_j^*)$$

is not affected anymore by permutations of the blocks; indeed, there is nothing to permute. Note also that the joint distribution $p(x, \Pi_n|y)$ becomes a function with a particular support. If we fix the number of blocks in Π_n to k, we have that $(x^*, \Pi_n) \mapsto p(x^*, \Pi_n|y) : \mathbb{X}^k \times \mathcal{O}_{[n]}^{(k)} \to [0, 1]$, where $\mathcal{O}_{[n]}^{(k)}$ is the set of all ordered set partitions of $[n]$ with exactly k blocks. As an illustration, consider the partitions $\tilde{\Pi}_4 = \{\{1\}, \{2\}, \{3, 4\}\}$ and $\tilde{\Pi}_4' = \{\{1\}, \{3, 4\}, \{2\}\}$, as before, and the parameters $x^* = (x_1^*, x_2^*, x_3^*)$. Therefore, under (10), we have

$$p(x^*, \tilde{\Pi}_4|y) \neq p(x^*, \tilde{\Pi}_4'|y),$$

since—unlike in the previous case—the likelihood is not the same. For the left-hand side, we have that

$$p(y|x^*, \tilde{\Pi}_4) = \kappa(y_1; x_1^*)\, \kappa(y_2; x_2^*)\, \kappa(y_3; x_3^*) \kappa(y_4; x_3^*),$$

which is different to the right-hand side, namely

$$p(y|x^*, \tilde{\Pi}_4') = \kappa(y_1; x_1^*)\,\kappa(y_2; x_3^*)\,\kappa(y_3; x_2^*)\kappa(y_4; x_2^*).$$

2.2 Towards interpretable mixture models

We have seen that the label-switching problem appears because standard mixture-model methodologies usually are defined to work in the space of set partitions, $\mathcal{P}_{[n]}$, instead of the complete space—given by the ordered set partitions, $\mathcal{O}_{[n]}$—and as a consequence, some information is lost. In this case, the information lost is the link between clusters and their respective governing parameter. But remember that label-switching and the lack of identifiability are related issues. Hence, this ordered-set-partition approach could also be applied in mixture models to make them identifiable, or even more, to make them interpretable.

Regarding identifiability, San Martín & Quintana (2002) explain that—given a sampling probability P^θ indexed by a parameter θ—a parametrization θ is said to be *identified* if the mapping $\theta \rightarrowtail P^\theta$ is injective. Applying this definition to our context of mixture modeling, we have that P is a mixture density and the parameter θ consists on component parameters x_j^*, $j = 1, \ldots, k$, and the partition $\Pi_n = \{\pi_1, \ldots, \pi_k\}$. Thus, when a mixture-model methodology is defined to work in the space of set partitions, $\mathcal{P}_{[n]}$, parameter θ takes the form

$$\theta = \left(x_1^*, \ldots, x_k^*, \pi_1, \ldots, \pi_k\right),$$

and the mapping obtained is not injective. This occurs due to the fact that we can permute every component parameter x_j^* with every partition block π_j, i.e. the label-switching problem. Therefore, the mixture model is not identifiable. On the other hand, if θ takes the form

$$\theta = \left((x_1^*, \pi_1), \ldots, (x_k^*, \pi_k)\right),$$

we are working with the space of ordered set partitions, $\mathcal{O}_{[n]}$ (cf. Equation 10), and the mapping $\theta \rightarrowtail P^\theta$ is injective; in other words, we obtain an identifiable mixture model.

Concerning interpretability, recall the approach of Mena & Walker (2015) discussed above. Such an approach aims to obtain identifiable mixture models by avoiding label-switching, but, also tries to obtain interpretable mixture models by favoring certain elements in the set of allocations $\mathcal{A}_n^{(m)}$. Nevertheless, it also has some concerns to care about: (i) the number of mixture components is fixed and finite, and (ii) there is a positive probability (although small) of having empty components.

Notice that ordered set partitions also solves this second point. Under the approach of words, $\mathcal{W}^{[k]}$, as k grows, the number of empty groups also increases; actually, given any sample of size n, if $k > n$ this always happens. But with ordered set partitions, $\mathcal{O}_{[n]}^{(k)}$, this number is bounded by n (cf. Tables 1.2 and 1.3). This—besides of making allocation problems more tractable—provides a more coherent model, because it is more logical to have at most $k = n$ groups to allocate n observations than having an arbitrary number of them. This philosophy is shared with the theory of random partitions; indeed, one of their attractiveness in clustering is their ability to infer about the number of groups. Furthermore, we can take advantage of the close relationship between set partitions and ordered set partitions to overcome the fist limitation. If we could find a way to extend the existent theory of random partitions to $\mathcal{O}_{[n]}$-valued random objects—besides of providing a novel methodology to deal with clustering-related problems—we would obtain an immediate approach to model also the number of groups.

Providing an extension to the $\mathcal{P}_{[n]}$-valued theory, as explained, we would obtain identifiable mixture models. However, to go further, that is, to have interpretable mixture models, we have to find a way to translate the prior (6) for d to our new context of ordered set partitions or to provide an alternative proposal.

Recalling, Mena & Walker (2015) define $p(d) \propto \lambda^D$, for some fixed $0 < \lambda < 1$, where

$$D = \sum_{i=1}^{n} d_i,$$

which can be written as

$$D = \underbrace{1 + 1 + \cdots + 1}_{n_1 \text{ times}} + \underbrace{2 + 2 + \cdots + 2}_{n_2 \text{ times}} + \cdots + \underbrace{k + k + \cdots + k}_{n_k \text{ times}} = \sum_{j=1}^{k} j n_j,$$

with n_j, $j = 1, \ldots, k$, the size of the jth block. This distribution is exchangeable with respect to $\mathcal{A}_n^{(m)}$, but its equivalent over $\mathcal{O}_{[n]}^{(m)}$ is not, even though the sum D is the same in both cases. In the first case, exchangeability means that

$$\mathbf{P}(\delta = (d_1, \ldots, d_n)) = \mathbf{P}(\delta = (d_{\rho(1)}, \ldots, d_{\rho(n)})),$$

for all permutation ρ of $[n] = \{1, \ldots, n\}$, and it is satisfied by $p(d) \propto \lambda^D$. On the other hand, an exchangeable distribution over ordered set partitions should satisfy, based on

Definition 1.3,

$$\mathbf{P}(\vartheta_n = \pi) = \mathbf{P}(\vartheta_n = \sigma(\pi)),$$

for any $\pi = \{\pi_1, \ldots, \pi_k\} \in \mathcal{O}_{[n]}$ and all permutation σ in the symmetric group acting on $[n]$, or equivalently,

$$\mathbf{P}(\vartheta_n = \pi) = q(n_1, \ldots, n_k),$$

where $q : \Delta_{n,k} \to [0,1]$ is a symmetric function, with $n_j = \#\pi_j$ for $j = 1, \ldots, k$. However, the function

$$q(n_1, \ldots, n_k) \propto \lambda^{\sum_{j=1}^{k} j n_j},$$

for a fixed $\lambda \in (0,1)$, is not exchangeable. As an illustration, take $n = 3$ and $m = 2$. Then

$$\mathcal{A}_3^{(2)} = \big\{(1,2,2),\ (1,1,2),\ (1,2,1),\ (2,1,1),\ (2,2,1),\ (2,1,2)\big\},$$

and

$$\begin{aligned}\mathcal{O}_{[3]}^{(2)} = \big\{ &\{\{1\},\{2,3\}\},\ \{\{2\},\{1,3\}\},\ \{\{3\},\{1,2\}\},\\ &\{\{1,2\},\{3\}\},\ \{\{1,3\},\{2\}\},\ \{\{2,3\},\{1\}\}\big\},\end{aligned}$$

with $\Delta_{3,2} = \{(1,2),\ (2,1)\}$. Therefore,

$$\begin{aligned}\mathbf{P}(\delta = (1,1,2)) = \mathbf{P}(\delta = (1,2,1)) = \mathbf{P}(\delta = (2,1,1)) = p(1,1,2) \propto \lambda^4,\\ \mathbf{P}(\delta = (1,2,2)) = \mathbf{P}(\delta = (2,2,1)) = \mathbf{P}(\delta = (2,1,2)) = p(1,2,2) \propto \lambda^5,\end{aligned}$$

as expected, but

$$\begin{aligned}\mathbf{P}(\vartheta_3 = \{\{1\},\{2,3\}\}) = \mathbf{P}(\vartheta_3 = \{\{2\},\{1,3\}\}) = \mathbf{P}(\vartheta_3 = \{\{3\},\{1,2\}\}) = q(1,2) \propto \lambda^4,\\ \mathbf{P}(\vartheta_3 = \{\{1,2\},\{3\}\}) = \mathbf{P}(\vartheta_3 = \{\{1,3\},\{2\}\}) = \mathbf{P}(\vartheta_3 = \{\{2,3\},\{1\}\}) = q(2,1) \propto \lambda^5,\end{aligned}$$

so exchangeability is lost since it is required that $q(1,2) = q(2,1)$. Hence, we require a different approach.

A second option to achieve interpretable mixture models is working directly on the subset of $\mathcal{A}_n^{(m)}$ where this feature is preserved. In such a set, it is required that

$$n_1 \geq n_2 \geq \cdots \geq n_k, \tag{11}$$

where n_j is the size of the jth group. Although it is not straightforward to give a symbolic

$n \setminus k$	1	2	3	4	5	6	7	8	9	10	$\#\Delta_n^*$
1	1										1
2	1	1									2
3	1	1	1								3
4	1	2	1	1							5
5	1	2	2	1	1						7
6	1	3	3	2	1	1					11
7	1	3	4	3	2	1	1				15
8	1	4	5	5	3	2	1	1			22
9	1	4	7	6	5	3	2	1	1		30
10	1	5	8	9	7	5	3	2	1	1	42

TABLE 1: Triangular array showing the counting sequences of the subset of Δ_n, denoted by Δ_n^*, in which the greatest part is $n_1 = r$, for different values of r and n such that $1 \leq r \leq n$ and $n = 1, \ldots, 10$.

specification of such a combinatorial class, we can analyze the support of the prior distribution for $\mathcal{O}_{[n]}$-valued random objects. Note that such a distribution should be restricted to the subset of $\Delta_{n,k}$ where (11) is satisfied for every $k = 1, \ldots, n$; equivalently, we can fix the size of the biggest group, n_1, and obtain the elements in $\Delta_n = \cup_{k=1}^{n} \Delta_{n,k}$ satisfying (11). Let Δ_n^* be such a subset. As an illustration, consider the case $n = 5$, so we have

$$\Delta_5^* = \{(1,1,1,1,1),\ \ (2,2,1),\ (2,1,1,1),\ \ (3,2),\ (3,1,1),\ \ (4,1),\ \ (5)\}\,.$$

The counting sequence for the underlying combinatorial class has been found in the Encyclopedia of Integer Sequences with the entry A008284. In Table 1, some values of this sequence are shown. Given this class, we can find the subset of $\mathcal{O}_{[n]}$ whose blocks sizes belong to it. Furthermore, all the mixture densities with groups in bijection with such a subset of $\mathcal{O}_{[n]}$ will be interpretable.

CONCLUDING REMARKS

Throughout this thesis, we have examined the importance of clustering and related approaches under a statistical perspective. There is a great number of methodologies dealing with these problems, in part, due to their versatility and applicability in many and different fields. In this respect, we have explained that random partitions are suitable candidates to deal with clustering problems since their support is isomorphic to the space where clustering is defined. Among the different available methodologies to work with random partitions, we chose the Bayesian nonparametric approach—and constrained the present thesis to follow such an approach—because—besides of justifying such an approach—de Finetti's representation theorem provides a direct form to model random partitions.

As it is known, a crucial part to perform any Bayesian analysis involves the correct choice of the prior distribution, which can be achieved through the study of all the possible values taken by the random object of interest. In the case of the clustering models, such support is given by particular structures better understood under a combinatorial perspective. Thus, we saw how set partitions, $\mathcal{P}_{[n]}$, serve as support for random partitions, and—derived from this combinatorial class—we obtained two additional classes emerging when the clustering rules are changed: segmentations, $\mathcal{S}_{[n]}$, and ordered set partitions, $\mathcal{O}_{[n]}$. Moreover, we have that $\mathcal{S}_{[n]} \subseteq \mathcal{P}_{[n]} \subseteq \mathcal{O}_{[n]}$, fact that has made us think that random partitions might not be enough or appropriate for particular clustering situations.

Indeed, the class of segmentations encodes clustering situations where observations are indexed by covariates taking values in some ordered set; this eliminates groupings with gaps. Chapter II was devoted to the study of these models—called segmentation models—focusing on an application in the field of change-point detection. Based on the theory of random partitions, we presented different approaches to build $\mathcal{S}_{[n]}$-valued prior distributions, and study their role in this kind of clustering. Furthermore, a sampling scheme to draw values of the corresponding $\mathcal{S}_{[n]}$-valued posterior distribution was provided. All this work served to present a novel proposal for change-point detection,

which happened to be an attractive alternative to the existing list of methodologies in this topic.

Moving on to set partitions, it can be seen that there is a great interest on this topic. Random partitions have been studied under different probabilistic approaches. Most of Chapter I was a survey on this topic, focused just on the de Finetti's theorem viewpoint. Following the study of set-partition-based cluster analysis, a complementary study was given in Chapter III. There, it has been pointed out the relationship between random partitions and mixture models and why these models can be thought as an extension to random partition models for continuous observations. We concentrated on the study of specific mixture densities having decreasing weights. On the one hand, these densities intend to deal with some of the challenges of mixture modeling, like identifiability, and—from the results obtained in this chapter—the generalization made of the geometric process mixture model promises to be a good starting point. On the other hand, the underlying random probability measure in these densities could result on appealing alternatives as random partition distributions.

The last clustering scenario corresponds to the one encoded by ordered set partitions. In Chapter IV, we continued with the study of mixture models focusing on other of their challenges which appears during simulations: the label-switching problem. In order to do that, we started analyzing a related problem consisting on the allocation of observations among a given number of mixture components. Under a suitable modification, the resulting space for this problem is encoded by ordered set partitions, but what was more interesting was the fact that if we study the equivalent clustering model based on mixture densities and consider a bigger space than set partitions, i.e. ordered set partitions, we can see that there are no more label-switching problems. We concluded from this observation that label-switching appears because sampling schemes often work on a smaller space. At this point, we have given the foundations of a different and new methodology to deal with the label-switching problem. Furthermore—given the relationship between label-switching and non identifiability—we could also deal with non identifiability issues by considering the complete space instead of the reduced one of set partitions. Another important observation obtained from this study was the convenience of thinking on a special class of mixture models where mixture components represent subpopulations in a clear way. This would recover one of the original purposes of mixture models: interpretability.

However, all these studies are not finished; there is more work to do in each of these lines of research. Turning back to segmentation models and under the context of our

proposed model, one of the main interests is the prediction. Actually, we have already devised a possible procedure to perform it. As explained in Section II.4, we may be interested in one of three classes of prediction: (i) given a new observation, we want to know whether it forms part of the current regime or it stars a new one, where in such a case a change point appears, (ii) we might require an estimate of the length of the current regime, that is, we want to known when the next change point will occur; or (iii) under certain contexts, we could need to know if a change point occurred inside an unobserved interval of time, namely between two consecutive observations.

On a different direction, segmentation models for multivariate data also represent an attractive extension to our proposal. One possibility is simply using a multivariate likelihood, but it could be too restrictive in the sense that the model would only detect changes occurring in all the variables at the same point location. A more flexible model should instead allow the detection of two types of changes: (i) those proper for each variable, and (ii) those common for all of them.

Regarding to mixture models with decreasing weights, they might represent an appealing option to deal with the problem of identifiability. We have presented two approaches to extend the geometric process. We have analyzed one of them, whereas the second one is part of our current work. On the other hand, the underlying random probability measure in these models—used as mixing distribution—could be useful in contexts like species sampling modeling. In this context, it is desirable to provide the probability distribution for the induced random partition, which leaves a broad line of research to explore.

Finally, the idea presented regarding to interpretable mixture models constitutes an important advance in the context of mixture modeling. As explained before, interpretability represents the original idea for a mixture model, however it seldom is achieved in practice. Thus, it is necessary to find an implementation for this model. Merging the studies of ordered set partitions with decreasing-weights mixture models correctly would result in a novelty in this context.

APPENDIX A

SAMPLING SCHEME FOR THE PARAMETERS OF THE TWO-PARAMETER POISSON-DIRICHLET PROCESS

When working with the two-parameter Poisson-Dirichlet process, it might be of interest assigning prior distributions to its parameters. Under the case $\theta > -\sigma$, this section provides a simulation scheme to sample from the corresponding posterior distributions.

Assume a beta prior with parameters (α, β), is assigned to σ and a shifted gamma prior with parameters $(\gamma, \delta, -\sigma)$, to θ. A shifted gamma random variable X is such that its probability density function is given by

$$f(x; a, b, c) = \frac{(x-c)^{a-1}}{\Gamma(a)b^a} e^{-b(x-c)} \mathbf{1}(x > c).$$

Furthermore, assume that likelihood function of (σ, θ) is taken from the EPPF of PD(σ, θ) given by

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1\uparrow}. \tag{1}$$

Simulating from the posterior of σ can be easily done via the ARMS method (Gilks et al. 1995) since this is given by

$$p(\sigma | \dots) \propto \sigma^{\alpha-1}(1-\sigma)^{\beta-1}(\theta+\sigma)^{\gamma-1} e^{-\delta\sigma} \prod_{i=1}^{k-1}(\theta + i\sigma) \prod_{j=1}^{k} (1-\sigma)_{n_j-1\uparrow}$$

with $\sigma \in (\max\{-\theta, 0\}, 1)$. For the posterior distribution of θ, the following result can be used.

Proposition 1. The augmented full conditional distribution of θ—with likelihood function (1) and prior distribution given by a shifted gamma distribution with parameters $(\gamma, \delta, -\sigma)$—is given by

$$p(\theta|y, z, \dots) \propto \sum_{j=0}^{k+1} w_j \mathrm{Ga}(\theta; \gamma + j, \delta + y - \log(z), -\sigma),$$

where the weights $(w_j)_{j=0}^{k+1}$ are given by

$$w_j \propto \frac{\big((n-\sigma)(n+1-\sigma)c_{k-1,j} + (2n+1-2\sigma)\sigma c_{k-1,j-1} + \sigma^2 c_{k-1,j-2}\big)\Gamma(\gamma+j)}{\big(\sigma(\delta + y - \log(z))\big)^j},$$

with k and n as in (1), $z \sim \mathrm{Be}(\theta+2, n)$, $y \sim \mathrm{Exp}(\theta+1)$ and $c_{k,j}$, $j = 1 \dots, k$, the absolute value of the Stirling number of the first kind, where $c_{k,r} = 0$ for $r \leq 0$ and $r > k$ and $c_{0,0} = 1$.

Demostración

The posterior distribution of θ is given by

$$p(\theta| \dots) \propto \frac{(\theta+\sigma)^{c-1} e^{-d\theta}}{(\theta+1)_{n-1\uparrow}} \prod_{i=1}^{k-1} (\theta + i\sigma) \mathbf{1}(\theta > -\sigma). \tag{2}$$

To derive the augmented full conditional distribution, notice that

$$\prod_{i=1}^{k-1} (\theta + i\sigma) = \sigma^{k-1} \sum_{j=0}^{k-1} c_{k-1,j} \Big(\frac{\theta+\sigma}{\sigma}\Big)^j,$$

where $c_{k-1,j}$ is the absolute value of a Stirling number of the first kind, and that

$$\frac{1}{(\theta+1)_{n-1\uparrow}} = \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} = \frac{(\theta+n)(\theta+n+1)}{(\theta+1)\Gamma(n)} \boldsymbol{B}(\theta+2, n),$$

where $\boldsymbol{B}$ is the beta function. Therefore, the posterior distribution (2) can be rewritten as

$$\sum_{j=0}^{k-1} \frac{c_{k-1,j}}{\sigma^j} \frac{(\theta+n)(\theta+n+1)}{(\theta+1)} \boldsymbol{B}(\theta+2, n)(\theta+\sigma)^{c+j-1} e^{-d\theta} \mathbf{1}(\theta > -\sigma).$$

Using data augmentation, let $y \sim \mathrm{Exp}(\theta+1)$ and—following the approach of West (1992)

to update the Dirichlet process's parameter—let $z \sim \text{Be}(\theta + 2, n)$. Thus,

$$p(\theta|y, z, \dots) \propto \sum_{j=0}^{k-1} \frac{c_{k-1,j}}{\sigma^j}(\theta + n)(\theta + n + 1)(\theta + \sigma)^{c+j-1} e^{-(d+y-\log(z))\theta} \mathbf{1}(\theta > -\sigma).$$

Expanding the product $(\theta + n)(\theta + n + 1)$ as function of $(\theta + \sigma)$ and rearranging the terms of the sum with respect to $(\theta + \sigma)^{c+j-1}$, $j = 0, 1, \dots, k + 1$, we have

$$p(\theta|y, z, \dots) \propto \sum_{j=0}^{k+1} w'_j(\theta + \sigma)^{c+j-1} e^{-(d+y-\log(z))\theta} \mathbf{1}(\theta > -\sigma),$$

where

$$w'_j = \frac{(n - \sigma)(n + 1 - \sigma)c_{k-1,j} + (2n + 1 - 2\sigma)\sigma c_{k-1,j-1} + \sigma^2 c_{k-1,j-2}}{\sigma^j},$$

with $c_{k,r} = 0$ for $r \leq 0$ and $r > k$. Finally, completing the density in each term leads us to the stated result. □

In the best of my knowledge, other available implementations to simulate from the posterior distribution of θ were limited to Metropolis-Hastings steps (see, for example, Jara et al. 2010 and Nieto-Barajas & Contreras-Cristan 2014). Hence, the above Proposition could also be useful in other Bayesian nonparametric implementations based on the two-parameter Poisson-Dirichlet process.

APPENDIX B

SIMULATION OF TRUNCATED DISTRIBUTIONS

In this section, simulation schemes for truncated distributions are provided. Specifically, two cases are considered: Poisson and negative binomial distributions, both left-truncated.

§1 Truncated Poisson distribution

Suppose a random number x from a truncated Poisson distribution is required. A random variable $X \sim \text{Poi}(\lambda)$, with $\lambda > 0$, has probability mass function given by

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \mathbf{1}(x \geq 0).$$

If truncation is from the right, namely, there is a $\tau \in \mathbb{N}$ such that $\mathbf{P}(X > \tau) = 0$, simulation is straightforward since the resulting density has a finite support $\{1, 2, \dots, \tau\}$. On the other hand, truncation from the left presents some issues, but these can be overcome via data augmentation. Therefore, simulation is done using a Gibbs sampler.

Suppose X is a random variable Poisson distributed with parameter λ and left-truncated at $\tau \in \mathbb{N}$, that is, $\mathbf{P}(X < \tau) = 0$. Using a Gibbs sampler, it is required to sampling from a density proportional to $\lambda^x / x! \, \mathbf{1}(x \geq \tau)$. Notice first that

$$\begin{aligned} \frac{1}{x!} =& \frac{1}{(x-\tau)!} \left\{ \prod_{i=0}^{\tau-1} (x - d + 1 + i) \right\}^{-1} = \frac{1}{(x-\tau)!} \frac{\Gamma(x-\tau+1)}{\Gamma(x+1)} \frac{\Gamma(\tau)}{\Gamma(\tau)} \\ =& \frac{\beta(x-\tau+1, \tau)}{\Gamma(\tau)(x-\tau)!}, \end{aligned} \tag{1}$$

where the expression between brackets is the Pochhammer symbol, $(x - \tau + 1)_{\tau\uparrow}$, and

$\beta(a, b)$ is the beta function, which additionally has the integral representation

$$\beta(a, b) = \int_0^1 z^{a-1}(1-z)^{b-1}\mathrm{d}z.$$

Using data augmentation with $z \sim \mathrm{Be}(x - \tau + 1, \tau)$, the sampler has to simulate from the bivariate density

$$p(x, z) \propto \frac{\lambda^x}{(x-\tau)!} z^{x-\tau}(1-z)^{\tau-1}\mathbf{1}(x \geq \tau).$$

From this, the full conditional distribution for x is given by

$$p(x|z) \propto \frac{(\lambda z)^{x-\tau}}{(x-\tau)!}\mathbf{1}(x \geq \tau),$$

which corresponds to a Poisson distribution with parameter λz shifted at τ. Therefore, a Gibbs sampler draws values at iteration t as follows:

$$\begin{aligned} z^{(t)} &\sim \mathrm{Be}(x^{(t-1)} - \tau + 1, \tau), \\ x^{(t)} &\sim \mathrm{Poi}(\lambda z^{(t)}) + \tau, \end{aligned}$$

for an initial value $x^{(0)}$ and given λ and τ.

§ 2 Truncated negative binomial distribution

In a similar way to the truncated Poisson distribution, simulating from a right-truncated distribution is straightforward, but not when it is left-truncated. The negative binomial probability mass function is given by

$$f(x) = \binom{x + r - 1}{x} p^r(1-p)^x\mathbf{1}(x \geq 0),$$

for $r > 0$ and $0 < p < 1$.

Suppose that it is required to sample from a negative binomial distribution left-truncated at $\tau \in \mathbb{N}$. This can be also done using a Gibbs sampler. In order to do it, notice that using (1),

$$p(x) \propto \frac{\beta(x - \tau + 1, \tau)\Gamma(x + r)}{(x-\tau)!}(1-p)^x\mathbf{1}(x \geq \tau).$$

If the beta an gamma functions are substituted by their integral expressions, the resulting density is

$$p(x, z, v) \propto \frac{(1-p)^x}{(x-\tau)!} z^{x-\tau}(1-z)^{\tau-1} v^{x+r-1} e^{-v}\mathbf{1}(x \geq \tau),$$

and the full conditional distribution of x is

$$p(x|z,v) \propto \frac{((1-p)vz)^{x-\tau}}{(x-\tau)!} e^{-(1-p)vz} \mathbf{1}(x \geq \tau),$$

which is a Poisson distribution with parameter $(1-p)vz$ and shifted at τ. Therefore, using data augmentation, with $z \sim \text{Be}(x-\tau+1,\tau)$ and $v \sim \text{Ga}(x+r,1)$, a Gibbs sampler draws values at iteration t as follows:

$$\begin{aligned} z^{(t)} &\sim \text{Be}(x^{(t-1)} - \tau + 1, \tau), \\ v^{(t)} &\sim \text{Ga}(x^{(t-1)} + r, 1), \\ x^{(t)} &\sim \text{Poi}((1-p)z^{(t)}v^{(t)}) + \tau, \end{aligned}$$

for an initial value $x^{(0)}$, and where r, p and τ are given values.

Similarly to the simulation of parameter θ explained in the previous section, we have only found one more procedure to sampling from these left-truncated distributions by Geyer (2007), which is based on rejection sampling.

BIBLIOGRAPHY

Aldous, D. J. (1985), Exchangeability and related topics, *in* 'École d'Été de Probabilités de Saint-Flour XIII', Lecture notes in mathematics, Springer-Verlag.

Allen, D. E., McAleer, M., Powell, R. J. & Kumar-Singh, A. (2013), 'Nonparametric Multiple Change Point Analysis of the Global Financial Crisis'.
URL: *http://dx.doi.org/10.2139/ssrn.2270029*

Antoniak, C. E. (1974), 'Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems', *The Annals of Statistics* **2**(6), 1152–1174.

Aoki, M. (2003), Applications of Exchangeable Random Partitions to Economic Modeling, UCLA Economics Online Papers 228, UCLA Department of Economics.

Barry, D. & Hartigan, J. A. (1992), 'Product partition models for change point problems', *The Annals of Statistics* **20**(1), 260–279.

Barry, D. & Hartigan, J. A. (1993), 'A Bayesian Analysis for Change Point Problems', *Journal of the American Statistical Association* **88**(421), 309–319.

Berestycki, N. (2009), 'Recent progress in coalescent theory', *Ensaios Matemáticos* **16**, 1–193.

Bissiri, P. G. & Ongaro, A. (2014), 'On the topological support of species sampling priors', *Electronic Journal of Statistics* **8**(1), 861–882.

Braun, J. V., Braun, R. K. & Müller, H. G. (2000), 'Multiple Changepoint Fitting via Quasilikelihood, with Application to DNA Sequence Segmentation', *Biometrika* **87**(2), 301–314.

Bubula, A. & Ötker-Robe, I. (2002), The Evolution of Exchange Rate Regimes Since 1990: Evidence from De Facto Policies, Working Paper 02/155, International Monetary Fund.

Bush, C. A. & MacEachern, S. N. (1996), 'A semiparametric Bayesian model for randomised block designs', *Biometrika* **83**(2), 275–285.

Caron, F., Doucet, A. & Gottardo, R. (2012), 'On-line changepoint detection and parameter estimation with application to genomic data', *Statistics and Computing* **22**, 579–595.

Cayley, A. (1859), 'On the analytical forms called trees, second part', *Philosophical Magazine, Series IV* **18**(121), 374–378.

Charalambides, C. A. (2005), *Combinatorial Methods in Discrete Distributions*, Wiley.

Chen, J. & Gupta, A. K. (1997), 'Testing and locating variance changepoints with application to stock prices', *Journal of the American Statistical Association* **92**(438), 739–747.

Chernoff, H. & Zacks, S. (1964), 'Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time', *Annals of Mathematical Statistics* **35**(3), 999–1018.

Collet, F., Leisen, F., Spizzichino, F. & Suter, F. (2013), 'Exchangeable Occupancy Models and Discrete Processes with the Generalized Uniform Order Statistics Property', *Probability in the Engineering and Informational Sciences* **27**, 533–552.

Comtet, L. (1974), *Advanced Combinatorics*, Reidel Publishing Company.

Consul, P. C. & Famoye, F. (2006), *Lagrangian Probability Distributions*, Birkhäuser.

Cox, J. C., Ingersoll, Jonathan E., J. & Ross, S. A. (1985), 'A Theory of the Term Structure of Interest Rates', *Econometrica* **53**(2), 385–407.

Crane, H. (2014), 'The cut-and-paste process', *The Annals of Probability* **42**(5), 1952–1979.

Dahl, D. B. (2006), Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model, *in* K.-A. Do, P. Müller & M. Vannucci, eds, 'Bayesian Inference for Gene Expression and Proteomics', Cambridge University Press, pp. 201–218.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. & Ruggiero, M. (2015), 'Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(2), 212–219.

de Finetti, B. (1937), 'La prévision: ses lois logiques, ses sources subjectives', *Annales de l'Institut Henri Poincaré* **7**(1), 1–68.

De Moivre, A. (1712), *The Doctrine of Chances*, Millar, London.

Dobigeon, N. & Toumeret, J. Y. (2007), 'Joint segmentation of wind speed and direction using a hierarchical model', *Computational Statistics and Data Analysis* **51**, 5603–5621.

Dobinski, G. (1877), 'Summierung der Reihe $\sum n^m/n!$ für $m = 1, 2, 3, 4, 5, \ldots$', *Grunert Archiv (Arch. Math. Phys.)* **61**, 333–336.

Epifani, I., Lijoi, A. & Prünster, I. (2003), 'Exponential functionals and means of neutral-to-the-right priors', *Biometrika* **90**(4), 791–808.

Escobar, M. D. & West, M. (1995), 'Bayesian Density Estimation and Inference Using Mixtures', *Journal of the American Statistical Association* **90**(430), 577–588.

Ewens, W. J. (1972), 'The sampling theory of selectively neutral alleles', *Theoretical Popu-*

lation Biology **3**(1), 87–112.

Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Prünster, I. & Teh, Y. W. (2015), 'On the Stick-Breaking Representation for Homogeneous NRMIs'.
URL: *http://doi.org/10.1214/15-BA964*

Favaro, S. & Teh, Y. W. (2013), 'MCMC for Normalized Random Measure Mixture Models', *Statistical Science* **28**(3), 335–359.

Fearnhead, P. & Liu, Z. (2007), 'Online Inference for Multiple Changepoint Problems', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 589–605.

Feller, W. (1943), 'On a General Class of "Contagious" Distributions', *Annals of Mathematical Statistics* **14**(4), 389–400.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, Vol. 1, third edn, John Wiley and Sons.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**(2), 209–230.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics* **7**(2), 179–188.

Flajolet, P. & Sedgewick, R. (2009), *Analytic Combinatorics*, Cambridge University Press.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer.

Fuentes-García, R., Mena, R. H. & Walker, S. G. (2009), 'A nonparametric dependent process for Bayesian regression', *Statistics & Probability Letters* **79**(8), 1112–1119.

Fuentes-García, R., Mena, R. H. & Walker, S. G. (2010*a*), 'A New Bayesian Nonparametric Mixture Model', *Communications in Statistics - Simulation and Computation* **39**(4), 669–682.

Fuentes-García, R., Mena, R. H. & Walker, S. G. (2010*b*), 'A Probability for Classification Based on the Dirichlet Process Mixture Model', *Journal of Classification* **27**, 389–403.

Geyer, C. J. (2007), 'Lower-Truncated Poisson and Negative Binomial Distributions'.
URL: *http://cran.r-project.org/web/packages/aster/vignettes/trunc.pdf*

Ghosal, S. & van der Vaart, A. (2015), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge University Press, to appear.

Gilks, W. R., Best, N. G. & Tan, K. K. C. (1995), 'Adaptive rejection Metropolis sampling', *Applied Statistics* **44**, 455–472.

Gnedin, A. & Pitman, J. (2005), Exchangeable Gibbs partitions and Stirling triangles, *in* 'Representation Theory, Dynamical Systems, Combinatorial and Algorithmic Methods. Part 12', Vol. 325 of *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI),*

PDMI, pp. 83–102.

Gnedin, A. V., Hansen, B. & Pitman, J. (2007), 'Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws', *Probability Surveys* **4**, 146–171.

Godbole, A., Goyt, A., Herdan, J. & Pudwell, L. (2014), 'Pattern Avoidance in Ordered Set Partitions', *Annals of Combinatorics* **18**(3), 429–445.

Godsill, S. J. (2001), 'On the Relationship Between Markov Chain Monte Carlo Methods for Model Uncertainty', *Journal of Computational and Graphical Statistics* **10**(2), 230–248.

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–732.

Green, P. J. & Richardson, S. (2001), 'Modelling Heterogeneity With and Without the Dirichlet Process', *Scandinavian Journal of Statistics* **28**(2), 355–375.

Griffin, J. & Holmes, C. (2010), Computational issues arising in bayesian nonparametric hierarchical models, *in* N. L. Hjort, C. C. Holmes, P. Müller & S. G. Walker, eds, 'Bayesian Nonparametrics', Cambridge University Press, pp. 208–222.

Griffiths, R. C. & Spanò, D. (2010), Diffusion processes and coalescent trees, *in* N. H. Bingham & C. M. Goldie, eds, 'Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman', Cambridge University Press.

Gross, O. A. (1962), 'Preferential Arrangements', *The American Mathematical Monthly* **69**(1), 4–8.

Gutiérrez, L., Gutiérrez-Peña, E. & Mena, R. H. (2014), 'Bayesian nonparametric classification for spectroscopy data', *Computational Statistics & Data Analysis* **78**, 56–68.

Haight, F. A. & Breuer, M. A. (1960), 'The Borel-Tanner distribution', *Biometrika* **47**, 145–150.

Halmos, P. R. (1944), 'Random Alms', *Annals of Mathematical Statistics* **15**(2), 182–189.

Harding, D. & Pagan, A. R. (2008), Business cycle measurement, *in* S. N. Durlauf & L. E. Blume, eds, 'The New Palgrave Dictionary of Economics', 2nd edn, Palgrave Macmillan.

Hartigan, J. A. (1990), 'Partition models', *Communications in Statistics - Theory and Methods* **19**, 2745–2756.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition*, 2 edn, Springer-Verlag New York.

Hershberger, S. L. (2005), History of Discrimination and Clustering, *in* 'Encyclopedia of Statistics in Behavioral Science', Vol. 2, John Wiley & Sons, pp. 840–842.

Hinkley, D. V. & Hinkley, E. A. (1970), 'Inference About the Change-Point in a Sequence of Binomial Variables', *Biometrika* **57**(3), 477–488.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G., eds (2010), *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Ishikawa, M., Kasraoui, A. & Zeng, J. (2008), 'Euler-Mahonian Statistics on Ordered Set Partitions', *SIAM Journal on Discrete Mathematics* **22**(3), 1105–1137.

Ishwaran, H. & James, L. F. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453), 161–173.

Jara, A., Lesaffre, E., De Iorio, M. & Quintana, F. (2010), 'Bayesian semiparametric inference for multivariate doubly-interval-censored data', *Annals of Applied Statistics* **4**(4), 2126–2149.

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

Jochmann, M. (2010), Modeling U.S. Inflation Dynamics: A Bayesian Nonparametric Approach, Technical Report 2010–06, Scottish Institute for Research in Economics (SIRE).

Kallenberg, O. (2005), *Probabilistic Symmetries and Invariance Principles*, Probability and its Applications, Springer, New York.

Kalli, M., Griffin, J. E. & Walker, S. G. (2011), 'Slice sampling mixture models', *Statistics and Computing* **21**(1), 93–105.

Kander, Z. & Zacks, S. (1966), 'Test Procedures for Possible Changes in Parameters of Statistical Distributions Occurring at Unknown Time Points', *Annals of Mathematical Statistics* **37**(5), 1196–1210.

Kaplan, A. Y. & Shishkin, S. L. (2000), Application of the change-point analysis to the investigation of the brain's electrical activity, *in* B. E. Brodsky & B. S. Darkhovsky, eds, 'Non-Parametric Statistical Diagnosis: Problems and Methods', Springer, pp. 333–388.

Karatzas, I. & Shreve, S. E. (1988), *Brownian Motion and Stochastic Calculus*, Springer-Verlag.

Kasraoui, A. & Zeng, J. (2009), 'Euler-Mahonian statistics on ordered set partitions (II)', *Journal of Combinatorial Theory, Series A* **116**(3), 539–563.

Kingman, J. F. C. (1967), 'Completely random measures', *Pacific Journal of Mathematics* **21**(1), 59–78.

Kingman, J. F. C. (1975), 'Random discrete distributions', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **37**(1), 1–22.

Kingman, J. F. C. (1978*a*), 'Random Partitions in Population Genetics', *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **361**(1704), 1–20.

Kingman, J. F. C. (1978*b*), 'The Representation of Partition Structures', *Journal of the London Mathematical Society* **s2-18**(2), 374–380.

Kingman, J. F. C. (1982), 'The coalescent', *Stochastic Processes and their Applications* **13**(3), 235–248.

Ko, S. I. M., Chong, T. T. L. & Ghosh, P. (2015), 'Dirichlet Process Hidden Markov Multiple Change-point Model', *Bayesian Analysis* **10**(2), 275–296.

Kolchin, V. F., Sevastyanov, B. A. & Chistyakov, V. P. (1978), *Random Allocations*, John Wiley and Sons.

Korwar, R. M. & Hollander, M. (1973), 'Contributions to the Theory of Dirichlet Processes', *The Annals of Probability* **1**(4), 705–711.

Lau, J. & Cripps, E. (2015), 'Stick-Breaking Representation and Computation for Normalized Generalized Gamma Processes', *Sankhya A* 77(2), 300–329.

Lee, J., Quintana, F. A., Müller, P. & Trippa, L. (2013), 'Defining Predictive Probability for Species Sampling Models Functions', *Statistical Science* **28**(2), 209–222.

Lijoi, A., Mena, R. H. & Prünster, I. (2005), 'Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors', *Journal of the American Statistical Association* **100**(472), 1278–1291.

Lijoi, A., Mena, R. H. & Prünster, I. (2007*a*), 'Bayesian nonparametric estimation of the probability of discovering a new species', *Biometrika* **94**, 769–786.

Lijoi, A., Mena, R. H. & Prünster, I. (2007*b*), 'Controlling the reinforcement in Bayesian non-parametric mixture models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(4), 715–740.

Lijoi, A. & Prünster, I. (2010), Models beyond the Dirichlet process, *in* N. L. Hjort, C. C. Holmes, P. Müller & S. G. Walker, eds, 'Bayesian Nonparametrics', Cambridge University Press, pp. 80–136.

Lijoi, A., Prünster, I. & Walker, S. G. (2008), 'Investigating nonparametric priors with Gibbs structure', *Statistica Sinica* **18**, 1653–1668.

Lo, A. Y. (1984), 'On a class of Bayesian nonparametric estimates: I. Density estimates', *The Annals of Statistics* **12**(1), 351–357.

Loschi, R. H. & Cruz, F. R. B. (2005), 'Extension to the product partition model: comput-

ing the probability of a change', *Computational Statistics & Data Analysis* **48**(2), 255–268.

Loschi, R. H., Cruz, F. R. B., Iglesias, P. L. & Arellano-Valle, R. B. (2003), 'A Gibbs sampling scheme to the product partition model: an application to change point problems', *Computers & Operations Research* **30**(3), 463–482.

MacEachern, S. N. (1994), 'Estimating normal means with a conjugate style Dirichlet process prior', *Communications in Statistics - Simulation and Computation* **23**(3), 727–741.

Mansour, T. (2013), *Combinatorics of set partitions*, Chapman & Hall/CRC.

Marin, J.-M., Mengersen, K. L. & Robert, C. (2005), Bayesian modelling and inference on mixtures of distributions, *in* D. Dey & C. R. Rao, eds, 'Handbook of Statistics: Volume 25', Elsevier.

Martínez, A. F. & Mena, R. H. (2014), 'On a Nonparametric Change Point Detection Model in Markovian Regimes', *Bayesian Analysis* **9**(4), 823–858.

McCullagh, P. & Yang, J. (2008), 'How many clusters?', *Bayesian Analysis* **3**(1), 101–120.

McLachlan, G. & Peel, D. (2000), *Finite Mixture Models*, Wiley-Interscience.

Mena, R. H. (2013), Geometric Weight Priors and their Applications in Bayesian Nonparametrics, *in* P. Damien, P. Dellaportas, N. G. Polson & D. A. Stephen, eds, 'Bayesian Theory and Applications', Oxford University Press, pp. 271–296.

Mena, R. H., Ruggiero, M. & Walker, S. G. (2011), 'Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling', *Journal of Statistical Planning and Inference* **141**(9), 3217–3230.

Mena, R. H. & Walker, S. G. (2009), 'On a construction of Markov models in continuous time', *METRON - International Journal of Statistics* **LXVII**(3), 303–323.

Mena, R. H. & Walker, S. G. (2012), 'An EPPF from independent sequences of geometric random variables', *Statistics & Probability Letters* **82**(6), 1059–1066.

Mena, R. H. & Walker, S. G. (2015), 'On the Bayesian mixture model and identifiability', *Journal of Computational and Graphical Statistics* **24**(4), 1155–1169.

Miller, J. W. & Harrison, M. T. (2014), 'Inconsistency of Pitman-Yor Process Mixtures for the Number of Components', *Journal of Machine Learning Research* **15**, 3333–3370.

Minin, V. N., Dorman, K. S., Fang, F. & Suchard, M. A. (2007), 'Phylogenetic Mapping of Recombination Hotspots in Human Immunodeficiency Virus via Spatially Smoothed Change-Point Processes', *Genetics* **175**(4), 1773–1785.

Mira, A. & Petrone, S. (1996), Bayesian hierarchical nonparametric inference for change–point problems, *in* J. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds, 'Bayesian Statistics 5', Oxford University Press, pp. 693–703.

Muliere, P. & Scarsini, M. (1985), 'Change–point problems: A Bayesian nonparametric approach', *Aplikace matematiky* **30**(6), 397–402.

Müller, P. & Quintana, Fernando A.and Rosner, G. L. (2011), 'A Product Partition Model With Regression on Covariates', *Journal of Computational and Graphical Statistics* **20**(1), 260–278.

Neal, R. M. (2000), 'Markov chain sampling methods for Dirichlet process mixture models', *Journal of Computational and Graphical Statistics* **9**(2), 249–265.

Nieto-Barajas, L. E. & Contreras-Cristan, A. (2014), 'A Bayesian Nonparametric Approach for Time Series Clustering', *BA* **9**(1), 147–170.

Noack, A. (1950), 'A class of random variables with discrete distributions', *Annals of Mathematical Statistics* **21**(1), 127–132.

Nobile, A. & Fearnside, A. T. (2007), 'Bayesian finite mixtures with an unknown number of components: the allocation sampler', *Statistics and Computing* **17**, 147–162.

Otter, R. (1949), 'The multiplicative process', *Annals of Mathematical Statistics* **20**(2), 206–224.

Page, E. S. (1954), 'Continuous Inspection Schemes', *Biometrika* **41**(1/2), 100–115.

Papaspiliopoulos, O. & Roberts, G. O. (2008), 'Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models', *Biometrika* **95**(1), 169–186.

Park, J.-H. & Dunson, D. B. (2010), 'Bayesian generalized product partition model', *Statistica Sinica* **20**(20), 1203–1226.

Patil, G. P. (1964), Estimation for the generalized power series distribution with two parameters and its application to binomial distribution, *in* C. R. Rao, ed., 'Contributions to Statistics', Statistical Publishing Society; Oxford: Pergamon, pp. 335–344.

Pearson, K. (1894), 'Contributions to the Mathematical Theory of Evolution', *Philosophical Transactions of the Royal Society of London A* **185**, 71–110.

Perman, M., Pitman, J. & Yor, M. (1992), 'Size–biased sampling of Poisson point processes and excursions', *Probability Theory and Related Fields* **92**, 21–39.

Perreault, L., Bernier, J., Bobée, B. & Parent, E. (2000), 'Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited', *Journal of Hydrology* **235**(3–4), 221–241.

Pitman, J. (1995), 'Exchangeable and partially exchangeable random partitions', *Probability Theory and Related Fields* **102**, 145–158.

Pitman, J. (1996), Some developments of the Blackwell-MacQueen urn scheme, *in* T. S. Ferguson, L. S. Shapley & J. B. MacQueen, eds, 'Statistics, Probability and Game Theory;

papers in honor of David Blackwell', Vol. 30 of *Lecture Notes-Monograph Series*, Institute of Mathematical Statistics, pp. 245–267.

Pitman, J. (2003), Poisson-Kingman partitions, *in* D. R. Goldstein, ed., 'Statistics and science: a Festschrift for Terry Speed', Vol. 40 of *Lecture Notes-Monograph Series*, Institute of Mathematical Statistics, pp. 1–34.

Pitman, J. (2006), *Combinatorial stochastic processes*, Ecole d'eté de probabilités de Saint-Flour XXXII - 2002, Springer.

Punskaya, E., Andrieu, C., Doucet, A. & Fitzgerald, W. J. (2002), 'Bayesian curve fitting using MCMC with applications to signal segmentation', *IEEE Transactions on Signal Processing* **50**, 747–758.

Quintana, F. A. & Iglesias, P. L. (2003), 'Bayesian Clustering and Product Partition Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2), 557–574.

Rao, C. R. (1947), 'A statistical criterion to determine the group to which an individual belongs', *Nature* **160**, 835–836.

Regazzini, E., Lijoi, A. & Prünster, I. (2003), 'Distributional results for means of normalized random measures with independent increments', *The Annals of Statistics* **31**(2), 560–585.

Reinhart, C. M. & Rogoff, K. S. (2004), 'The Modern History of Exchange Rate Arrangements: A Reinterpretation', *The Quarterly Journal of Economics* **119**(1), 1–48.

Richardson, S. & Green, P. J. (1997), 'On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(4), 731–792.

Rodríguez, C. E. & Walker, S. G. (2014), 'Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies', *Journal of Computational and Graphical Statistics* **23**(1), 25–45.

San Martín, E. & Quintana, F. A. (2002), 'Consistency and identifiability revisited', *Brazilian Journal of Probability and Statistics* **16**(1), 99–106.

Schervish, M. J. (1995), *Theory of statistics*, Springer series in statistics, Springer, New York, Berlin, Heidelberg.

Sethuraman, J. (1994), 'A constructive definition of Dirichlet priors', *Statistica Sinica* **4**(2), 639–650.

Shiryaev, A. N. (1963), 'On Optimum Methods in Quickest Detection Problems', *Theory of Probability & Its Applications* **8**(1), 22–46.

Smith, A. F. M. (1975), 'A Bayesian approach to inference about a change–point in a sequence of random variables', *Biometrika* **62**(2), 407–416.

Stirling, J. (1717), 'Methodus Differentialis Newtoniana Illustrata. Authore Jacobo Stirling, e Coll. Balliol. Oxon', *Philosophical Transactions (1683-1775)* **30**, 1050–1070.

Titterington, D., Smith, A. & Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

Truyen, T. T., Phung, D. Q. & Venkatesh, S. (2011), Probabilistic Models over Ordered Partitions with Applications in Document Ranking and Collaborative Filtering, *in* 'Proceedings of the 2011 SIAM International Conference on Data Mining', pp. 426–437.

Tryon, R. (1939), *Cluster Analysis*, McGraw-Hill, New York.

Uhlenbeck, G. E. & Ornstein, L. (1930), 'On the theory of Brownian motion', *Physical Review* **36**(5), 823–841.

von Mises, R. (1944), 'On the classification of observation data into distinct groups', *Annals of Mathematical Statistics* **16**(1), 68–73.

Wade, S., Walker, S. G. & Petrone, S. (2014), 'A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting', *Scandinavian Journal of Statistics* **41**(3), 580–605.

Walker, S. G. (2007), 'Sampling the Dirichlet Mixture Model with Slices', *Communications in Statistics - Simulation and Computation* **36**(1), 45–54.

Welch, B. L. (1939), 'Note on discriminant functions', *Biometrika* **31**(1/2), 218–220.

West, M. (1992), Hyperparameter estimation in Dirichlet process mixture models, Technical report, Institute of Statistics and Decision Sciences, Duke University.

Zantedeschi, D., Damien, P. & Polson, N. G. (2011), 'Predictive Macro-Finance With Dynamic Partition Models', *Journal of the American Statistical Association* **106**(494), 427–439.

Zhou, M. & Walker, S. G. (2014), Sample size dependent species models. arXiv:1410.3155.

Zubin, J. A. (1938), 'A technique for measuring like-mindedness', *Journal of Abnormal Psychology* **33**, 508–516.