Politecnico di Milano

Facoltà di Ingegneria dei Sistemi

Corso di Studi in INGEGNERIA MATEMATICA

Tesi di Laurea Magistrale

Model-based clustering via Bayesian nonparametric mixture models

Relatore: Prof. Alessandra GUGLIELMI

Correlatore: Dott. Raffaele ARGIENTO

Candidato:

Andrea CREMASCHI Matr. 751137

Anno Accademico 2011-2012

Contents

List of Figures

List of Tables

Introduction

In this work we discuss Bayesian methods for data clustering, reviewing some of the most popular ones in the literature, and propose some extensions. As an introduction to the work, a brief general overview (i.e. not only Bayesian) of the main classes of clustering methods is given, both from a heuristic and model-based clustering point of view. Heuristic methods encompass the well known agglomerative hierarchical clustering, the K-Means algorithm and the quite new "density-based" DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise, Ester et al., 1996). On the other hand, as far as model-based clustering is concerned, we give a parametric characterization in Chapter 1, in order to proceed with a deeper discussion on Bayesian nonparametric models in the first part of Chapter 2. There, the Bayesian nonparametric approach is presented, recalling some important mathematical aspects of the approach. Furthermore, two nonparametric priors are introduced, that is the Dirichlet process prior (Ferguson, 1973) and the Normalized Generalized Gamma process prior (Regazzini, Lijoi and Prünster, 2003) which is a generalization of the former. In particular, we discuss their mathematical definition, properties (such as the discreteness of their trajectories), and relationships. The second part of Chapter 2 is focused on two popular models involving nonparametric priors that will be used in the clustering later analysis: the Product Partition Model (or PPM, see Hartigan, 1990) and the Dirichlet Process Mixtures model (or DPM model, see Lo, 1984). In DPM models, data are supposed to be generated from a mixture of kernel densities, indexed by the vector of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, which is a finite sample from a realization of the nonparametric prior used (Dirichlet or NGG process prior in this work). In this case, each observation corresponds to a latent variable θ_i. Differently, the PPMs are nonparametric models in which the random partition represents the parameter and hence a prior for it has to be chosen. Usually the prior distribution of the random partition is called Exchangeable Partition Probability Function (or briefly EPPF). Moreover, we discuss the relationship between

the PPMs and the DPM models, showing how every DPM model can be related to a PPM with particular features (see Quintana and Iglesias, 2003). When we adopt a DPM model, thanks to the a.s. discreteness of the trajectories of the nonparametric prior introduced, ties in the vector $\boldsymbol{\theta}$ will appear with positive probability, thus yielding a random partition of the dataset. The index set $\{1, ..., n\}$ of the data is partitioned into $\Pi = \{C_1, \ldots, C_K\}$, which is random since $\boldsymbol{\theta}$ is. Each C_j in Π contains the labels of the latent variables in $\boldsymbol{\theta}$ with the same value.

The last part of Chapter 2 presents the most popular Bayesian nonparametric clustering methods using PPM or DPM models: the Stochastic Search by Posterior Sampling of Partitions (SS), the Bayesian Hierarchical Clustering Procedure (BH) and the Loss-Function Minimization methods. As far as the last method is concerned, this consists in the minimization of the posterior expected value of a loss function, where the parameter is the random partition; here we consider the loss function proposed in Binder (1978) and Lau and Green (2007), which is the sum of the costs of the misclassification errors that could occur. After a detailed presentation of these three methods, we propose a generalization of the last one, based on a different clustering criteria. We define a new decision rule for clustering, which is not based on the ties of the latent vector, but on the distance between its components. Instead of defining a $n \times n$ matrix $M = [m_{ij}]$ with $m_{ij} = 1$ when the two correspondent latent variables coincide [have the same value] and $m_{ij} = 0$ otherwise, we define $m_{ij} = 1$ if some distance between the kernel densities, corresponding to θ_i and θ_j, is smaller then a fixed threshold ϵ. In practice, this definition would not define an equivalence relation among the elements of $\{1, ..., n\}$; to overcome this problem, the DBSCAN is applied, defining a new equivalent relation among $\{1, ..., n\}$ and providing the correspondent partition. Thus, a matrix M is obtained, and then posterior expected value of the loss function by Binder (1978) can be evaluated. As before, the posterior expected value of the loss function is minimized with respect to the random partition, sampled using a MCMC algorithm. Of course, new parameters must be introduced into the analysis, such as the threshold ϵ, and N, which is the minimum number of elements in order to define a group cluster, via the DBSCAN algorithm. Additionally, the choice of the distance in the definition of the new equivalence relation deeply influences the estimates. Hence, different distances will be taken into account in this work.

All the methods introduced so far (SS, BH, loss-minimization and the new one) will be tested

on three different datasets. The Galaxy dataset, analyzed in Chapter 3, is very popular in the density estimation literature and contains observed velocities of $n = 82$ galaxies (univariate observations). We adopt a conjugate DPM model with Gaussian kernels and compute the Bayesian clustering (and density) estimates for all the proposed methods. In particular, we assume both Dirichlet and NGG process priors as mixing measure for the mixture. As far as the choice of the hyperparameters is concerned, a robustness analysis is carried on. As a second example, in Chapter 4 we analyze the Kevlar dataset, which consists of $n = 108$ univariate lifetimes of Kevlar fibres under different levels of stress. The stress levels represent the covariates. We adopt in this dataset both DPM models and NGG-mixture with Weibull kernel densities. Then we present the Bayesian clustering estimates resulting from the loss-function minimization methods with the standard and the new matrices M. Concerning the choice of the hyperparameters of this model, a vast robustness analysis has been performed, following the work by Argiento et al. (2010). In Chapter 5, we analyze a simulated bivariate dataset whose elements lie on a non-convex region. Tha analysis of this dataset is usually a difficul test for clustering methods in the literature. Therefore Chapter 5 shows comparisons between some of the heuristic methods presented in Chapter 1 and the new loss-function minimization method, applied to the test dataset.

Finally, in Chapter 6, some details on the algorithm used here for posterior sampling and density estimation are discussed. In particular, a description of the Polya urn sampling scheme for conjugate models is given, for both Dirichlet and NGG process prior. Furthermore, density estimates for the Galaxy and the simulated bivariate datasets are provided.

The main original contributions of this thesis are:

- The proposal of the new model-based clustering procedure based on Bayesian nonparametric mixture models and the DBSCAN algorithm. The proposed method seems to be completely new in the literature.

- Original coding in R and C of all the algorithms to compute posterior distribution and clustering estimates for all the application presented.

The clustering method presented in this work needs further developments; it seems interesting to elicit a prior distribution for the threshold value ϵ, on the basis of eventual prior information, and to design a MCMC algorithm for posterior sampling.

Sommario

In questo lavoro vengono studiati metodi di clustering bayesiano, introducendo alcuni dei più popolari in letteratura e proponendo alcune estensioni di questi. Come introduzione, viene proposta una rassegna delle due principali classi di metodi di clustering (non necessariamente bayesiani), ovvero della classe dei metodi euristici e di quelli "*model-based*". La prima di queste classi comprende tecniche quali i metodi gerarchici, l'algoritmo K-Means e il recente algoritmo "density-based" DBSCAN (Density-Based Spatial Clustering of Applications with Noise, Ester et al., 1996). Per quanto riguarda la seconda classe di metodi, ovvero quelli *model-based*, nel Capitolo 1 viene fornita una caratterizzazione di tipo parametrico, per poi procedere con una più approfondita presentazione dei modelli bayesiani non parametrici, nel Capitolo 2, assieme ad un'introduzione all'approccio bayesiano non parametrico. Inoltre, vengono introdotte le due prior non parametriche: il processo di Dirichlet (Ferguson, 1973) e il processo Normalized Generalized Gamma (Regazzini, Lijoi and Prünster, 2003), che è una generalizzazione del primo. In particolare, vengono presentate le loro definizioni formali, alcune proprietà di rilievo (quali la discretezza delle loro traiettorie) e le relazioni che li legano. La seconda parte del Capitolo 2 si concentra su i due più famosi modelli di tipo non parametrico, che saranno usati nella successiva *cluster analysis*: i Product Partition Models (o PPMs - Hartigan, 1990) e i Dirichlet Process Mixture models (o DPM - Lo, 1984). Nei modelli DPM, i dati sono generati da una mistura di densità, indicizzate da un vettore di parametri $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, il quale, a sua volta, è un campione finito da una realizzazione della prior non parametrica inserita nel modello (in questo lavoro, o processo di Dirichlet, o processo NGG). In questo caso, ogni osservazione corrisponde ad una variabile latente θ_i. Invece, i PPMs sono modelli nonparametrici in cui la partizione aleatoria rappresenta il parametro, e alla quale quindi va assegnata una distribuzione a priori. Solitamente, la distribuzione a priori per le partizioni aleatorie è chiamata Exchangeable Partition Probability Function (o brevemente EPPF). Inoltre, sempre nel

Capitolo 2, viene presentata la relazione che lega le due classi di modelli, illustrando come un modello DPM può essere associato ad un PPM con particolari caratteristiche (vedi Quintana e Iglesias, 2003). Quando utilizziamo un modello DPM, grazie alla proprietà di discretezza delle traiettorie della prior non parametrica, il vettore $\boldsymbol{\theta}$ ha probabilità positiva di contenere delle coincidenze, inducendo una partizione aleatoria sul nostro dataset. L'insieme di indici $\{1, ..., n\}$ dei dati è partizionato attraverso $\Pi = \{C_1, \ldots, C_K\}$, che è anch'essa aleatoria, poichè lo è il vettore $\boldsymbol{\theta}$. Ogni C_j in Π contiene le etichette delle variabili latenti presenti in $\boldsymbol{\theta}$ che hanno lo stesso valore numerico. L'ultima parte del Capitolo 2 presenta i principali metodi di clustering bayesiano non parametrico che utilizzano PPMs o modelli DPM, e che sono: il metodo Stochastic Search by Posterior Sampling of Partitions (SS), il metodo Bayesian Hierarchical Clustering Procedure (BH) e i metodi di minimizzazione di funzionali di costo. L'ultimo di questi metodi consiste nella minimizzazione del valore atteso a posteriori di una funzione di costo della partizione aleatoria. Qui consideriamo quella proposta da Binder (1978) e Lau e Green (2007), che è definita come la somma dei costi associati alle misclassificazioni. Dopo un'esposizione dettagliata dei tre metodi, viene proposta una generalizzazione dell'ultimo, basata su un diverso criterio di clustering. Definiamo una nuova regola per classificare i dati, che non è basata sulle coincidenze nel vettore di variabili latenti, ma sulla loro distanza. Invece di definire una matrice $n \times n$ $M = [m_{ij}]$, dove $m_{ij} = 1$ se le due corrispondenti variabili latenti coincidono e $m_{ij} = 0$ altrimenti, definiamo $m_{ij} = 1$ se un'opportuna distanza tra le corrispondenti densità parametriche è minore di una soglia fissata ϵ. Questa definizione non definisce una relazione di equivalenza sull'insieme $\{1, ..., n\}$ perchè la proprietà di transitività non è verificata; per superare questa limitazione, applichiamo l'algoritmo DBSCAN, che definisce una nuova relazione di equivalenza su $\{1, ..., n\}$ e utilizziamo la partizione così generata. Dunque, otteniamo una matrice M con la quale è possibile valutare il valore atteso a posteriori della funzione di costo di Binder (1978). Come per la vecchia relazione di equivalenza, il valore atteso a posteriori della funzione di costo viene minimizzato rispetto alla partizione aleatoria, campionata attraverso un algoritmo MCMC. Ovviamente, devono essere introdotti nuovi parametri in questa analisi, come la soglia ϵ o il valore N, che rappresenta il minimo numero di elementi necessari per poter definire un *cluster* nell'algortimo DBSCAN. Inoltre, la scelta della distanza nella definizione della nuova relazione di equivalenza influenza molto le stime fornite dal metodo. Per questo

verranno considerate diverse distanze.

Tutti i metodi introdotti finora (SS, BH, minmizzazione di un funzionale di costo) sonostati verificati su tre diversi dataset. Il dataset Galaxy, analizzato nel Capitolo 3, è molto utilizzato nella letterature sulla stima di densità e contiene i valori (univariati) di $n = 82$ velocità di galassie. Abbiamo adottati un modello DPM di tipo coniugato con kernel Gaussiani e abbiamo calcolato le stime di clustering bayesiano e di densità per tutti i metodi proposti. In particolare, utilizzeremo sia il processo di Dirichlet che il processo NGG come misure misturanti. Per quanto riguarda la scelta degli iperparametri del modello, è stata svolta un'analisi di robustezza. Come secondo esempio, nel Capitolo 4, abbiamo analizzato il dataset Kevlar che consiste di $n = 108$ tempi di vita (univariati) di altrettante fibre Kevlar, sottoposte a differenti livelli di stress, che rappresentano le covariate in questo modello. Abbiamo adottato un modello DPM con misturante data da un processo NGG e kernel Weibull. Dunque, abbiamo calcolato le stime bayesiane risultanti dall'applicazione dei metodi di minimizzazione della funzione di costo. L'analisi di robustezza per la scelta degli iperparametri è stata svolta seguendo il lavoro di Argiento et al. (2010). Nel Capitolo 5, abbiamo presentato l'esempio di un dataset bivariato i cui elementi sono disposti in una regione non convessa. L'analisi di questo dataset è solitamente un test difficile per i metodi di clustering che si trovano in letteratura. Per questo motivo mostriamo in questo capitolo alcuni confronti tra i metodi euristici presentati nel Capitolo 1 e il nuovo metodo di minimizzazione della funzione di costo. Il capitolo si conclude con alcuni commenti.

Infine, nel Capitolo 6, vengono forniti alcuni dettagli sul campionamento a posteriori e dei risultati di stime di densità. In particolare, viene descritto il funzionamento dell'algoritmo di campionamento ad urna di Polya, sia per il processo di Dirichlet che per il processo NGG, accompagnato dai risultati di stima di densità ottenuti per il caso del dataset univariato Galaxy e del dataset simulato bivariato.

I principali contributi originali che si trovano in questo lavoro sono:

- La proposta di un nuovo metodo di clustering bayesiano di tipo *model-based*, basato su modelli non parametrici di tipo mistura e sull'algoritmo DBSCAN. Il metodo sembra essere completamente nuovo in letteratura.

- La programmazione in R e in C degli algoritmi per il campionamento a posteriori e per

l'implementazione dei metodi di clustering per le applicazioni presentate.

Il metodo di clustering presentato in questo lavoro necessità di ulteriore lavoro; ci sembra promettente l'idea di assumere una prior per la soglia ϵ, sulle basi di eventuali informazioni a priori, e di calcolare le corrispondenti stime a posteriori, disegnando un algoritmo di tipo MCMC.

Chapter 1

Data Clustering

One of the most important issues in modern Data Analysis is to get knowledge of the inner structure of the data which are going to be analyzed. This would be a solid starting point from which to describe the composition of the data-set, making the comprehension of the problem much easier, as well as the development of the analysis itself. Unfortunately, it is often impossible to describe the relationships which stand between the data precisely, making very hard to get the information needed to treat them in the proper manner.

To overcome this problem, many explorative methods have been proposed and studied, to form the well known Data Mining techniques. The name *Data Mining* refers to the explorative processes that are executed by analysts to get a better knowledge of the problem, in the sense of data representation and meaning. These theories and methods include Discriminant Analysis and Principal Components Analysis, as well as Clustering Techniques, which are what concern this work.

Cluster analysis is the statistical field involving all the theories, methods, algorithms and techniques whose goal is to re-organize the data-set, exploiting its peculiar features. From the word *cluster* it is clear that the aim of cluster analysis is to gather the data into distinct groups (or clusters), according to some similarity measure, or following a particular statistical model. What we need is to characterize the data assigning a label to each observation, aiming at enlightening the "true" representation of the entire data-set (or the most close to that). To do so, a lot of techniques have been presented in the last decades, yielding to many different results, as a proof of the importance of cluster analysis in Applied Statistics. One of the consequences of this fast growth is the necessity of finding a proper classification of these methods, based on

their functional, mathematical or methodological aspects.

As far as this work concerns, it is useful to start presenting a classification of the clustering techniques based on the statistical definition of the problem. To do so, it is useful to distinguish between *model-based* and *heuristic* clustering techniques. The first class refers to those methods that require a mathematical model describing the problem; the latter includes those algorithms defined from a given starting point, and carried on following some heuristic scheme (such as hierarchical or greedy scheme). It is useful to point out that the discriminant factor between these two classes of methods is not the usage of mathematical tools, but the presence of a valid statistical model underneath the analysis of the data. In fact, heuristic techniques use a lot of mathematical instruments to face cluster analysis, and often the process itself is based on some mathematical results (i.e., definitions and properties of particular metrics, minimization of some functionals...). Another aspect to mention concerns the existence of some methods which combine these two main characteristics, in which a statistical valid model is defined and, starting from it, a heuristic algorithm is used to perform the clustering. In this work, the attention is focused on model-based clustering techniques, in particular using a *Bayesian nonparametric* approach.

1.1 Heuristic Clustering

As mentioned before, many clustering techniques have been proposed in recent decades, and most of them are heuristic-based. So, it would be too onerous to describe them in details. For this reason the methods presented in this section are those of any relevance in this work, having connections with the statistical analysis presented or being used to offer a comparison of the results. For further details on heuristics clustering algorithms see Jain and Dubes (1988).

Heuristic methods can be seen as algorithms which take as an input the data-set and an initialization of interesting variables (sometimes also a terminal condition is given, such as a threshold condition upon some parameters), and give as output the final "grouping" of the data. The different schemes of these algorithms lead to a further subdivision: *hierarchial*, *partitive* and *density-based* clustering algorithms. The last class has been recently introduced by Ester, Kriegel and Xu (1995), and takes its name from the disposition of the observed data (from here the word *density*, not related with the density function of any random variable).

Hierarchical algorithms have been firstly proposed by Johnson (1967), and take their name from the procedure followed to group the data. Starting as n distinct clusters (where n is the size of the data-set), the objects are sequentially unified into sub-clusters, in order to minimize the *dissimilarity* between the data that are gathered in the same cluster. Step by step, always bigger sub-clusters are created, until all the data-set represents a unique cluster. The evaluation criterion of this algorithm is the dissimilarity between the data. It is important to indicate that to find a proper dissimilarity to describe the relationships between the data is often not easy, especially because it depends on problem features and on data characterization. In most of the cases, the measure of the dissimilarity is represented by the distances between the data (Euclidean distance is the most used), or by some customer-defined function of the elements (useful when working with categorical data).

Partitive algorithms find their most famous representative in the *K-means* (or *K-centroids*) algorithm. The term "K-means" was first used by James MacQueen (1967). It is an iterative process that, at each step, performs these two actions: 1) assigning the observations to the closer among K given points, called centroids; 2) finding the new set of centroids, defined as those points minimizing the sum of the distances in each of the clusters found at the previous point (when working with the Euclidean distance, these points are the sample means). The input of the algorithm consists of the number of clusters K and the relative centroids. Even if this second information can be quite general and with poor influence on the final result, thanks to the convergence of the algorithm, the same cannot be said for the first one. In fact, a different fixed number of clusters lead to very different partitioning of the data-set. Actually, this represents a drawback of this algorithm, together with the strong dependence on the choice of the distance to use.

The group of the density-based algorithms is the newest of the three, and finds an example in the *DBSCAN* method (Density-Based Spatial Clustering of Applications with Noise, Ester et al. (1996)). This technique relies on a density-based notion of clusters which is designed to discover clusters of arbitrary shape (an advantage with respect to a lot of methods). As mentioned before, the term "density" is not related to any density function of some random variable, but refers to the spatial disposition of the observations. DBSCAN requires only two input parameters: the minimum number of points to define a group to be a cluster, and the

maximum distance ϵ between the elements of the same cluster. One of the advantages of this method is to succeed in finding those data for which is difficult to decide a cluster label, therefore classified as "noise" (which can be seen as a cluster itself).
In many situations, lots of information are required to carry out a good heuristic cluster analysis, such as the number of clusters, or the distance to use. Usually, in heuristic clustering, the choice of the algorithm depends on the computational effort and on the purpose of the analysis.

1.2 Model-Based Clustering

Before introducing the main features of model-based clustering methods, it is useful to point out some statistical issues, concerning not only this work, but also the statistical field in which it is collocated. This is the reason why a sharp distinction between frequentist and Bayesian framework is done. In this section, a short overview of parametric frequentist model-based clustering techniques is given, with some recall to Bayesian parametric results, in order to present, in details in the next chapter, the *model-based clustering via Bayesian nonparametrics*.
Model-based clustering differs from heuristic clustering for the starting mathematical specification: the mathematical model for the problem is specified, instead of initializing some parameters according to available information about the problem or the data. This is a crucial difference in the way of explaining the data and the final results.
In frequentist model-based clustering, the data are often modelled through a finite mixture of kernel densities, in order to represent the partition into clusters as the sum of K distinct distributions. In case of real data vectors, these kernel densities are often chosen to be Gaussian, to approximate the real model structure in the smoother possible way. Similarly to heuristics, the number of clusters (i.e., the number of densities in the mixture) has to be fixed. Nevertheless, in this case, this number is treated as a parameter, and then an estimate for it is given. As presented by Fraley and Raftery (1998, 2002), the model in case of Gaussian mixture is the

following:

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{K} \tau_j N(\cdot; \boldsymbol{\mu}_j, \Sigma_j)$$

$$\tau_j = \mathbb{P}(\boldsymbol{X}_i \in C_j), \text{ for } i = 1, ..., n$$

$$\sum_{j=1}^{K} \tau_j = 1,$$

where K stands for the number of clusters, C_j is the j-th cluster, and $N(\cdot; \boldsymbol{\mu}_j, \Sigma_j)$ refers to the j-th normal density with mean vector $\boldsymbol{\mu}_j$ and covariance matrix Σ_j.

From a frequentist point of view, the determination of the model structure can be performed by model selection techniques and parameter estimation using maximum likelihood algorithms, such as the *EM algorithm* (Dempster et al., 1977): this consists of a two-steps iterative algorithm used to find the ML estimates of parametric models involving a latent set of variables, in this case the group labels. Despite the EM algorithm converges to a local optimum under proper conditions, estimation for mixture models has a number of limitations. First, the rate of convergence can be slow. Second, the EM algorithm for multivariate normal mixtures breaks down when the covariance matrix associated with one or more components is ill-conditioned. An example of the application of the EM algorithm for clustering purposes has been proposed by Fraley and Raftery (1998, 2002), where a proper likelihood ratio is used in order to chose between different models, not only as far as the number of components of the mixture is concerned, but also for the parameters of the underling normal densities of the various components.

A possible variation in the model above is achieved by assuming a Bayesian approach, consisting in fixing a prior distribution for the model's parameters, such as kernel densities' parameters $(\boldsymbol{\mu}, \Sigma)$, the vector $\boldsymbol{\tau}$ and the number of clusters. An example is given in Heard, Holmes and

Stephens (2006), which present the following *parametric Bayesian model*:

$$
\begin{aligned}
&\boldsymbol{X}^{(j)}|\boldsymbol{\mu}_j, \Sigma_j, n_j, K \overset{\text{i.i.d.}}{\sim} N_{n_j}(\boldsymbol{\mu}_j, \Sigma_j), \quad \text{for } j = 1:K \\
&\boldsymbol{\mu}_j|\Sigma_j, n_j, K \sim N_{n_j}(\boldsymbol{m}, \frac{\Sigma_j}{k_0}), \quad \text{for } j = 1:K \\
&\Sigma_j|\boldsymbol{n}, K \sim IW(\nu_1, \Psi_1), \quad \text{for } j = 1:K \\
&\boldsymbol{n} = (n_1, ..., n_K)|K, \boldsymbol{\tau} \sim Multinomial(\tau_1, ..., \tau_K) \\
&\boldsymbol{\tau} = (\tau_1, ..., \tau_K)|K \sim Dir(\alpha_1, ..., \alpha_K) \\
&K \sim U(\{1, ..., n\})
\end{aligned}
$$

In this model, the vector $\boldsymbol{n} = (n_1, ..., n_K)$ is the vector containing the number of elements for each cluster. The j-th group of data is represented by the multivariate random variable $\boldsymbol{X}^{(j)}$, having a multivariate normal kernel density.

The analysis in Bayesian model-based clustering techniques is performed by evaluating the posterior estimates of the interesting parameters (such as the number of clusters). For this purpose, a wide range of models have been studied, in order to find the best result for many clustering problems, providing better solutions with respect to the frequentist methods. This is the reason why, in this work, the Bayesian approach is adopted.

So far, only parametric models have been presented, both form a frequentist and Bayesian point of view. Nevertheless, the main approach adopted in this work is the nonparametric one, allowing more flexibility and robustness to the parameter estimation. To introduce Bayesian nonparametric models, some mathematical instruments are required, such as the definition of nonparametric priors, presented in the next chapter.

Chapter 2

Bayesian nonparametric model-based clustering

2.1 The Bayesian nonparametric approach

Usually, in classical statistics, the term *nonparametric* refers to those techniques that do not rely on the hypothesis that the data belong to some particular distribution, the probability model chosen to describe the observations being totally general. This means to assume that the data are distributed according to an element of a family of distributions, which cannot be put into a bijection with a finite-dimensional parameter. In this sense, the distribution itself becomes the parameter of the model, and the classical nonparametric inference aims at estimating it.

The same generalization can be achieved in a Bayesian fashion, where the *parameter* of the model is the unknown distribution. The random variable representing the unknown distribution in the model is called *random probability measure* (or briefly r.p.m.). In order to introduce the most common r.p.m.'s, the definition of probability measures on a collection of distribution functions is required, together with some notation.

Let $\mathcal{P}(\mathbb{X})$ be the space of all probability measures on $\mathbb{X}$, where $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ is a Polish space ($\mathcal{B}(\mathbb{X})$ indicates the Borel σ-algebra on $\mathbb{X}$). In order to define a σ-field of subsets of $\mathcal{P}(\mathbb{X})$, it is necessary to introduce the definition of *weak convergence*.

A sequence $\{P_n\}$ of probability measures on $\mathbb{X}$ is said to **converge weakly** to a probability measure P, written as $P_n \xrightarrow{\text{w}} P$, if $\int f dP_n \to \int f dP, \forall f \in \mathcal{C}(\mathbb{X})$, where $\mathcal{C}(\mathbb{X})$ is the set of all

bounded continuous functions on $\mathbb{X}$.

Under the topology of weak convergence, the space $\mathcal{P}(\mathbb{X})$ is metrizable, complete and separable (i.e., Polish) (for further details see Ghosh and Ramamoorthy, 2003). After this specification, the notion of random probability measure (r.p.m.) can be introduced.

Definition 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A* random probability measure *is a measurable function $P : (\Omega, \mathcal{F}) \to (\mathcal{P}(\mathbb{X}), \mathcal{B}(\mathcal{P}(\mathbb{X})))$.*

Thus, P is a random variable with values in the space of all probability measures, and, for each $\omega \in \Omega$, $P(\omega)$ is a probability measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$.

Any random probability measure can be seen as the parameter for the model or, equivalently, as a random variable taking values in the space of all probability measures, following its law Π; therefore it can be introduced in the model as done in the parametric case. In this case, the random vectors representing the observations are (conditionally) independent and identically distributed according to the unknown probability distribution P which is random, with distribution Π, i.e.

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n | \boldsymbol{P} = P \overset{\text{i.i.d.}}{\sim} P$$

$$\boldsymbol{P} \sim \Pi.$$

This approach is equivalent to assume only that the sequence $\{\boldsymbol{X}_n\}$ is *exchangeable*.

Definition 2. *Let $\{X_1, ..., X_n\}$ be a finite set of random variables taking values in the space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. The set is said to be* ***exchangeable*** *if, for every permutation σ of the indices $\{1, ..., n\}$, the joint probability distribution of the permuted sequence $(X_{\sigma(1)}, ..., X_{\sigma(n)})$ is the same of the original sequence. An infinite collection $(X_n)_{n \geq 1}$ of random variables is exchangeable if every finite sub-sequence is exchangeable.*

Provided this fundamental definition, we can now state the theorem allowing the formalization of the nonparametric model, called the *de Finetti's representation theorem.*

Theorem 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ a Polish space. For each $n \in \mathbb{N}$, consider the measurable functions $X_n : (\Omega, \mathcal{F}) \to (\mathbb{X}, \mathcal{B}(\mathbb{X}))$. A sequence $(X_n)_{n \geq 1}$ is exchangeable if and only if there exists a r.p.m. $\boldsymbol{P}$ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ such that, conditionally on $\boldsymbol{P}$=P, $(X_n)_{n \geq 1}$ are independent and identically-distributed (i.i.d.) according to P. Furthermore,*

if the sequence is exchangeable, then Π*, the distribution of* $\boldsymbol{P}$*, is unique, and the following convergence result holds:*

$$\boldsymbol{P}_n(B) = \frac{1}{n}\sum_{i=1}^{n} I_B(X_i) \overset{\text{a.s.}}{\to} \boldsymbol{P}(B), \qquad \forall B \in \mathcal{B}(\mathbb{X})$$

This result takes a central place into the Bayesian nonparametric approach, relating the random probability measure to the data description and hence formalizing the model construction. In Bayesian applications, the distribution of a r.p.m. represents the *nonparametric prior*. We point out that, in this work, it is enough to consider $\mathbb{X} = \mathbb{R}^k$, where k is a positive integer.

2.2 Nonparametric priors

Let $\mathcal{D}$ be the class of nonparametric priors on $\mathcal{P}(\mathbb{X})$. According to Ferguson (1973) and Antoniak (1974), three desirable properties of a nonparametric prior distribution can be highlighted:

1. The class $\mathcal{D}$ of random prior distributions on $\mathcal{P}(\mathbb{X})$ should be analytically tractable in three respects:

   (a) It should be reasonably easy to determine the posterior distribution on $\mathcal{P}(\mathbb{X})$ given a "sample";

   (b) It should be possible to express conveniently the expectations of simple loss functions;

   (c) The class $\mathcal{D}$ should be closed, in the sense that if the prior is a member of $\mathcal{D}$, then the posterior is a member of $\mathcal{D}$.

2. The class $\mathcal{D}$ should be "rich", so that there will exist a member of $\mathcal{D}$ capable of expressing any prior information or belief.

3. The class $\mathcal{D}$ should be parameterized in a manner which can be readily interpreted in relation to prior information and belief.

These requirements are not mutually exclusive, although they seem to be antagonist, in the sense that some property may be reached in the expense of another. Furthermore, points 1(a) and 1(b) do not represent a problem nowadays, thanks to application of MCMC techniques. Ferguson (1973) presented a process (i.e., the *Dirichlet process*) which satisfies the above stated

requirements. Another important process is the *Normalized Generalized Gamma process*, or NGG, presented for the first time by Regazzini, Lijoi and Prünster (2003). This process can be seen as a generalization of the Dirichlet process, tuned by an additional parameter.
In next two sections, definitions and properties of the two processes are given.

2.2.1 The Dirichlet process prior

In this section a very useful family of prior distributions on $\mathcal{P}(\mathbb{X})$ is developed: the Dirichlet process priors. The Dirichlet process arises naturally as an infinite-dimensional analogue of the finite-dimensional Dirichlet prior, which generalizes in more dimensions the beta distribution. A review of the finite-dimensional case is now given.

Finite Dimensional Dirichlet Distribution

Let $\mathbb{X} = \{1, 2, ..., k\}$ be a finite set of elements. Then the space of all the probability distributions on $\mathbb{X}$ is represented by the $(k-1)$-dimensional simplex:

$$S_k = \{\boldsymbol{p} = (p_1, ..., p_{k-1}) : p_i \geq 0 \text{ for } i = 1, 2, ..., k-1, \sum_{i=1}^{k-1} p_i \leq 1\}$$

where each $\boldsymbol{p}$ in S_k is a suitable prior probability vector for the elements of $\mathbb{X}$. It is evident that $p_k = 1 - \sum_{i=1}^{k-1} p_i$ (from here the (k-1)-dimensionality of the simplex). A "natural" prior distribution for the vector $\boldsymbol{p}$ is the *k-dimensional Dirichlet distribution.*

Definition 3. *Let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$, with $\alpha_i > 0$ for $i = 1, 2, ..., k$. Then $\boldsymbol{p} = (p_1, ..., p_k)$ has **Dirichlet distribution** with parameters $(\alpha_1, ..., \alpha_k)$ if its density is:*

$$f(p_1, ..., p_{k-1}) = \frac{\Gamma(\alpha_1 + ... + \alpha_k)}{\Gamma(\alpha_1) \cdot ... \cdot \Gamma(\alpha_k)} (\prod_{j=1}^{k-1} p_j^{\alpha_j - 1})(1 - \sum_{j=1}^{k-1} p_j)^{\alpha_k - 1} \mathbb{I}_{S_k}(p_1, ..., p_{k-1}).$$

If any $\alpha_i = 0$, the Dirichlet distribution still exists, fixing the corresponding $p_i = 0$, and degenerating on a lower-dimensional set.
An equivalent definition of the finite-dimensional Dirichlet distribution can be obtained starting from the gamma one.

Definition 4. *Let X be a random variable, taking values on the positive real line. Let α, β be two positive real numbers. X is said to have a **gamma distribution** with parameters (α, β),*

writing Gamma(α, β) *if its density is of the form:*

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta} I_{(0,+\infty)}(x).$$

Let now consider the independent r.v.'s $Z_1, ..., Z_k$, each one distributed as $Gamma(\alpha_j, \beta)$, for j=1,...,k, given the set of real positive numbers $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$.

Definition 5. *The random vector* $\boldsymbol{Y} = (Y_1, ..., Y_k)$*, where*

$$Y_j = \frac{Z_j}{\sum_{i=1}^k Z_i} \qquad j = 1, ...k$$

has k-dimensional Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$*. Furthermore,* $\boldsymbol{Y}$ *is independent of* $\sum_{i=1}^k Z_i$*.*

The Dirichlet process

As a final step, the infinite-dimensional generalization of the Dirichlet distribution can be presented. It is enough to replace the finite space $\mathbb{X}$ with the real space $\mathbb{R}^k$. Let $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ be the real line with the Borel σ-algebra $\mathcal{B}$ and let $\mathcal{P}(\mathbb{R}^k)$ be the set of probability measures on $\mathbb{R}^k$, equipped with the proper Borel σ-algebra $\mathcal{B}(\mathcal{P}(\mathbb{R}^k))$. The definition of Dirichlet process can now be given.

Definition 6. *Let* α *be a finite measure on* $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$*. A r.p.m.* P *on* $\mathbb{R}^k$ *has a* ***Dirichlet process prior*** *with parameter* α *if, for every finite and measurable partition* $B_1, B_2, ..., B_m$ *of* $\mathbb{R}^k$*, then* $(P(B_1), P(B_2), ..., P(B_m)) \sim D(\alpha(B_1), \alpha(B_2), ..., \alpha(B_m))$*.*

In the following we will write $\mathcal{D}_\alpha$ to denote the Dirichlet process with parameter α. Some important results about the Dirichlet process are now presented:

1. Conditionally on P in $\mathcal{P}(\mathbb{R}^k)$, let $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ be i.i.d. from P and let P itself be distributed as $\mathcal{D}_\alpha$, where α is a finite measure. Then, the posterior distribution of P given $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ is $\mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$.

2. From simple computations, it is clear that $\boldsymbol{X}_i \sim \overline{\alpha}$, where $\overline{\alpha}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathbb{R}^k)}$. For simplicity of notation, $\overline{\alpha}$ will be denoted as P_0. Clearly, this is a probability measure on the sample space $\mathbb{R}^k$. In the same way, $\mathbb{E}[P] = P_0$, where P stands for the random vector originated

from a given partition of the sample space. To see this, note that for each $A \subset \mathbb{R}^k$, $P(A)$ is $Beta(\alpha(A), \alpha(A^c))$ and hence $\mathbb{E}[P(A)] = \frac{\alpha(A)}{(\alpha(A)+\alpha(A^c))} = \frac{\alpha(A)}{\alpha(\mathbb{R}^k)}$. This parameter will be included in the notation writing $\mathcal{D}(\alpha, P_0)$.

The aim of this work is to exploit the power and the flexibility of nonparametric priors to support the cluster analysis. In order to do that, it is necessary to characterize the sample from a Dirichlet process, that is, to specify what kind of probability measures are induced on the space $\mathbb{R}^k$ by the Dirichlet process prior. The next theorem gives a discrete representation of Dirichlet process' trajectories.

Theorem 2. $\mathcal{D}_\alpha\{P : P \textit{ is discrete }\} = 1$

This theorem yields that there is a non-zero probability to observe common values in a sample $(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ from the Dirichlet process, which induces a subdivision (i.e., a clustering) of the sample, based on the value of the observations. The coincidence of observed values can be represented through a *generalized Polya Urn scheme*, which means that a sample from the Dirichlet process can be viewed as one from a Polya Urn allowing a continuum set of colours (see Blackwell and MacQueen, 1973). In other words, the joint law of the sample $\mathcal{L}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ can be computed from the following full conditionals:

$$\begin{aligned} \boldsymbol{X}_1 &\sim P_0 \\ \boldsymbol{X}_i|\boldsymbol{X}_1, ..., \boldsymbol{X}_{i-1} &\sim \frac{\alpha + \sum_{j=1}^{i-1} \delta_{\boldsymbol{X}_j}}{\alpha(\mathbb{R}^k) + i - 1} \text{ for } i = 1, ..., n, \end{aligned} \tag{2.1}$$

Iterating, we have:

$$\mathcal{L}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \mathcal{L}(\boldsymbol{X}_n|\boldsymbol{X}_{n-1}, ..., \boldsymbol{X}_1)\mathcal{L}(\boldsymbol{X}_{n-1}, ..., \boldsymbol{X}_1) = (...) = \mathcal{L}(\boldsymbol{X}_1)\prod_{j=1}^{n-1}\mathcal{L}(\boldsymbol{X}_{j+1}|\boldsymbol{X}_j, ..., \boldsymbol{X}_1) \tag{2.2}$$

With this scheme, it is easier to see how the Dirichlet process induces a partition on the observations. Of course, the partition of the data itself represents a random variable, as well as the number of clusters K_n. This is the key point of the *Bayesian nonparametric clustering* via Dirichlet process, and represents the basis of this work. Referring to Antoniak (1974), the density of the random variable K_n (i.e., the number of distinct values in $(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$) is the following:

$$\mathbb{P}(K_n = k) = |S_1(n,k)|a^k \frac{\Gamma(n)}{\Gamma(a+n)}, \text{ for } k \in \{1,...,n\} \tag{2.3}$$

where a stands for $\alpha(\mathbb{X})$, n is the data size, and $\Gamma(\cdot)$ is the Eurler's gamma function. $S_1(n,\cdot)$ are the absolute values of the Stirling numbers of the first kind.

2.2.2 The NGG process prior

In this section, the NGG process is presented, following the definition given in Argiento, Guglielmi and Pievatolo (2010).

Let μ be a random measure on $(\Theta, \mathcal{B}(\Theta))$, let $\sigma \in [0,1]$, ω, κ be non-negative parameters, and $P_0(\cdot)$ a non-atomic probability measure on Θ; Θ can be any Polish space. According to the notation in Brix (1999), we say that μ is a *generalized gamma measure* if

$$\mu(B) = \int_{\mathbb{R}^+} yN(dy, B), \qquad B \in \mathcal{B}(\Theta),$$

where N is a Poisson random measure on $\mathbb{R}^+ \times \Theta$ with mean measure ν defined by

$$\nu(A \times B) = \kappa P_0(B) \int_A \rho(ds), A \in \mathcal{B}(\mathbb{R}^+), B \in \mathcal{B}(\Theta),$$

and

$$\rho(ds) = \frac{1}{\Gamma(1-\sigma)} s^{-\sigma-1} e^{-\omega s} ds, s > 0.$$

By definition, μ is a random measure with independent increments, alternatively called *completely random measure*, i.e., $\mu(B_1), ..., \mu(B_k)$ are mutually independent if $B_1, ..., B_k$ are disjoint. A random probability measure P can be built from a generalized gamma measure μ according to a standard construction via normalization of completely random measures, which dates back to Kingman (1975). In fact, it can be shown that $\mathbb{P}(0 < \mu(\Theta) =: T < +\infty) = 1$ (see Regazzini, Lijoi and Prünster, 2003), so that $P(\cdot) := \mu(\cdot)/T$ defines a random probability measure on Θ, which will be called *normalized generalized gamma process*, $P \sim NGG(\sigma, \kappa, \omega, P_0)$, with parameters $(\sigma, \kappa, \omega, P_0)$, where $0 \leq \sigma \leq 1, \omega, \kappa \geq 0$. This parametrization is not unique, as the scaling property in Pitman (2003) shows, since $(\sigma, \kappa, \omega, P_0)$ and $(\sigma, s^\sigma \kappa, \omega/s, P_0)$ (for any $s > 0$) yield the same distribution for P.

It is well known that the process P can be represented as

$$P = \sum_{i=1}^{+\infty} P_i \delta_{\tau_i} = \sum_{i=1}^{+\infty} \frac{J_i}{T} \delta_{\tau_i} \tag{2.4}$$

where $P_i := \frac{J_i}{T}$, $(J_i)_i$ are the ranked points of a Poisson process on $\mathbb{R}^+$ with mean density $\rho(ds)$, and $T = \sum_i J_i$; the sequences $(P_i)_{i\geq 1}$ and $(\tau_i)_{i\geq 1}$ in (2.4) are independent, and τ_i are i.i.d. from P_0. Since the NGG process selects discrete distribution with probability one (similarly to the Dirichlet process, as stated in Theorem 2), sampling from P induces an exchangeable random partition on the positive integers; see, for instance, Pitman (2006).

Generally, the finite-dimensional distributions of P are not available in closed analytic form, but the firts two moment measures of P are given (see James, Lijoi and Prünster, 2006) by

$$\begin{aligned} &\mathbb{E}[P(B)] = P_0(B), \\ &Var[P(B)] = P_0(B)(1 - P_0(B))\mathcal{I}(\sigma, \kappa) \end{aligned}$$

where

$$\mathcal{I}(\sigma, \kappa) := (\frac{1}{\sigma} - 1)(\frac{\kappa}{\sigma})^{\frac{1}{\sigma}} e^{\frac{\kappa}{\sigma}} \Gamma(\frac{1}{\sigma}, \frac{\kappa}{\sigma}) = (\frac{1}{\sigma} - 1) \int_1^{+\infty} e^{-\frac{\kappa}{\sigma}(y-1)} y^{-\frac{1}{\sigma}-1} dy$$

and $\Gamma(a, x) := \int_x^{+\infty} e^{-t} t^{a-1} dt$ denotes the incomplete gamma function. The factor $\mathcal{I}(\sigma, \kappa)$ is decreasing as a function of each variable alone, and goes to 0 as $\sigma \to 1$ or $\kappa \to +\infty$, so that $P(B)$ converges in distribution to $P_0(B)$ for any $B \in \mathcal{B}(\Theta)$. On the other hand, it can be shown that

$$\begin{aligned} &\lim_{\sigma \to 0,\ \kappa \to 0} \mathcal{I}(\sigma, \kappa) = 1, \\ &P(B) \xrightarrow{\mathrm{d}} \delta_\tau(B) \end{aligned}$$

where τ is a random variable with distribution P_0. If $\sigma = 0$ and $\kappa > 0$ we recover the Dirichlet process with measure parameter $\alpha(\cdot) = \kappa P_0(\cdot)$.

2.3 Models

In order to deal with r.p.m.'s, various models have been proposed. Among the others, the Product Partition model (PPM) and the Dirichlet process mixture (DPM) model are the most

popular, and result to be very useful in understanding the problem of clustering, allowing also the implementation of the NGG process. At the same time, through the Polya Urn scheme presented in the previous section, they make easier the sampling procedure and the implementation of MCMC techniques. The interesting feature of these models is that they induce a prior distribution on random partitions, thanks to the discreteness property of Dirichlet and NGG's trajectories.

In this section, the two models are presented, with an explanation of the relationship between them.

2.3.1 Product partition models

A clustering of n objects into K_n groups can be represented by a set partition $\pi = \{C_1, ..., C_{K_n}\}$ of a set $C_0 = \{1, ..., n\}$, having the following properties:

1. $C_i \neq \emptyset$, for $i = 1, ...K_n$;
2. $C_i \cap C_j = \emptyset$, for $i \neq j$;
3. $\bigcup_{j=1}^{K_n} C_j = C_0$.

The sets $\{C_1, ..., C_{K_n}\}$ are referred to as partition components.

The number of all the possible partitions of n elements is the Bell number $B(n)$ (see, for instance, Bell, 1934), and represents the cardinality of the space of all possible partitions of the integers $\{1, ..., n\}$, $\mathcal{P}_n$.

The Product Partition model (PPM) is a particular model parameterized by the set partition, introduced for the first time by Hartigan (1990). The likelihood and the prior for the model, based on the random partition of the elements, are described as follows:

$$\begin{aligned} p(\boldsymbol{x}|\pi) &= \prod_{j=1}^{K_n} m(\boldsymbol{x}_{C_j}) \\ p(\pi) &= M \prod_{j=1}^{K_n} h(C_j) \\ M &= 1/\sum_{\pi \in \mathcal{P}_n} \prod_{j=1}^{K_n} h(C_j) \end{aligned} \tag{2.5}$$

where $m(\boldsymbol{x}_{C_j})$ stands for the marginal density of those data belonging to the j-th group identified by the partition, and $h(C_j)$ are called *cohesionfunctions* and describe the prior of the random partition. The prior distribution for the random partition is also called *Exchangeable Partition Probability Function* (or briefly *EPPF*).

By Bayes theorem, the posterior distribution is proportional to a product over partition components:

$$p(\pi|\boldsymbol{x}) \propto p(\boldsymbol{x}|\pi)p(\pi) \propto [\prod_{j=1}^{K_n} m(\boldsymbol{x}_{C_j})][\prod_{j=1}^{K_n} h(C_j)] = \prod_{j=1}^{K_n} m(\boldsymbol{x}_{C_j})h(C_j) \tag{2.6}$$

All the inferences about π are made using this proportionality. In this model, the partition π is the only parameter under consideration; all the other parameters have been integrated out over their priors.

2.3.2 Dirichlet/NGG process mixture models

In parametric cluster analysis, mixture models are often used to describe the partition of the data into K given clusters, represented by K kernel densities. Dirichlet (and then NGG) process mixture model, introduced for the first time by Lo (1984), is a generalization along the nonparametric direction. In this model indeed the data are supposed to be generated from a mixture of kernel densities, indexed by the vector of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, which is, in turn, a finite sample from the nonparametric prior. The model is the following:

$$\begin{aligned} &\boldsymbol{X}_1, ..., \boldsymbol{X}_n|\theta_1, ..., \theta_n \overset{\text{ind.}}{\sim} K(\boldsymbol{x}_i|\theta_i) \\ &\theta_1, ..., \theta_n|P \overset{\text{i.i.d.}}{\sim} P \\ &P \sim q(\alpha(\cdot), P_0) \end{aligned} \tag{2.7}$$

where q stands for a proper nonparametric prior (in this work it will be either the Dirichlet process prior, or the NGG process prior), depending on the measure $\alpha(\cdot)$ and on the mean distribution P_0. It can be equivalently written as $q(aP_0)$, where $a = \alpha(\mathbb{R}^k)$ is the *total mass parameter*.

The tuning parameters $\boldsymbol{\theta}$ are called *latent variables*, and follow a generalized Polya Urn scheme,

as shown in (2.1) and (2.2). This scheme is of great relevance for posterior computations. In fact, in a conjugate model case, it is rather simple to calculate the full conditionals and to implement a posterior MCMC sampling algorithm. Of course, from the description above, the sample $\boldsymbol{\theta}$ could present some values in common; this is the clustering part of the model: the mixture model induces a partition π of the data, based on the values of the parameters of the kernel densities (i.e., the latent variables), though unobserving the specific values of those parameters (from here the *latent* characterization). In order to classify the data, is enough to know which latent variables are equal and which ones are not, avoiding the knowledge of their specific values $\boldsymbol{\phi}$, called vector of unique values.

In the mixture models, to give the latent variables is equivalent to give the two vectors $\boldsymbol{\phi}$ (unique values) and $\boldsymbol{c}$ (configurations), the last one representing the labels of the n observations. From these two vectors it is possible to reconstruct the partition of the data induced by the vector of latent variables.

2.3.3 Relationship between PPMs and mixture models

As mentioned before, a relationship exists between PPM and mixture models, and has been presented for the first time by Quintana and Iglesias (2003), regarding the Dirichlet process prior, but can also be found in the work by Lijoi, Mena and Prünster (2007), dealing with the NGG process.

Consider the nonparametric component of the DPM model (2.7):

$$\theta_1, ..., \theta_n | P \overset{\text{i.i.d.}}{\sim} P$$

$$P \sim \mathcal{D}(aP_0)$$

where $\mathcal{D}(aP_0)$ stands for the Dirichlet process prior, depending on the mass parameter a and the mean distribution P_0. As mentioned before, the Polya Urn representation can be used to obtain the joint law of a sample from the Dirichlet process prior. Substituting the predictive densities in (2.1) into the equation (2.2), then the joint law can be obtained, and results:

$$p(\theta_1, ..., \theta_n) = \prod_{i=1}^{n} \{ \frac{aP_0(\theta_i) + \sum_{j<i} \delta_{\theta_j}(\theta_i)}{a + i - 1} \} \tag{2.8}$$

Thanks to the discreteness of the trajectories of the Dirichlet process (see Theorem 2), the vector $\boldsymbol{\theta}$ of latent variables has non-zero probability to contain equal elements, inducing a partition of the data and the vector itself. Consider now a possible partition $\pi = \{C_1, ..., C_{K_n}\}$ of the sample $\{\theta_1, ..., \theta_n\}$ into K_n clusters, and let $\boldsymbol{\phi} = (\phi_1, ..., \phi_{K_n})$ be the vectors of its unique values (each different value corresponds to a different component in the mixture, i.e. to a different cluster). Similarly to what is done in Lo (1984) and in Quintana and Iglesias (2003), the equation (2.8) can be expressed as:

$$\begin{aligned} p(\theta_1, ..., \theta_n) &= \sum_{\pi \in \mathcal{P}_n} \frac{a^{K_n}}{\prod_{l=1}^{n}(a+l-1)} \prod_{j=1}^{K_n} (|C_j|-1)! P_0(\phi_j) \prod_{i \in C_j} \delta_{\phi_j}(\theta_i) \\ &= \overline{K} \sum_{\pi \in \mathcal{P}_n} \prod_{j=1}^{K_n} h(C_j) p_{C_j}(\theta_{C_j}) \end{aligned} \tag{2.9}$$

where $\overline{K} = \prod_{i=1}^{n}(a+i-1)^{-1}$, $h(C_j) = a(|C_j|-1)! = a\Gamma(|C_j|)$ for $j = 1, ..., K_n$, the vectors $\boldsymbol{\theta}_{C_1}, ..., \boldsymbol{\theta}_{C_{K_n}}$ are independent and $p_{C_j}(\boldsymbol{\theta}_{C_j})$ for $j = 1, ..., K_n$ is defined as the distribution such that all the elements in $\boldsymbol{\theta}_{C_j}$ are identical to the value ϕ_j drawn from P_0.

It is easy to see the PPM description arising from this equivalent equation. In fact, the functions $h(C_j)$, for $j = 1, ..., K_n$, represent the cohesion functions of the prior in (2.5). Hence, it can be argued that the DPM model is equivalent to a PPM with particular cohesion functions, equal to $h(C_j) = a\Gamma(|C_j|)$, for $j = 1, ..., K_n$.

As far as the NGG process mixture model concerns, it can be shown, as presented in Lijoi et al. (2007) and in Gnedin and Pitman (2005), that, with a particular choice of the cohesion functions, a PPM model can be obtained. In particular in this case the exchangeable random partition π has distribution of the form:

$$p(\pi) \propto V_{n,K_n} \prod_{j=1}^{K_n} h(C_j) \tag{2.10}$$

if and only if

$$h(C_j) = [1-\sigma]_{|C_j|-1}, \text{ for } j = 1, ..., K_n \tag{2.11}$$

for some $\sigma \in [0, 1]$ and $V_{n,K_n} = (n - \sigma K_n)V_{n+1,K_n} + V_{n+1,K_n+1}$. Here, $[x]_n$ is the ascending factorial given by $[x]_n = x(x+1)\cdots(x+n-1)$. It is easy to see how, putting $\sigma = 0$, the Dirichlet process is recovered, as mentioned before.

The mixture models generalization made by the PPMs can be expressed through the following model:

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n | \boldsymbol{\phi}, \pi \overset{\text{ind.}}{\sim} \sum_{j=1}^{K_n} K(\boldsymbol{x}_i | \phi_j)$$

$$\phi_1, ..., \phi_k | \pi \overset{\text{i.i.d.}}{\sim} P_0, \text{ for } i = 1, ..., K_n$$

$$\pi \sim p(\pi) \propto \prod_{j=1}^{K_n} h(C_j)$$

The model above can be seen as a mixture model involving the PPM description of the partition, making easier to understand the relationship between the two model descriptions.

As stated before, these two models are used to describe the data and their partitioning, allowing MCMC computations via the Polya Urn scheme, in order to sample from the posterior distribution of the random partition. In this work, both conjugate and non-conjugate models will be used, following the general pattern just presented.

2.4 Posterior choice of one single Bayesian estimate of the random partition

As stated in the previous chapter, a prior over the data partition can be induced using PPMs or mixtures models, allowing a more flexible analysis. All the inferences should be based on the posterior distribution of the random partition. Furthermore, these models are quite easy to handle, especially in the conjugate case, making the MCMC sampling affordable. A different problem is to find one single suitable posterior estimates of the partition of the data, in order to summarize the posterior distribution. Of course, this depends on the random partition's sampling space, which is the set of all possible partitions of n objects. The cardinality of this space is the Bell number, $B(n)$ (see Bell, 1934), and becomes very large as n increases. Notice that this space is not a totally ordered set. Hence, it is hard to treat distributions on it in the usual way.

To solve the problem of choosing a partition of the data, avoiding to deal directly with the analytical form of its posterior density, many methods have been proposed. In this section some of them are presented, in relation with the applications of this work. All the methods

described will be implemented on different data-sets in the following chapters.

Using PPMs and the equation (2.6), the posterior density of the random partition can be evaluated (up to a positive constant) for any given partition. This will be the key point of the first two procedures presented here, introduced in Lau and Green (2007), called the *stochastic search by posterior sampling of partitions* (or briefly **SS**) and the *Bayesian hierarchical clustering procedures* (or briefly **BH**).

Beyond methods based on the evaluation of the posterior density, there is a more "standard" Bayesian estimate, that corresponds to the *loss-function minimization* method. In this approach, a particular function is defined, in order to account the relevance of the partition examined, in terms of the loss resulting from choosing it among all the possibilities. The goal of these techniques is to find the data configuration minimizing the posterior expected value of the proposed loss-function.

2.4.1 Stochastic Search by Posterior Sampling of Partitions

The **SS** method uses the partitions sampled from the posterior distribution of the random partition, in order to find the *Maximum a Posteriori* (or *MAP*) among them. The density evaluation is made by exploiting the formula (2.6) (in a finite set, proportionality is enough to establish which element is the maximum).

Of course, a drawback of the method is related to the sample size, which would unlikely contain all the possible partitions of the objects, making difficult to find the true MAP. Dahl (2009) proposed an algorithm to find the exact MAP, but restricted only to a particular class of models, satisfying two specific conditions; the first one is about the maximality of non-overlapping components of the partition, while the second one assumes that the cohesion functions depend only on the number of elements of each cluster. Another drawback is related to the spread of the posterior density. As said before, the space of all possible partitions is a discrete space, with cardinality $B(n)$, the Bell number, which is considerably large, even for rather small values of n. So, the frequencies over the possible outcomes of the posterior distribution could be very low and therefore very similar from an element of the space to another, making very hard to establish which one is "better". Furthermore, the absence of a total order relation excludes the possibility of ranking the elements with the same mass, or with similar masses, in order to

choose which one is more desirable.

2.4.2 Bayesian Hierarchical Clustering Procedures

The second method involving the evaluation of the posterior density of the random partition is the **BH** method. The method was firstly presented by Heard et al. (2006) into a Bayesian parametric framework, and subsequently discussed by Lau and Green (2007) in a Bayesian nonparametric fashion. The term "hierarchical" refers to hierarchical clustering methods used in heuristic cluster analysis, as briefly presented in Chapter 1. Of course, here the model-based aspect of the analysis is not abandoned, leading to a method which is a combination of model-based and heuristic features. The description of the data follows the product partition model structure, presented in the previous chapter. The heuristic aspect, instead, is represented by the hierarchical procedure used to gather the observations.

The **BH** method proceeds in the same way as the hierarchical one, starting with n singleton and ending with a unique cluster, but with one fundamental difference: the choice of the *similarity* matrix between the data. In order to implement the hierarchical clustering algorithm, a matrix of similarities between clusters is needed, in order to evaluate, at each step, which elements to merge. Thanks to the product partition model, this evaluation between different partitions can be made in terms of posterior densities. At each step of the algorithm, the ratio between the posterior densities of two distinct partitions of the data is evaluated, using the equation (2.6), in order to chose which one is "better" (in this case, most likely a posteriori). The denominator of the ratio is the value of the posterior density at the current configuration (i.e., no cluster is moved), while the numerator is the value at the configuration merging the two selected clusters. This ratio R results:

$$R_{l,h} = \frac{p(\pi^{'}|\boldsymbol{x})}{p(\pi|\boldsymbol{x})} \propto \frac{\prod_{j=1}^{K_n^{'}} m(\boldsymbol{x}_{C_j}) h(C_j)}{\prod_{j=1}^{K_n} m(\boldsymbol{x}_{C_j}) h(C_j)} = \frac{h(C_l \cup C_h) m(\boldsymbol{x}_{C_l \cup C_h})}{h(C_l) h(C_h) m(\boldsymbol{x}_{C_l}) m(\boldsymbol{x}_{C_h})}, \text{ for } l, h = 1, ..., K_n, \quad (2.12)$$

where the $\pi^{'}$ represents the new configuration, that is the one in which the two selected clusters are merged. Of course, $K_n^{'} = K_n - 1$, because two clusters are going to be added. The letters l and h represent the indexes of the clusters to merge (they are identified by their unique values ϕ_l and ϕ_h). The last equivalence follows from the independence of the elements in the clusters,

as stated by the PPM, yielding to a useful simplification of the ratio formula (only the decision terms survive).

With this redefinition of similarity, the hierarchical algorithm can be implemented, following the usual path, starting with n singleton. At each step, the best merging is performed, among all the possible choices, according to the ratios R. The algorithm has a $O(n^3)$ complexity, which is rather high, but manageable. At the end $(n-1)$ different configurations are provided, representing the best choices at every step of the algorithm. The last thing to do is to choose the most likely a posteriori among them, which surely exists because they represent a finite set. In terms of posterior density, many possible configurations are excluded (it is enough to think about the difference between $(n-1)$ and the Bell number $B(n)$). This could be a serious drawback, because some interesting configurations could be missed. At the same time, this method provides an efficient way to find an optimal partition, allowing little complexity in computations.

2.4.3 Loss-Function Minimization methods

A popular method used to identify the optimal partition of the data consists in the minimization of a loss-function representing the misclassification cost generated by choosing a particular partition $\hat{\pi}$ instead of the "true" partition π. In particular, as proposed for the first time by Binder (1978), the optimal partition is the one minimizing the total average cost, i.e. the expected posterior value of the loss-function, computed with respect to the random partition π. We introduce the label vectors $\boldsymbol{c}$ and $\hat{\boldsymbol{c}}$, which, for a given partition, contain the label associated with each observations. The general form of the loss-function is the following:

$$L(\pi, \hat{\pi}) = \sum_{i<j} (a\mathbb{I}_{[c_i=c_j, \hat{c}_i \neq \hat{c}_j]} + b\mathbb{I}_{[c_i \neq c_j, \hat{c}_i = \hat{c}_j]}) \tag{2.13}$$

where π stands for the random partition, and $\hat{\pi}$ for the estimated partition. The label vectors $\boldsymbol{c}$ and $\hat{\boldsymbol{c}}$ can be used to define the corresponding "incidence matrixes", that is matrixes whose entries are binaries indicating whether two elements are in the same cluster or not. Thus, the function above counts how many times a wrong labeling happens, and assigns a different weight to each kind of misclassification. Being (2.13) a function of the random partition, it is a random variable itself. Hence, because of the complexity of the posterior density of the random

partition, it is impossible to provide any evaluation of it without integrating out the random part of the function, i.e. to compute the expected value of this function.

$$l(\hat{\pi}) = \mathbb{E}[L(\pi, \hat{\pi})|\boldsymbol{x}] = \sum_{i<j} (a\mathbb{I}_{[\hat{c}_i \neq \hat{c}_j]}\mathbb{P}(c_i = c_j|\boldsymbol{x}) + b\mathbb{I}_{[\hat{c}_i = \hat{c}_j]}\mathbb{P}(c_i \neq c_j|\boldsymbol{x})) \tag{2.14}$$

where we used the fact that, given a r.v. Y, then $\mathbb{E}[\mathbb{I}_{\{Y \in A\}}] = \mathbb{P}(Y \in A)$. Now l is a function of the proposed partition only, and then can be evaluated. Of course, the quantity $\mathbb{P}(c_i = c_j|\boldsymbol{x})$ is supposed to be known, or , as in this work, estimated from previous sampling.

As reported in Lau and Green (2007), let ρ_{ij} be the *posterior coincidence probabilities* $\mathbb{P}(c_i = c_j|\boldsymbol{x})$ and $\hat{K} = \frac{b}{a+b} \in [0, 1]$, then (2.15) can can be written as:

$$l = a\sum_{i<j} \rho_{ij} - (a+b)\sum_{i<j} \mathbb{I}_{[\hat{c}_i = \hat{c}_j]}(\rho_{ij} - \hat{K}) = a\sum_{i<j} \rho_{ij} - (a+b)f(\hat{\pi})$$

Minimizing $l(\hat{\pi})$ corresponds to maximizing $f(\hat{\pi})$, with respect to $\hat{\pi}$. In order to do so, two samples from the posterior distribution of the random partition are used, one to estimate the quantity ρ_{ij}, which is a mean of incidence matrixes computed via MCMC methods, and the other to evaluate the gain-function and finding the partition maximizing it. Of course, the goodness of the result is affected by the sample size and the choice of the parameter $\hat{K}$, which can be seen as the prior probability of putting together two elements, when they should be separated.

As far as the choice of $\hat{K}$ concerns, it is useful to refer another loss-function proposal by Dahl (2006). This method suggests to choose that partition minimizing the sum of the squares between the element of the incidence matrix and the correspondent posterior coincidence probabilities. In this case, the loss function is of the form $l(\hat{\pi}) = \sum_{i<j}(\mathbb{I}_{[\hat{c}_i = \hat{c}_j]} - \rho_{ij})^2$. As shown in Fritsch and Ickstadt (2009), letting $\hat{K} = 0.5$ in (2.15) is the same as minimizing the function proposed by Dahl (2006). In this work, $\hat{K}$ will be often fixed at this value.

2.5 A new class of Bayesian estimates for the random partition

In the next chapters, the methods discussed so far will be implemented, using different datasets. Furthermore, extensions of such methods and algorithms towards a generalizing direction are proposed. It is worth to specify that all of these new features are not proposed in the references cited in this work, and therefore they represent the innovating part of the project.

First of all, in order to to robustify the analysis, a higher level of hierarchy is considered. In this case, prior densities over the mass parameter a of the Dirichlet process are tested. Secondly, not only the Dirichlet process prior will be used, but also one of its generalizations, the NGG process prior, allowing more flexibility to the model.

Another important extension is made in relation with the choice of the loss-function to minimize. Of course, the one presented by Binder (1978) and Lau and Green (2007) will be considered (very often in the specific case of $\hat{K} = 0.5$, recalling Dahl (2006)). The classical specification of the loss-function involves a decision rule to establish whether two objects are in the same cluster or not. This rule is represented by the equivalence relation "=", which means that two elements are in the same group if and only if their latent variables $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are equal (or their labels c_i and c_j). In fact, in order to define a partition of the observations, an equivalence relation is needed, so that the induced partition can be used. The extension is made in this direction: a new equivalent relation is found, based on the distance between two elements, and not only on their peculiar values, relaxing the equality constraint. Of course, this new relation must include the old one, in order to see it as a generalization. In terms of distances, it can be said that $\boldsymbol{x}_i$ will be in the same cluster of $\boldsymbol{x}_j$ if and only if $d_{ij} = d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \leq \epsilon$, where $d(\cdot, \cdot)$ is a proper distance between the two elements, and ϵ is a threshold parameter, used to tune the flexibility of the new partition, defining the ϵ-neighborhood of a point p as $N_\epsilon(p) = \{q \in \boldsymbol{O} | d(p, q) \leq \epsilon\}$, being $\boldsymbol{O}$ the set of objects considered. The problem in this case is that this relation is not an equivalent relation (the transitivity is not holding), and then it cannot be used to define a new partition method, so it must be modified in order to have an equivalent relation. To do so, the connectivity relation defined by Ester et al. (1996) will be used, involving a new parameter N, which represents the minimum number of elements required to define a group

cluster. This relation is exploited to define the DBSCAN algorithm as in Ester et al. (1996), presented briefly in the first chapter, and can be achieved using three definitions:

Definition 7. ***(Directly density-reachable)***

A point p is directly density-reachable *from a point q with respect to ϵ and N if*

1. $p \in N_\epsilon(q)$ *and*
2. $|N_\epsilon(q)| \geq N$ *(core point condition).*

Of course, the word *density* here is not used with a probabilistic accent, but only in the sense of number of elements nearby a selected point. The second condition is called "core object condition". If this condition holds for an object p, then p is called "core object".

Definition 8. ***(Density-reachable)***

A point p is density-reachable *from a point q w.r.t. ϵ and N if there is a chain of points $p_1, ..., p_n$ with $p_1 = q$ and, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i.*

Density reachability is the transitive hull of direct density-reachability. This relation is not symmetric in general (only core objects can be mutually density-reachable).

Definition 9. ***(Density-connected)***

A point p is density-connected *to a point q w.r.t. ϵ and N if there is a point s such that both, p and q are density-reachable from s w.r.t. ϵ and N.*

A final remark is about the parameter N. This is useful to find cluster of arbitrary shape, and to calibrate the spatial density of the clusters, with respect to the threshold ϵ. In our case, though, even a singleton would be considered a cluster, because we are working with the unique latent vectors, which possibly identify groups of observations. So, it is enough to put $N = 1$. In this special case density-reachable and density-connected definitions are equivalent. Furthermore, it is important to notice that, in the case of $N > 1$, the density connectivity defined above does not identify the partitions uniquely, possibly leading to situations of uncertainty about some data (it is enough to consider an element equally distant from two different core points of two different clusters). In fact, the density-connection is not a symmetric relation if $N > 1$.

Now the new loss-function can be defined. Substituting the equivalence relation by $d_{ij}(\epsilon)$, where $d_{ij}(\epsilon) = 1$ if and only if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are density-connected with respect to ϵ (and $N = 1$), and 0 otherwise, we have:

$$L(\pi, \hat{\pi}) = \sum_{i<j} (a\mathbb{I}_{[d_{ij}(\epsilon)=1, \hat{d}_{ij}(\epsilon)=0]} + b\mathbb{I}_{[d_{ij}(\epsilon)=0, \hat{d}_{ij}(\epsilon)=1]}),$$

where the hat represents the deterministic part of the function. The mean value is:

$$l(\hat{\pi}) = \mathbb{E}[L(\pi, \hat{\pi})|\boldsymbol{x}] = \sum_{i<j} (a\mathbb{I}_{[\hat{d}_{ij}(\epsilon)=0]}\mathbb{P}(d_{ij}(\epsilon) = 1|\boldsymbol{x}) + b\mathbb{I}_{[\hat{d}_{ij}(\epsilon)=1]}\mathbb{P}(d_{ij}(\epsilon) = 0|\boldsymbol{x})), \tag{2.15}$$

and, reasoning as before:

$$l(\hat{\pi}) = a\sum_{i<j} \rho_{ij} - (a+b)\sum_{i<j} \mathbb{I}_{[\hat{d}_{ij}(\epsilon)=1]}(\rho_{ij} - \hat{K}) = a\sum_{i<j} \rho_{ij} - (a+b)f(\hat{\pi}),$$

where $\rho_{ij} = \mathbb{P}(d_{ij}(\epsilon) = 1|\boldsymbol{x})$.

With this new equivalent relation, there is more flexibility in the choice of the partition, allowing elements whose values are close to be in the same cluster.

Unfortunately, as happens in heuristic methods, the choice of the distance is of a great relevance for the resulting estimates. In this work, various distances will be considered, such as the Euclidean distance, Kullback-Leibler I-divergence (see, Csiszar 1975) and L^2 metric. The first is used to evaluate the distances between different vectors of unique values of the latent variables, while the others evaluate the distances between correspondent kernel densities. As far as the Kullback-Leibler I-divergence concerns, the distance between two observed latent variables $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ results:

$$KL(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) = \int k(\boldsymbol{x}|\theta_i) \log \frac{k(\boldsymbol{x}|\theta_i)}{k(\boldsymbol{x}|\theta_j)} d\boldsymbol{x} \tag{2.16}$$

where $k(\cdot|\boldsymbol{\theta}_i)$ and $k(\cdot|\boldsymbol{\theta}_j)$ represent the two kernel densities associated with the two latent variables $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. Observe that $KL(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) \geq 0$, and $KL(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) = 0$ if and only if $k(\cdot|\boldsymbol{\theta}_i) = k(\cdot|\boldsymbol{\theta}_j)$, that is if and only if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ a.s. (i.e., they are in the same cluster according to the old equivalence relation). Furthermore, $KL(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j)$ is not a symmetric function. In order to define a symmetric function, the sum $KL_{ij} = KL(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) + KL(\boldsymbol{\theta}_j||\boldsymbol{\theta}_i)$ will be considered in this work. Using the Kullback-Leibler I-divergence, elements belonging to a distribution which is "close"

to another one will be considered in the same cluster. A similar result is achieved using other distances, such as the L^2 distance or the Mahalanobis distance, which are defined as follows:

$$d_{L^2}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \sqrt{\int (k(\boldsymbol{x}|\boldsymbol{\theta}_i) - k(\boldsymbol{x}|\boldsymbol{\theta}_j))^2 d\boldsymbol{x}}, \tag{2.17}$$

$$d_M(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{2.18}$$

L^2 and Mahalanobis distances will be used in a bivariate example, involving Gaussian kernel densities (for the last one a symmetric version is used). Furthermore, substituting $k(\boldsymbol{x}|\boldsymbol{\theta}_i)$ and $k(\boldsymbol{x}|\boldsymbol{\theta}_j)$ in 2.17 with their square roots, we obtain the Hellinger distance, which will be used in the bivariate example, too. An interested property of this last distance is the it ranges in the finite set [0,1], making it a feasible metric to express and understand the results. Kullback-Leibler I-divergence will be used in all the examples.

Chapter 3

Galaxy Data

In this chapter, the models and procedures previously discussed are applied to the Galaxy dataset, popular in literature (see, for instance, Roeder, 1990). These data are observed velocities of $n = 82$ different galaxies, belonging to six well-separated conic sections of space. Values are expressed in [Km/s], scaled by a factor of 10^{-3}, in order to deal with thousands of kilometers. The error from sampling the velocities is estimated to be less than 50 Km/s.
The model chosen for the data is a DPM model, involving Gaussian kernel densities. As far as the mean parameter of the nonparametric prior P_0 concerns, a normal inverse-gamma distribution is taken into account. Briefly:

$$\begin{aligned}
&X_1, ..., X_n | \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_n} \overset{\text{ind.}}{\sim} N(x_i | \boldsymbol{\theta_i}), \boldsymbol{\theta_i} = (\mu_i, \sigma_i^2) \\
&\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_n} | P \overset{\text{i.i.d.}}{\sim} P \\
&P \sim \mathcal{D}(a \cdot P_0) \text{ or } NGG(\sigma, \kappa, P_0) \\
&(a \sim Gamma(\gamma_1, \gamma_2)) \\
&P_0(d\mu, d\sigma^2) = N(d\mu | m_0, \frac{\sigma^2}{k_0}) IG(d\sigma^2 | \nu_1, \Psi_1)
\end{aligned} \tag{3.1}$$

where $N(\mu, \sigma^2)$ represents the univariate normal distribution with mean μ and variance σ^2; $Gamma(\gamma_1, \gamma_2)$ stands for the univariate gamma distribution having mean $\frac{\gamma_1}{\gamma_2}$ and variance $\frac{\gamma_1}{\gamma_2^2}$ and $IG(\nu_1, \Psi_1)$ stands for the univariate inverse-gamma distributions with mean $\frac{\Psi_1}{\nu_1 - 1}$ (for $\nu_1 > 1$) and variance $\frac{\Psi_1^2}{(\nu_1-1)^2(\nu_1-2)}$ (for $\nu_1 > 2$).
Notice that a prior on the mass parameter a can be given. Of course, we will consider this case when the prior is a Dirichlet process. For the NGG case no priors are imposed on its

hyperparameters. We considered the following cases:

- Dirichlet process - fixed mass parameter a
- Dirichlet process - random mass parameter: $a \sim Gamma(\gamma_1, \gamma_2)$
- NGG process - fixed parameters σ and κ

Notice that each one of the three nonparametric priors taken into account in this work belongs to the same nonparametric family, that is the one of the NGG process prior. As mentioned in Section 2.2.2, when $\sigma = 0$, the NGG process prior recovers the Dirichlet process prior with fixed mass parameter $a = \kappa$. Similar considerations can be made when dealing with a random mass parameter. For each situation, the choice of the values of the hyperparameters is based on the value of the prior mean of the number of clusters, denoted here by K_n. In particular, we opted for those choice of hyperparameters leading to 1, 3 or 10 prior expected clusters. Finally, we choose three sets of hyperparameters for the mean distribution P_0, $(m_0, k_0, \nu_1, \Psi_1)$. The first set is the one proposed by Escobar and West (1995), which is $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$, while for the third and second second set we tried to incorporate our prior beliefs. We opted for a prior distribution P_0 imposing large variability on μ and a small variability on σ^2, in order to represent the prior information that many different clusters are necessary to explain the observations. In other words, we look for values of the hyperparameters such that $Var[\mu]$ results large, while $\mathbb{E}[\sigma^2]$ and $Var[\sigma^2]$ turn out to be small, leading to a combination of Gaussian kernels with many different locations and stretched shapes. Keeping in mind the relationship stated in the model (3.1) between μ and σ, we have:

$$\begin{aligned}
&\mathbb{E}[\sigma^2] = \frac{\Psi_1}{(\nu_1 - 1)} \quad \text{for } \nu_1 > 1 \\
&Var[\sigma^2] = \frac{\Psi_1^2}{(\nu_1 - 1)^2(\nu_1 - 2)} \quad \text{for } \nu_1 > 2 \\
&Var[\mu] = \mathbb{E}[Var[\mu|\sigma^2]] + Var[\mathbb{E}[\mu|\sigma^2]] = \mathbb{E}[\frac{\sigma^2}{k_0}] + Var[m_0] = \frac{\Psi_1}{(\nu_1 - 1)k_0} \quad \text{for } \nu_1 > 1
\end{aligned} \tag{3.2}$$

where the last equality holds thanks to the variance decomposition formula. Notice that the hyperparameters from Escobar and West (1995) lead to a prior distribution for σ^2 with infinite variance ($\nu_1 = 2$). We fixed $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$ for the second set of hyperparameters, which implies $\mathbb{E}[\sigma^2] \approx 1$, $Var[\sigma^2] \approx 0.06$ and $Var[\mu]$ larger than 1000, and

$(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$ for the third set of hyperparameters (this is equivalent to fix both the prior mean and variance of σ^2 equal to 0.1, and the prior variance of μ to 100). For both these two sets of hyperparameters, we have that, a priori, the distribution of μ is rather non-informative, while the distribution of σ^2 is approximately degenerate on a constant value (the variance is rather low). This assumption translates our prior belief that many clusters are necessary in order to explain the observations. In this case, we let it fixed at the null value, for simplicity of computations. A final remark concerns the correlation between the random variables μ and σ^2, which are the two components of the latent variable vector $\boldsymbol{\theta}$. If we compute $Cov(\mu, \sigma^2)$, we obtain:

$$Cov(\mu, \sigma^2) = \mathbb{E}[\mu \cdot \sigma^2] - \mathbb{E}[\mu]\mathbb{E}[\sigma^2] = \mathbb{E}[\mathbb{E}[\mu \cdot \sigma^2|\sigma^2]] - \mathbb{E}[\mathbb{E}[\mu|\sigma^2]]\mathbb{E}[\sigma^2] =$$
$$= \mathbb{E}[\sigma^2\mathbb{E}[\mu|\sigma^2]] - \mathbb{E}[\mathbb{E}[\mu|\sigma^2]]\mathbb{E}[\sigma^2] = \mathbb{E}[\sigma^2 \cdot m_0] - m_0\mathbb{E}[\sigma^2] = 0,$$

i.e. μ and σ^2 have zero prior correlation, but they are not a priori independent.

3.1 Dirichlet process with fixed mass parameter

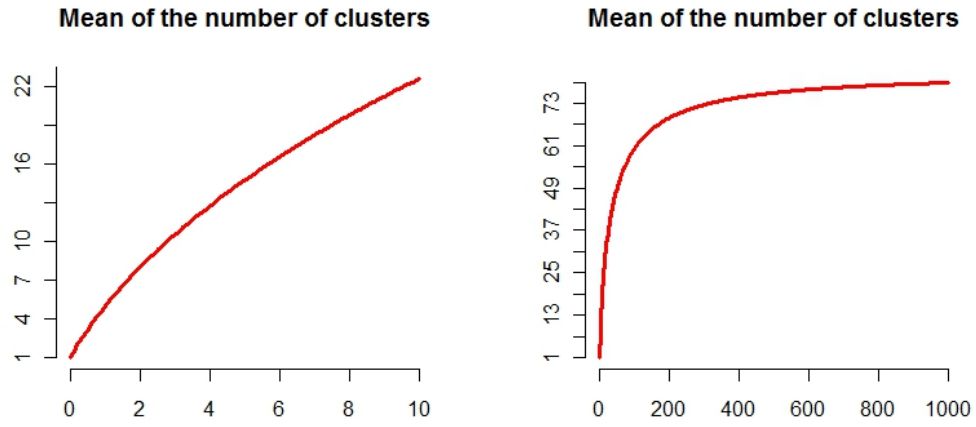
In the case of Dirichlet process prior with fixed mass parameter, an analytical expression of the density of the number of clusters is known (see, Antoniak 1974), and it is reported in the formula (2.3). From this result, it is possible to evaluate the mean value of the number of clusters, which results:

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{a}{a + i - 1}$$

Figure 3.1 displays the plot of the prior expected number of clusters, as a function of the mass parameter a on the x-axis, while Table 3.1 reports the values of the mass parameter which will be used in the next analysis.

$\mathbb{E}[K_n]$	a
1	0.001
3	0.455
10	2.755

Table 3.1: Values of the mass parameter a of the Dirichlet process prior, and corresponding expected number of clusters.



(a) $a \in (0, 1000)$ (b) $a \in (0, 10)$

Figure 3.1: Prior expected number of clusters K_n, varying the value of the mass parameter a of the Dirichlet process prior.

3.2 Dirichlet process with random mass parameter

Here, the mass parameter a is taken as a random variable, thus a probability distribution has to be chosen for it. As often proposed in literature (see, for instance, Escobar and West, 1995), we opted for the "standard" choice of a univariate gamma distribution $Gamma(\gamma_1, \gamma_2)$. Arguing as before, values for the hyperparameters γ_1 and γ_2 have to be found, in order to have 1, 3 or 10 expected number of clusters, which can be written as follows:

$$\mathbb{E}[K_n] = \mathbb{E}[\mathbb{E}[K_n|a]] = \mathbb{E}[\sum_{i=1}^{n} \frac{a}{a+i-1}]$$

where the first equality holds for the double mean property of expected values, and the last mean is evaluated with respect to the random variable a. This calculus can be executed numerically, via a Monte Carlo simulation. Alternatively, the expected number of clusters can be written in the following way:

$$\mathbb{E}[K_n] = \sum_{k=1}^{n} k \cdot \mathbb{P}(K_n = k) = \sum_{k=1}^{n} k \cdot \int_{\mathbb{R}^+} \mathbb{P}(K_n = k|a)p(da) \tag{3.3}$$

where the expression of $\mathbb{P}(K_n = k|a)$ is known form the formula (2.3), and $p(da)$ stands for the prior distribution of a. The inner integrals can be calculated numerically using adaptive quadrature methods, with a very low effort and high precision, and then summed to give the

estimated expected value. Therefore, the value of the mean of the number of clusters depends on two additional parameters, γ_1 and γ_2. In Figure 3.2(a) the evolution of the prior expected number of clusters can be observed, together with its contour plot (lighter colours mean higher values) in Figure 3.2(b). Of course, there will be infinite couples $(\gamma_1, \gamma_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ such that $\mathbb{E}[K_n]$= 1, 3 or 10. We choose these values according to the corresponding mean and variance of the prior distribution of the mass parameter. For the case of one expected cluster, a prior with a low mean and variance has been chosen, in order to reflect the same choice made in the case of fixed mass parameter a (see values in Table (3.1)). In fact, in that case, the value of the mass parameter is very low, leading to an approximatively parametric model, that is:

$$\begin{aligned} &X_1, ..., X_n | \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_n} \overset{\text{ind.}}{\sim} N(x_i | \boldsymbol{\theta_i}), \boldsymbol{\theta_i} = (\mu_i, \sigma_i^2) \\ &\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_n} | X_0 \overset{\text{i.i.d.}}{\sim} \delta_{X_0} \\ &X_0 \sim P_0 \\ &P_0(d\mu, d\sigma^2) = N(d\mu | m_0, \frac{\sigma^2}{k_0}) IG(d\sigma^2 | \nu_1, \Psi_1) \end{aligned} \tag{3.4}$$

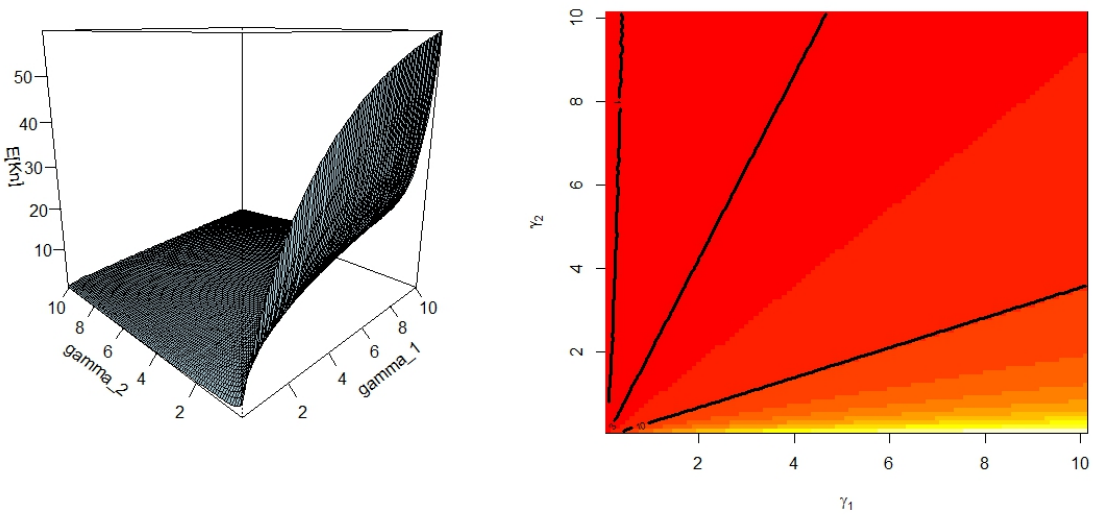
where $P_0(\cdot)$ is the mean distribution of the Dirichlet process prior. In the same way we want to preserve the condition leading to a larger prior expected number of clusters (10 in this case), imposing a quite large variance on the mass parameter (third line of Table 3.2). Finally, in Table 3.2, the ultimate choices of the hyperparameters is summarized. The second couple of hyperparameters has been assigned following the work of Escobar and West (1995).

$\mathbb{E}[K_n]$	γ_1	γ_2	$\mathbb{E}[a]$	$Var(a)$
1	2	100	1/50	2e-04
3	2	4	0.5	0.125
10	3	1	3	3

Table 3.2: Hyperparameters of the random mass parameter of the Dirichlet process prior and corresponding prior expected number of clusters. The last two columns report the values of the prior mean and variance of the mass parameter a.

3.3 NGG process with fixed parameters

As far as the NGG process prior concerns, integrals (3.3) for the prior expectations of K_n must be numerically evaluated. To choose the values of the hyperparameters, we considered



(a) (b)

Figure 3.2: (a) Surface and (b) contour plot of the mean values of the number of clusters K_n, varying the value of the hyperparameters γ_1 and γ_2 of the mass parameter distribution: $a \sim Gamma(\gamma_1, \gamma_2)$. In (b) the black lines represent those couples (γ_1, γ_2) for which the mean of the number of clusters is equal to 1, 3 and 10 respectively.

the lower bound existing for the expected number of clusters for a NGG process prior with hyperparameters σ and κ, that is $\mathbb{E}_{0,\sigma}[K_n] \leq \mathbb{E}_{\kappa,\sigma}[K_n]$. At the same time, large values of σ lead to very high computational efforts. So, we choose not to exceed the value $\sigma = 0.5$, and to fix it at 0.01, 0.25 and 0.5 (arbitrarily). Finally, we found those values for κ that allow us to have the prior expected number of clusters equal to 1, 3 or 10. Table 3.3 reports the final choices for the hyperparameters.

$\mathbb{E}[K_n]$	σ	κ
1	0.01	0.01
3	0.25	0.05
10	0.5	0.01

Table 3.3: Values of the hyperparameters σ and κ of the NGG process prior, and corresponding prior expected number of clusters.

After a deep analysis of the results of applying the methods here proposed to the Galaxy dataset, we decided to show only those concerning the prior choice of $\mathbb{E}[K_n] = 3$. In fact, all the methods proposed in this work turned out to be very robust with respect to the choice of the hyperparameters of the nonparametric priors, giving very similar outcomes. In particular, in the case of Dirichlet process prior with random mass parameter, this choice of $\mathbb{E}[K_n]$ recalls

the one made in Escobar and West (1995), often leading to a partition of the observations into three different clusters. Furthermore, the result of three clusters for Galaxy data is supported in Roeder (1990), as a lower bound for the number of modes of the estimated distribution.

3.4 SS method

In this section, the stochastic search method (SS) will be applied to the Galaxy data. We recall that this method searches for the best partition among those sampled by a MCMC algorithm, evaluating the corresponding proportional posterior densities using (2.6). To do so, we need to compute the values of the cohesion functions $h(C_j)$ and of the marginal densities $m(\boldsymbol{x}_{C_j})$, for $j = 1, ..., K_n$. The cohesion function depends on the hyperparameters of the nonparametric prior, while the densities $m(\boldsymbol{x}_{C_j})$, for $j = 1, ..., K_n$, only depend on the hyperparameters of the mean distribution P_0, that is $(m_0, k_0, \nu_1, \Psi_1)$. An analytical expression of the marginal densities is as follows:

$$m(\boldsymbol{x}_{C_j}) = \int_{\Theta} \prod_{i \in C_j} K(x_i|\phi_j) P_0(d\phi_j) = \int_{\mathbb{R}} \int_{\mathbb{R}^+} \prod_{i \in C_j} N(x_i|\mu_j, \sigma_j^2) N(d\mu_j|m_0, \frac{\sigma_j^2}{k_0}) IG(d\sigma_j^2|\nu_1, \Psi_1) d\mu_j d\sigma_j^2 \tag{3.5}$$

This marginalization is computed with respect to the Dirichlet process prior throughout the latent variables (or better, their unique values), preserving the dependency on the number of clusters. The integral above is very common in literature, being the marginal distribution of a normal inverse-gamma conjugate model, leading to a generalized n_j-dimensional Student's t-distribution density, being $n_j = |C_j|$. In the next sections we will give expressions of the cohesion functions, for particular specifications of the nonparametric priors involved in the analysis.

3.4.1 Dirichlet process with fixed mass parameter

As mentioned before (see Section 2.3.3), a strong relationship between PPMs and DPM models exists. Thanks to this relationship, it is possible to characterize the nonparametric process priors via the definition of their corresponding cohesion functions $h(C_j)$, for $j = 1, ..., K_n$. In the case of the Dirichlet process prior with fixed mass parameter a we have $h(C_j) = a \cdot \Gamma(|C_j|)$,

for $j = 1, ..., K_n$. Notice that these functions only depend on the cardinality of each cluster, not involving the values of the observations.

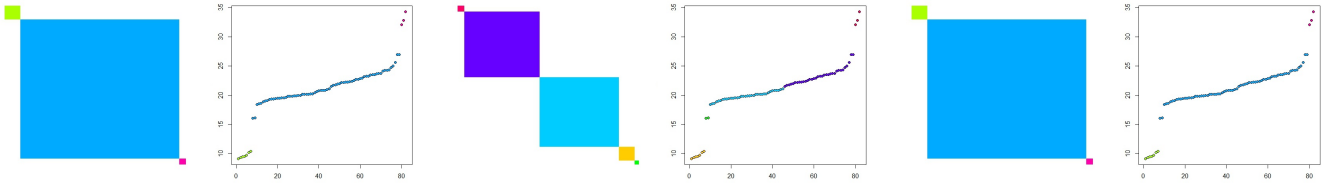
We are now able to perform the SS method for clustering Galaxy data, in the case of Dirichlet process prior with fixed mass parameter. The estimated partitions are shown in Figure 3.3, through their incidence matrixes and data representation. An incidence matrix M is a particular matrix whose entries $[m_{ij}]_{i,j=1,...,n}$ are binaries indicating whether two observations are clustered together ($m_{ij} = 1$) or not ($m_{ij} = 0$). Given a particular partition $\hat{\pi}$, this matrix can be calculated using the correspondent vector of labels for the observations. In order to obtain plots such as the ones in Figure 3.3, we have to re-order the incidence matrixes into clusters, and assign to them different colours (only the elements corresponding to positive entries are shown). This method represents an efficient way to display a partition clearly, showing how many clusters are identified and their dimensions. Unfortunately, this kind of plot misses the information about the assignment of the observations (i.e., which elements are grouped together and which not). To avoid this problem, incidence matrixes will be always presented accompanied by a plot of the Galaxy dataset, with the elements coloured according to the partition taken into account. This kind of figures will be often used during all this work, in order to show the estimates found by the methods proposed.

As mentioned before, the SS method turns out to be very robust with respect to the choice of the value of the mass parameter a. For this reason, we only show results concerning $a = 0.455$, leading to 3 prior expected number of clusters. The hyperparameters $(m_0, k_0, \nu_1, \Psi_1)$ were chosen as specified at the beginning of this chapter.

As can be easily inferred from Figure 3.3, the SS method is not very robust with respect to the choice of the set of hyperparameters of P; in particular, the estimated partition in Figure 3.3 (c) and (d) shows lack of robustness. This could be explained since the second set of hyperparameters leads to a very large variance a priori for μ, which would explain the highest number of clusters found by the method.

3.4.2 Dirichlet process with random mass parameter

When a prior is assumed for the total mass parameter a of the Dirichlet process prior, computations of the posterior density of the random partition becomes more difficult. In fact, starting



(a) (b) (c) (d) (e) (f)

Figure 3.3: Application of the SS method to the Galaxy dataset under Dirichlet process prior with fixed mass parameter ($\mathbb{E}(K_n) = 3$, $a = 0.455$). (a) and (b): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$; (c) and (d): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$; (e) and (f): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$.

from the formula (2.6), we have:

$$p(\pi|\boldsymbol{x}) \propto p(\boldsymbol{x}|\pi)p(\pi) = \int_{(0,+\infty)} p(\boldsymbol{x}|\pi, a)p(\pi|a)p(da)$$

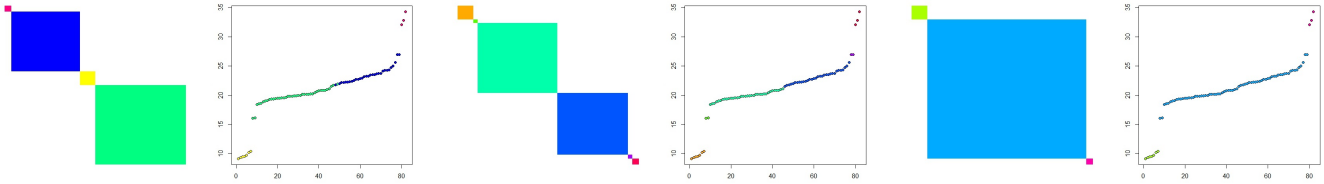
Of course, $p(\boldsymbol{x}|\pi, a) = \prod_{j=1}^{K_n} m(\boldsymbol{x}_{C_j})$ in (3.5), not directly involving the mass parameter a. Therefore, the integrand above contains only of the prior term of the mass parameter a, and the term $p(\pi|a)$. This last term is the $EPPF$ (Exchangeable Partition Probability Function) of the corresponding Dirichlet process prior with fixed mass parameter a, that is $p(\pi|a) \propto \prod_{j=1}^{K_n} h(C_j)$. So, the integral is proportional to:

$$\begin{aligned} &\int_{(0,+\infty)} p(\boldsymbol{x}|\pi, a)p(\pi|a)p(da) \propto \prod_{j=1}^{K_n} m(\boldsymbol{x}_{C_j}) \int_{(0,+\infty)} \prod_{j=1}^{K_n} \frac{a}{[a]_n} \Gamma(|C_j|) Gamma(da|\gamma_1, \gamma_2) = \\ &= \prod_{j=1}^{K_n} \Gamma(|C_j|) m(\boldsymbol{x}_{C_j}) \int_{(0,+\infty)} \frac{a^{K_n}}{[a]_n} Gamma(da|\gamma_1, \gamma_2) = \prod_{j=1}^{K_n} \Gamma(|C_j|) m(\boldsymbol{x}_{C_j}) \mathbb{E}_{p(a)} \left[\frac{a^{K_n}}{[a]_n} \right] \end{aligned} \tag{3.6}$$

where $[a]_n = \frac{\Gamma(a+n)}{\Gamma(a+1)}$ is the rising factorial (here $\Gamma(\cdot)$ denotes the Euler's gamma function). The computation of this integral is the same as the one necessary to compute the mean number of clusters for the same DPM model, except for the Stirling number term. Once again, numerical methods must be applied in order to evaluate this quantities.

In Figure 3.4 the results of the estimated partition provided by the SS method for $(\gamma_1, \gamma_2) = (2, 4)$ are shown (this choice of the hyperparameters of the Dirichlet process prior recalls the one made in Escobar and West (1995), leading to three prior expected clusters). Once again, the choice of the set of hyperparameters for the mean distribution P_0 strongly influences the results. Additionally, in this case, the choice of the first set of hyperparameters (i.e., the one

proposed by Escobar and West, 1995) does not give the same estimate found in the case of fixed mass parameter, leading to a partition with a larger number of clusters. This difference is due to the choice of the hyperparameters for the prior distribution of the mass parameter a. In the case of fixed mass parameter, the prior expected number of clusters was equal to 3 when $a = 0.455$; with a prior distribution on a we have to set the values for the two hyperparameters γ_1 and γ_2, and we choose only one of the infinite couples available, respecting the fact that $\mathbb{E}[K_n] = 3$ (referring to Table 3.2, line two). This choice leads to a prior distribution with mean $\mathbb{E}[a] = 0.5$, which is close enough to the value of 0.455, and a variance of 0.125, that is probably too high to emulate the situation of fixed mass parameter.



(a) (b) (c) (d) (e) (f)

Figure 3.4: Application of the SS method to the Galaxy dataset under Dirichlet process prior with random mass parameter ($\mathbb{E}(K_n) = 3$, $(\gamma_1, \gamma_2) = (2, 4)$). (a) and (b): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$; (c) and (d): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$; (e) and (f): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$.

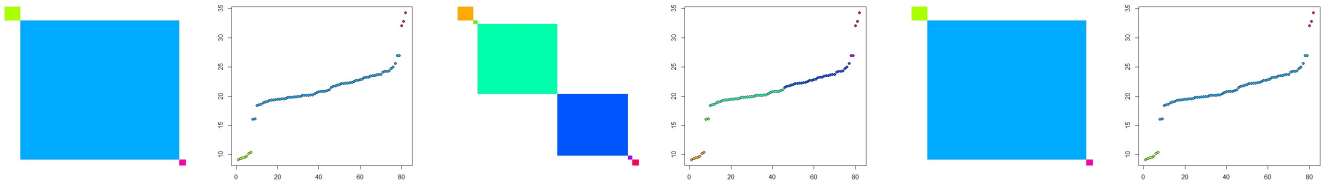
3.4.3 NGG process with fixed hyperparameters

As a final example, we consider the NGG process prior, with fixed hyperparameters σ and κ, and show the resulting estimated partitions. In order to evaluate the posterior densities, we need to know the values of the weights V_{nk} in the formulas (2.10) and (2.11). For this reason, the evaluation of corresponding weights A_{nk} is performed, and then exploited the following relationship:

$$V_{nk} = \frac{e^{\kappa/\sigma}\sigma^{k-1}}{\Gamma(n)} \exp(A_{nk}) \tag{3.7}$$

where k represents the current value of K_n. The logarithm of the weights is computed in order to avoid computational problems. A very complex relation exists between the coefficients of the cohesion functions for a NGG process prior. To compute this values, we refer to the algorithms used by Argiento et al. (2010).

Figure 3.5 shows the results. The estimates resulting for the case of NGG process prior turn out to be very similar to the ones in the case of Dirichlet process prior with fixed mass parameter, though the value of σ is set to 0.25, which is not very low (we recall that, when $\sigma = 0$, a NGG process prior recovers the Dirichlet process prior with fixed mass parameter $a = \kappa$). In this case, in order to have $\mathbb{E}[K_n] = 1, 3, 10$, we found low values for the hyperparameter κ, leading to an approximately parametric model (see, for example, the model in (3.4)). This could explain why, even with rather high values of σ, we still have results similar to the Dirichlet process prior.



(a) (b) (c) (d) (e) (f)

Figure 3.5: Application of the SS method to the Galaxy dataset under NGG process prior with fixed hyperparameters ($\mathbb{E}(K_n) = 3$, $(\sigma, \kappa) = (0.25, 0.05)$). (a) and (b): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$; (c) and (d): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$; (e) and (f): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$.

3.4.4 Final considerations

A final remark about the application of the SS method to the Galaxy data concerns the choice of the estimate for the random partition. The method has proved to be very robust with respect to the choice of the mass parameter, while slightly sensitive to the choice of the nonparametric prior, mostly due to the difficulty of expressing the same prior beliefs in different nonparametric environments. Additionally, the choice of the set of hyperparameter for P_0 strongly influences the results, sometimes clearly overestimating the number of clusters. We believe that the 3-clusters partition is the one better representing the "real" classification of the data, supported by the graphical characterization of the observations (sub-figures (b), (d) and (f) next to the incidence matrixes) and by prior belief. Finally, we recall that the SS method only explores those partitions sampled by the Gibbs sampler algorithm, yielding to unreliable estimates due to the dimension of the MCMC sample analyzed. We cannot exclude that greater MCMC samples would have given different results, though here we ran MCMC samples of 5.000 and

10.000 iterations, with a thinning of 15 and a burn-in of 50.000, not encountering relevant variations in the results. Furthermore, in terms of posterior density, two completely different partitions could have similar weights, leading to a not efficient way to choose the partition.

3.5 BH method

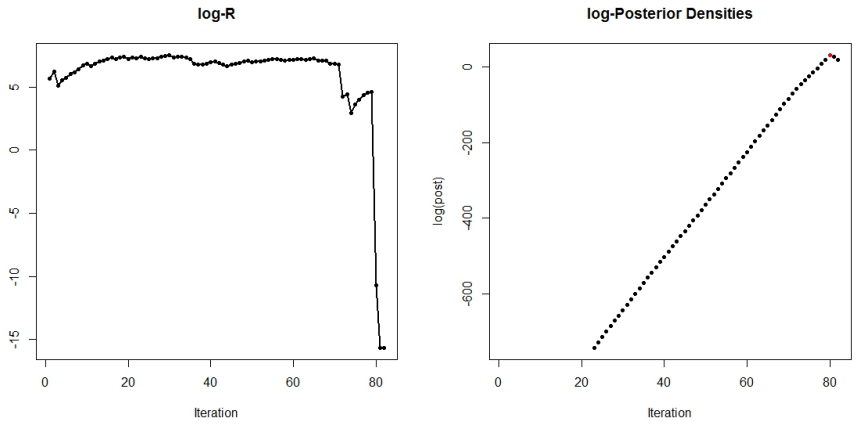
To apply the BH method to the Galaxy data, we need to compute the ratios in formula (2.12). To do so, we need to know the expression of the posterior densities we want to compare at each step (or better, their proportional form). Keeping in mind (2.6), together with the results of the previous chapter, we are able to give expressions of these ratios, for each nonparametric prior we consider. The equation of such ratio between two different partitions is:

$$R_{i,j} = \frac{p(\pi_i|\boldsymbol{x})}{p(\pi_j|\boldsymbol{x})} \propto \frac{p(\pi_i)}{p(\pi_j)} \underbrace{\frac{p(\boldsymbol{x}|\pi_i)}{p(\boldsymbol{x}|\pi_j)}}_{BF_{ij}}$$

It is important to point out that the ratio used to perform the algorithm does not represent a Bayes Factor, though its expression could look very similar. This equation shows that the ratio would equal the Bayes Factor if and only if the prior over the possible configurations of the data (i.e., partitions) was uniformly distributed. Such a prior would clearly be counterintuitive, assigning the same weight to all possible configurations, and will not be taken into account here. Furthermore, Bayes Factor compares two hypothesis which are disjointed and exhaustive, while the ratio used in the BH algorithm compares two data configurations for which clearly these characteristics do not hold.

As mentioned in Section 2.4.2, at each step the BH algorithm looks for the two sub-clusters which maximize the ratio and then joins them together. A plot of the maximal ratios observed during a run of the algorithm is showed in (3.6)(a). At the end of the algorithm, n different partitions are provided, and we choose the one whose posterior density is the highest (up to a positive constant). See figure (3.6)(b) for an example.

We performed a vast analysis of this clustering method, for different choices of priors and hyperparameters, and the BH method for the Galaxy dataset leads always to the same estimate. In particular, we implemented the method for all the three nonparametric prior examined (Dirichlet process prior with fixed and random mass parameter and NGG process prior), varying



(a) Logarithm of the maximal ratios. (b) Logarithm of the proportional posterior densities. The red point spots the maximum.

Figure 3.6: Examples of BH results. Dirichlet process prior with fixed mass parameter. $\mathbb{E}(K_n) = 1$, $a = 0.001$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$.

their hyperparameters as discussed at the beginning of this chapter and according to Tables 3.1, 3.2 and 3.3. In Figure (3.7) an incidence matrix of one analysis is reported (in particular, concerning the case of Dirichlet process prior with random mass parameter and prior expected number of clusters equal to 3, for the third set of hyperparameters of the mean distribution P_0), which is the same as the partition often resulting by applying the SS method. This shows how robust this method turned out to be, no matter what nonparametric prior is used, at least for hyperparameters we fixed here.

In the following sections, formulas to compute the ratio used in the BH method are presented, for different nonparametric priors.

3.5.1 Dirichlet process with fixed mass parameter

In the case of the simple Dirichlet process prior, the ratio results:

$$R_{l,h} = \frac{p(\pi^{'}|\boldsymbol{x})}{p(\pi|\boldsymbol{x})} \propto \frac{\Gamma(|C_l| + |C_h|)}{\Gamma(|C_l|)\Gamma(C_h)} \frac{m(\boldsymbol{x}^{l \cup h})}{m(\boldsymbol{x}^{(l)})m(\boldsymbol{x}^{(h)})} \frac{1}{a}$$

where letters l and h refer to the two sub-clusters evaluated at the current iteration of the BH algorithm. Of course, the factor $\frac{1}{a}$ can be avoided in the calculus, representing a proportionality term.

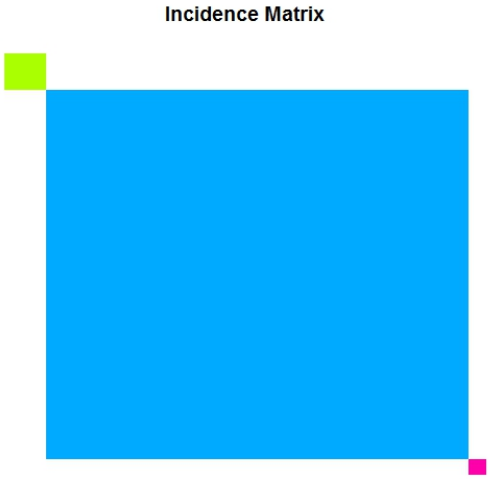


Figure 3.7: Estimate given by the application of the BH method to the Galaxy dataset. We obtained this result for all the examined configurations of nonparametric priors and hyperparameters.

3.5.2 Dirichlet process with random mass parameter

A slightly different formula has to be used when the mass parameter of the Dirichlet process prior is random. In this case, referring to the calculations made for the SS method in (3.6), we have:

$$R_{l,h} = \frac{p(\pi^{'}|\boldsymbol{x})}{p(\pi|\boldsymbol{x})} \propto \frac{\Gamma(|C_l| + |C_h|)}{\Gamma(|C_l|)\Gamma(C_h)} \frac{m(\boldsymbol{x}^{l\cup h})}{m(\boldsymbol{x}^{(l)})m(\boldsymbol{x}^{(h)})} \frac{\mathbb{E}_{p(a)}\left[\frac{a^{k^{'}}}{[a]_n}\right]}{\mathbb{E}_{p(a)}\left[\frac{a^{k}}{[a]_n}\right]}$$

where k and $k^{'}$ stand for the numbers of clusters in the two compared configurations. Notice how the factor $\frac{1}{a}$ has been replaced by the ratio of the two expected values, where $k^{'} = k - 1$. To evaluate the expected values, numerical methods have been exploited (the same as (3.3)).

3.5.3 NGG process with fixed parameters

Recalling the expression of the EPPF for a NGG process prior in (2.10) and (2.11), we can compute the ratio $R_{l,h}$ in the following way:

$$R_{l,h} = \frac{p(\pi^{'}|\boldsymbol{x})}{p(\pi|\boldsymbol{x})} \propto \frac{V_{n,k-1}}{V_{n,k}} \frac{[1-\sigma]_{|C_l|+|C_h|-1}}{[1-\sigma]_{|C_l|-1}[1-\sigma]_{|C_h|-1}} \frac{m(\boldsymbol{x}^{l\cup h})}{m(\boldsymbol{x}^{(l)})m(\boldsymbol{x}^{(h)})}$$

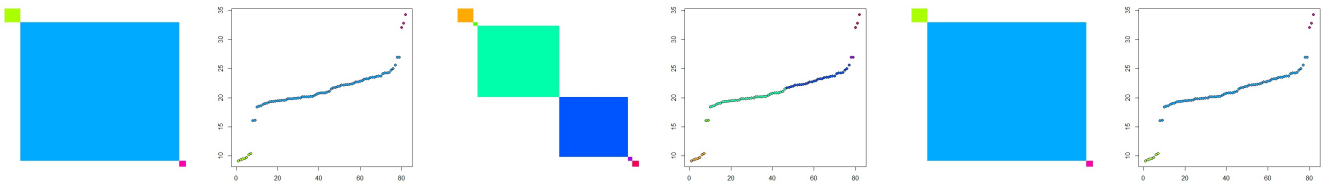
where k stands for the number of clusters K_n in the configuration examined, and $[1-\sigma]_n$ is

the rising factorial.

3.6 Loss-function minimization method

This section shows the Bayesian estimates of the random partition obtained by minimizing the standard expected posterior loss-function in (2.13) (here standard refers to the case of clusters identified by the ties in the latent vectors), proposed by Binder (1978) and Lau and Green (2007). Once again, to enlighten different backgrounds and computations, we analyze the situations of the three usual nonparametric priors, with different sets of hyperparameters of P_0. Nevertheless, the outcomes of implementing this method turned out to be very robust with respect to the choice of both the nonparametric priors and its hyperparameters. What really influences the final estimates is the choice of the set of hyperparameters of P_0.

Figure 3.8 shows the results of applying the standard loss function minimization method for different choices of the set of hyperparameters of P_0. We will not present the results for each nonparametric prior used, being all the estimates very similar. In particular, we obtained similar estimates when varying both the nonparametric prior and the values of σ and κ (we recall that the Dirichlet process prior is a particular case of the NGG one), reflecting a high robustness of the method. Figure 3.8 shows the estimates referring to $\mathbb{E}[K_n] = 3$, when the nonparametric prior is a Dirichlet process with random mass parameter. The estimates here are very similar to those presented in Section 3.4 on the SS method, in particular for the case of Dirichlet process prior with fixed mass parameter and NGG process prior (see Figures 3.3 and 3.5).



(a) (b) (c) (d) (e) (f)

Figure 3.8: Application of the standard loss function minimization method to the Galaxy dataset. Results holding for all the process prior choices ($\mathbb{E}(K_n) = 3$). (a) and (b): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$; (c) and (d): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$; (e) and (f): $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$.

3.7 A new loss-function method

In this section we will show a first application of the extended loss-function minimization method, proposed in this work. We will consider the Euclidean distance between vectors of latent variables (means and variances of the kernel densities), and the Kullback-Leibler I-divergence between kernel densities, and we apply such method to the model chosen for the Galaxy data in (3.1). When the distance is the Kullback-Leibler I-divergence, simple calculations show that:

$$KL_{ij} = \frac{((\mu_i - \mu_j)^2(\sigma_i^2 + \sigma_j^2) + (\sigma_i^2 - \sigma_j^2)^2)}{2\sigma_i^2\sigma_j^2} \tag{3.8}$$

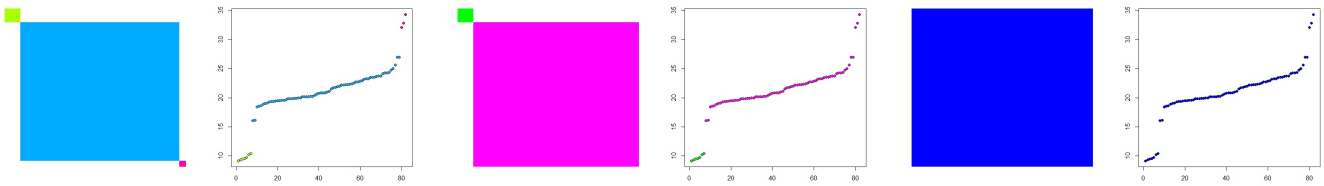
We will apply the new method in order to cluster the observations. We recall that, to perform such an analysis, we need to chose values for the parameters N and ϵ for the DBSCAN algorithm. While for the first parameter it is enough to set $N = 1$ (presence of "noise" is not allowed and partitions are uniquely defined), more difficult is to choose the second one. As expected, little values of ϵ lead to a large number of clusters, with a few elements into; conversely, large values of ϵ tend to gather all the observations in one group. Moreover, the threshold between these different behaviors is not easy to determine. Figures from 3.9 to 3.14 show different estimates provided by the new clustering method for various ϵ and for $N = 1$, for both the Euclidean distance and the Kullback-Leibler I-divergence. Differently from the methods previously presented in this chapter, we will only display the incidence matrixes of the resulting estimates. This is due to the fact that the vector of latent variables $\boldsymbol{\theta}$ cannot be identify by the new method, because the new equivalent relation is based on distances and not on ties between the values of $\boldsymbol{\theta}$.

3.8 Euclidean distance

In this section we will show the estimates of the new clustering method when the prior is Dirichlet or NGG process and the distance is the Euclidean distance between vectors of latent variables. We only present the results for the Dirichlet process prior with random mass parameter, being the estimates all very similar to each other. Furthermore, in all the estimates, we only consider the case of $\mathbb{E}[K_n] = 3$ (the other estimates are the same, showing high robustness

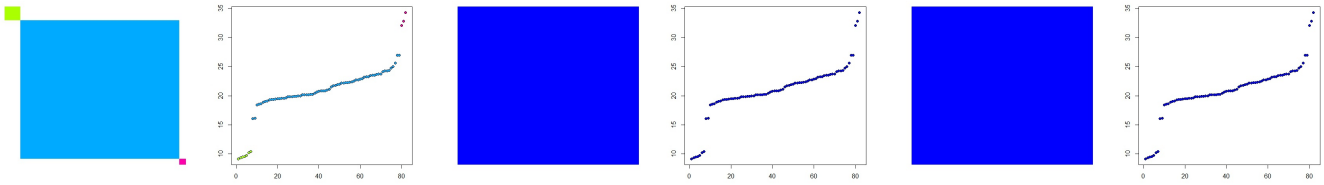
of the method). As usual, we take into account the three different sets of hyperparameters of P_0.

Values of the parameter ϵ have been chosen in order to show the difference of the resulting partition when it increases. As expected, the number of clusters decreases. We choose $\epsilon = 10, 12, 15$ for the first and third set of hyperparameters of P_0 and $\epsilon = 5, 7, 10$ for the second set.



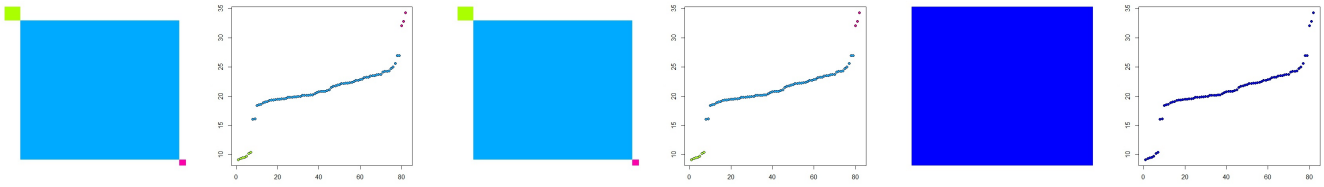
(a) $\epsilon = 10$ (b) (c) $\epsilon = 12$ (d) (e) $\epsilon = 15$ (f)

Figure 3.9: Partitions resulting from applying the new loss function minimization method (Euclidean distance). Dirichlet process prior with random mass parameter $a \sim Gamma(2, 4)$, $\mathbb{E}(K_n) = 3$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$.



(a) $\epsilon = 5$ (b) (c) $\epsilon = 7$ (d) (e) $\epsilon = 10$ (f)

Figure 3.10: Partitions resulting from applying the new loss function minimization method (Euclidean distance). Dirichlet process prior with random mass parameter $a \sim Gamma(2, 4)$, $\mathbb{E}(K_n) = 3$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$.



(a) $\epsilon = 10$ (b) (c) $\epsilon = 12$ (d) (e) $\epsilon = 15$ (f)

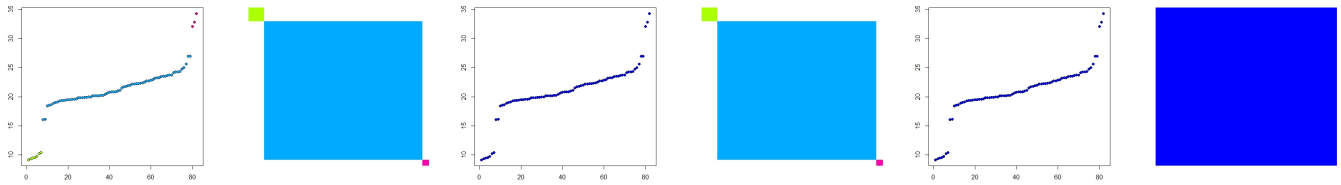
Figure 3.11: Partitions resulting from applying the new loss function minimization method (Euclidean distance). Dirichlet process prior with random mass parameter $a \sim Gamma(2, 4)$, $\mathbb{E}(K_n) = 3$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$.

3.9 Kullback-Leibler I-divergence

The estimates shown in this section refer to the application of the new method considering the symmetrized Kullback-Leibler I-divergence, defined in formula (2.16). After further analysis, we observed that the new method defined using the Kullback-Leibler I-divergence is barely influenced by the choice of the nonparametric prior and its hyperparameters, this showing a high robustness of the method. Hence, we decided to present only the estimates in case of $\mathbb{E}[K_n] = 3$, which are similar for both Dirichlet and NGG process priors. In particular, Figures from 3.12 to 3.14 refer to the Dirichlet process prior with random mass parameter.

As far as the choice of ϵ is concerned, here we include analysis where the estimated partitions are enough different from each other, in order to show how the method works. In particular, we choose to put $\epsilon = 2, 6, 8$. We recall that, in the case of Kullback-Leibler I-divergence, we consider its log-transformation, that is $\log(1 + KL_{ij})$, and consider the inequality $\log(1 + KL_{ij}) \leq \epsilon$ to define the new equivalence relation.

The figures of this section show that the estimates are more sensitive to the choice of the hyperparameters of P_0 with respect to the case of the Euclidean distance. Furthermore, the estimated configuration with 3 clusters is often found by the method.

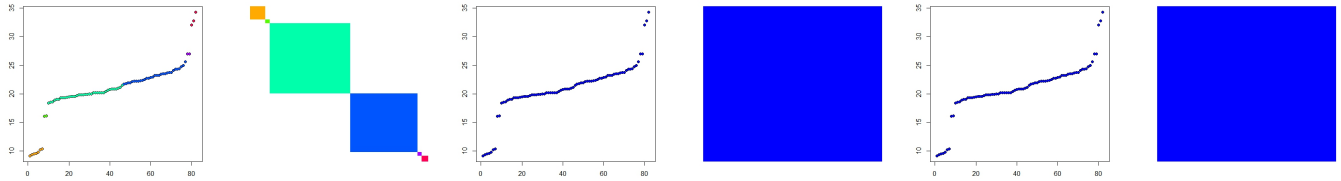


(a) $\epsilon = 2$ (b) (c) $\epsilon = 6$ (d) (e) $\epsilon = 8$ (f)

Figure 3.12: Partitions resulting from applying the new loss function minimization method (logarithm of Kullback-Leibler I-divergence). Dirichlet process prior with random mass parameter $a \sim Gamma(2, 4)$, $\mathbb{E}(K_n) = 3$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$.

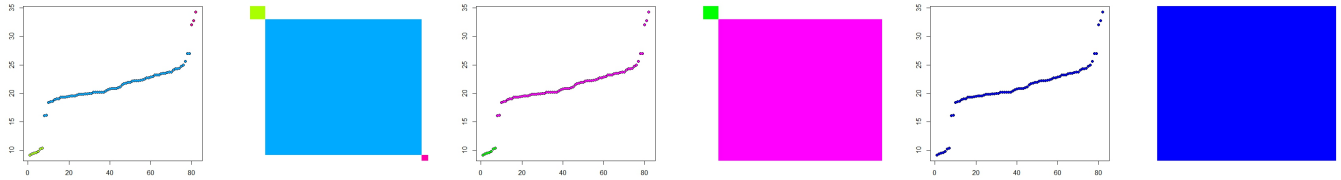
3.9.1 Final Considerations

In this case, we need to fix the further hyperparameter ϵ to provide an estimate of the random partition. From the definition of the method itself, and as shown in figures above, when fixing a value of $\epsilon > 0$, the method tends to find less clusters than the standard loss function method (i.e., by setting $\epsilon = 0$).



(a) $\epsilon = 2$ (b) (c) $\epsilon = 6$ (d) (e) $\epsilon = 8$ (f)

Figure 3.13: Partitions resulting from applying the new loss function minimization method (logarithm of Kullback-Leibler I-divergence). Dirichlet process prior with random mass parameter $a \sim Gamma(2, 4)$, $\mathbb{E}(K_n) = 3$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 20, 20)$.



(a) $\epsilon = 2$ (b) (c) $\epsilon = 6$ (d) (e) $\epsilon = 8$ (f)

Figure 3.14: Partitions resulting from applying the new loss function minimization method (logarithm of Kullback-Leibler I-divergence). Dirichlet process prior with random mass parameter $a \sim Gamma(2, 4)$, $\mathbb{E}(K_n) = 3$ and $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$.

In all the results presented in this section, the method turned out to be very influenced by the choice of the hyperparameters of P_0. Nevertheless, we can argue that the method is very robust with respect to the choice of the nonparametric prior used. In fact, we only displayed the estimates concerning the case of Dirichlet process prior with random mass parameter. In order to define the new equivalent relation for the new clustering method we proposed two different choices of the distance to use: the Euclidean distance and the Kullback-Leibler I-divergence between latent vectors $\boldsymbol{\theta}$. As Figures 3.9 to 3.14 show, the estimates in case of different distances are quite similar (an exception is made for the case presented in Figure 3.12 (a) and (b)).

Chapter 4

Kevlar Data

The next dataset we will analyze consists of $n = 108$ lifetimes of pressure vessels, wrapped with a Kevlar yarn, coming from 8 different spools, at different levels of pressure (23.4, 25.5, 27.6 and 29.7 MPa). Eleven lifetimes with the lowest level of pressure are right censored at the time 41.000 hours. The model chosen to describe the observed lifetimes is the one proposed by Argiento et al. (2010), consisting of a semiparametric Bayesian Weibull regression model, where the random effect is not induced by the spool classification, but is inferred via a nonparametric component. In this model, the log-lifetimes are taken as the response, while the covariates are represented by a proper function of the stress levels ($x_i = \log \frac{Stress_i}{\min Stress} = \log \frac{Stress_i}{23.4}$). The model is the following:

$$
\begin{aligned}
&T_i = e^{x_i\beta} \cdot V_i, \text{ for } i = 1, ..., n \\
&V_1, ..., V_n | \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_n} \overset{\text{ind.}}{\sim} Weibull(v_i | \theta_{i1}, \theta_{i2}) \\
&\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_n} | P \overset{\text{i.i.d.}}{\sim} P \\
&P \sim NGG(\sigma, \kappa P_0) \\
&P_0(d\theta_{i1}, d\theta_{i2}) = Gamma(d\theta_{i1} | a, b) \times Gamma(d\theta_{i2} | c, d) \\
&\beta \sim N(0, 10^3)
\end{aligned}
\tag{4.1}
$$

where by a $Weibull(\theta_1, \theta_2)$ random variable we mean the variable with survival function

$$
S(w) = exp\{-\left(\frac{w}{\theta_2}\right)^{\theta_1}\}, w \geq 0; \tag{4.2}
$$

note that $\theta_1 > 0$ and $\theta_2 > 0$ represent the shape and the scale parameters, respectively.

Of course, on the log-scale the model can be expressed as:

$$\log T_i = x_i\beta + \log\theta_{i2} + \frac{W_i}{\theta_{i1}}, \text{ for } i = 1, ..., n,$$

where W_i represents the error in the regression model, and $W_i \sim Gumbel(0, 1)$, i.e. a random variable having survival function e^{-e^w}, with $\mathbb{E}[W_i] = -\gamma$ (minus the Euler-Mascheroni constant) and $Var(W_i) = \frac{\pi^2}{6}$.

In this model, the nonparametric component induces a grouping criterion on the observations. In fact, thanks to the discreteness of the trajectories of the nonparametric prior imposed on $\boldsymbol{\theta}$, the sequence $\{\log\theta_{i2}\}$ contains ties with positive probability, thus inducing a partition on the observations. As seen before, the grouping of the lifetimes is not fixed, but random, as well as the number of clusters.

In the next sections, we will show the estimates provided by the loss-function minimization method using both the standard and the new similarity matrix.

Notice that the model presented here is very different for the one in (3.1), used in Chapter 3 to describe the Galaxy data. First of all, the model in (4.1) is not a conjugate one, and so we cannot apply clustering method involving analytical computations of the posterior density of the random partition (such as SS and BH method presented in Sections 2.4.1 and 2.4.2). Additionally, here we have the presence of covariates, differently from model (3.1).

4.1 Cluster analysis using the standard similarity matrix

In this section the Bayesian estimates under posterior loss-function minimization are presented, considering the usual classification based on the equality of the latent variables $\boldsymbol{\theta}$. Even if we did the analysis for different sets of values for the hyperparameters, here we only report the estimates for two different configurations of the hyperparameters of P_0, (a, b, c, d). The first configuration correspond to $(a, b, c, d) = (1, 1, 1, 1)$; the second one to $(a, b, c, d) = (0.5, 0.044, 2, 2)$. For the other NGG hyperparameters, we set $(\sigma, \kappa) = (0.1, 10)$. For further details about the choice of these values, see Argiento et al. (2010). As mentioned in Section 2.4.3, we need to fix a value for the parameter of the loss-function $\hat{K} = \frac{b}{a+b}$, where a and b represent the two

misclassification costs. Here, it will be held equal to 0.5 (corresponding to the Dahl quadratic loss-function (2006), i.e. equal misclassification costs).

In Figures 4.1 and 4.2, the incidence matrixes of the clustering estimates are shown, together with the posterior kernel density estimates of $\log \theta_{2i}$, $i = 1, ..., n$. These quantities have a random intercept interpretation into the log-scale model of the Kevlar lifetime fibres. Thus, it is useful to see their posterior distributions, in order to better understand the performed clustering. In particular, the plot of the estimated posterior densities of the quantities $\log \theta_{2i}$ on the left of Figures 4.1 and 4.2 shows a very clear clustering of the observations into three different groups (i.e., each evident mode represents a cluster), which is coherent with both the Bayesian analysis carried out by Argiento et al. (2012) and the frequentist study reported in Crowder et al. (1991). The first set of hyperparaters of P_0, $(a, b, c, d) = (1, 1, 1, 1)$, leads to a more dispersive partition of the data (see Figure 4.1 (b)), involving 9 different clusters of different shapes, while from Figure 4.2 (b) (second set of hyperparameters of P_0, $(a, b, c, d) = (0.5, 0.044, 2, 2)$), the subdivision into three groups is more evident. Therefore we argue that the loss function minimization method provided by the standard similarity matrix, based on the ties in the vectors of latent variables, is not robust.

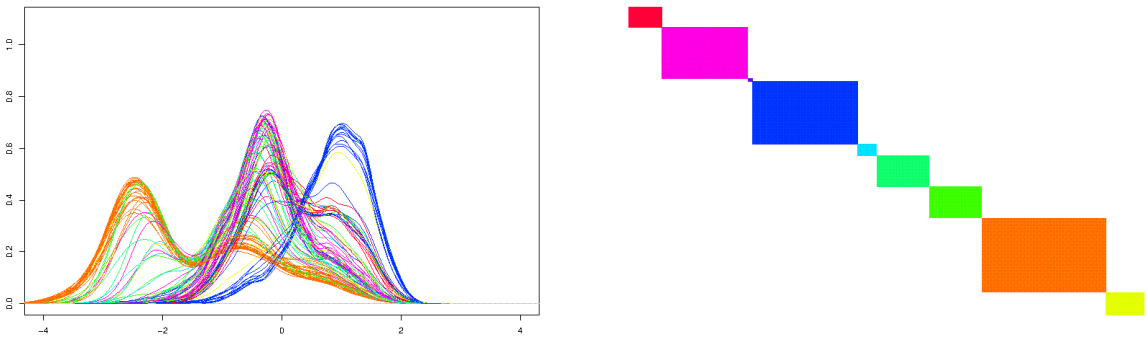


Figure 4.1: Application of the standard loss-function minimization method. $(a, b, c, d) = (1, 1, 1, 1)$.

4.2 Cluster analysis using the new similarity matrix

It is now possible to show the estimates from the new clustering method, described in Section 2.5. First of all, in order to introduce a new equivalence relation based on the distance between

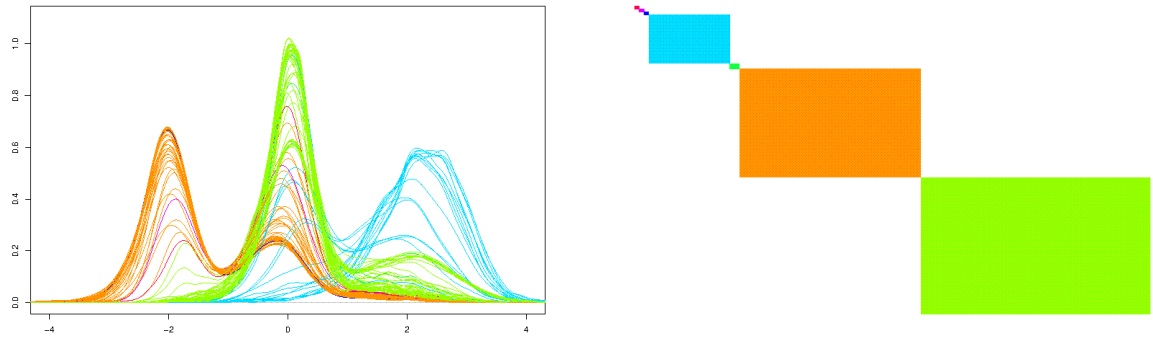


Figure 4.2: Application of the standard loss-function minimization method. $(a, b, c, d) = (0.5, 0.04, 2, 2)$.

the latent variables, it is necessary to specify which distance will be used. In this section, Euclidean distance and Kullback-Leibler I-divergence in (2.16) are considered. As far as the method based on the Euclidean distance is concerned, we will use the distance between means and variances of the corresponding Weibull densities $W(\theta_1, \theta_2)$. Regarding the choice of N, we fixed it equal to 1.

4.2.1 Euclidean Distance

The expressions of the mean and variance of a random variable $W \sim Weibull(\alpha, \beta)$ defined in formula (4.2) are:

$$\begin{aligned} \mathrm{E}[W] &= \beta\Gamma(1 + \frac{1}{\alpha}) \\ Var[W] &= \beta^2(\Gamma(1 + \frac{2}{\alpha}) - \Gamma(1 + \frac{1}{\alpha})^2) \end{aligned} \tag{4.3}$$

In order to apply the new equivalence relation, we had to choose values for the additional parameters N and ϵ. In this case we fixed $N = 1$, and $\epsilon = 0.5, 1, 1.5$. We recall that the DBSCAN algorithm (the clustering algorithm associated with the new equivalence relation), when the parameter N is grater than 1, is able to locate "noise" elements, i.e. those elements not satisfying the equivalence relation with any other object. In our case, we do not allow the existence of such elements and decided to put always $N = 1$ (even a singleton, when found by the algorithm, is considered as a cluster). Furthermore, when $N > 1$, the partitions found by the DBSCAN algorithm are not uniquely determined, and can verify some situations of

labelling conflict.

In Figures 4.3, 4.4 and 4.5, we report the estimates, with the Euclidean distance between means and variances of the Weibull kernel densities to define the new equivalence relation in (2.5), applied to the model (4.1) with the hyperparameters $(a, b, c, d) = (1, 1, 1, 1)$. On the other hand, when hyperparameters are $(a, b, c, d) = (0.5, 0.044, 2, 2)$, we found that the estimate is about the same for all different values of ϵ (the estimate is shown in Figure 4.6), showing a high robustness with respect to the value of ϵ.

As one can immediately see from the incidence matrixes reported in Figures 4.3 - 4.6, the estimated number of clusters is not far from 3 (sometimes 4 or 5 clusters are found, some of them composed of a few elements; only Figure 4.3 shows a clear overestimation of the number of clusters). This result not only supports the Bayesian analysis carried out by Argiento et al. (2012), and the frequentist analysis of Crowder et al.(1991), but also shows how our new method is more robust. Of course, the choice of the parameter ϵ is a key point of such an analysis, though, in this case, the new method proves to be quite robust with respect to it (only Figure 4.3 is unsatisfactory, since it is not coherent with previous frequentist and Bayesian analysis).

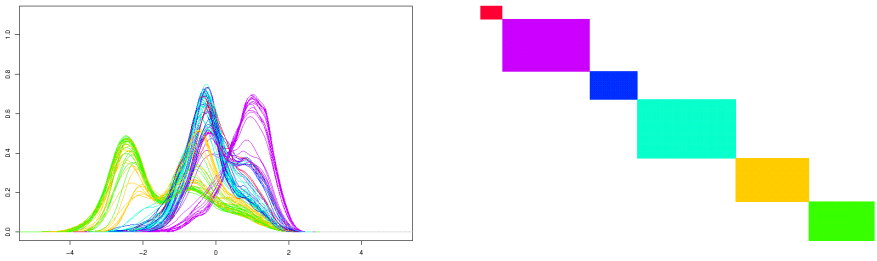


Figure 4.3: Estimate given by the new loss-function minimization method (Euclidean Distance). $N = 1$, $\epsilon = 0.5$ and $(a, b, c, d) = (1, 1, 1, 1)$.

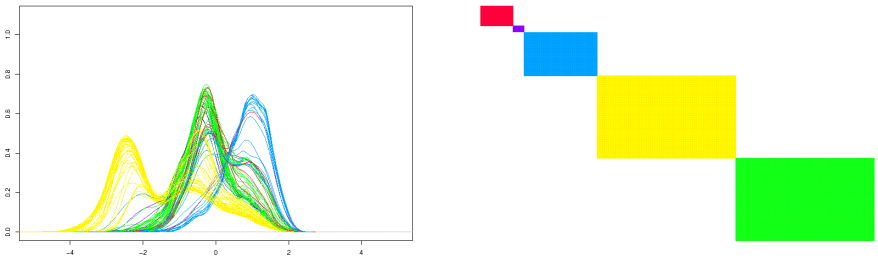


Figure 4.4: Estimate given by the new loss-function minimization method (Euclidean Distance). $N = 1$, $\epsilon = 1$ and $(a, b, c, d) = (1, 1, 1, 1)$.

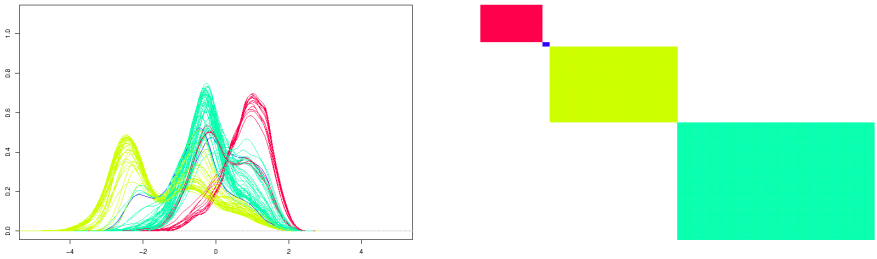


Figure 4.5: Estimate given by the new loss-function minimization method (Euclidean Distance). $N = 1$, $\epsilon = 1.5$ and $(a, b, c, d) = (1, 1, 1, 1)$.

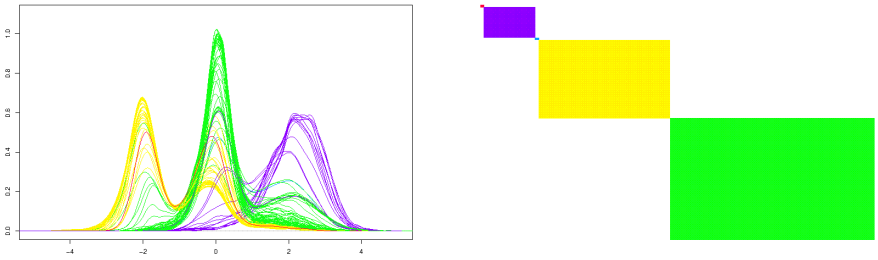


Figure 4.6: Estimate given by the new loss-function minimization method (Euclidean Distance). $N = 1$ and $(a, b, c, d) = (0.5, 0.044, 2, 2)$.

4.2.2 Kullback-Leibler I-divergence

The other distance we consider here is the Kullback-Leibler I-divergence, defined in (2.16). In the case of Kevlar data, following the model above, Weibull kernel densities are used to compute the I-divergence. The $Weibull(\alpha, \beta)$ densities have the following expression:

$$f_W(w|\alpha, \beta) = \frac{\alpha}{\beta^\alpha} w^{\alpha-1} e^{-(\frac{w}{\beta})^\alpha} \mathrm{I}_{(0,+\infty)}(w).$$

The symmetrized Kullback-Leibler I-divergence in (2.16) between two latent variables results:

$$KL_{ij} = (\alpha_i - \alpha_j)(\log \frac{\beta_i}{\beta_j} - \gamma(\frac{1}{\alpha_i} - \frac{1}{\alpha_j})) - 2 + \frac{\beta_i}{\beta_j}^{\alpha_j} \Gamma(1 + \frac{\alpha_j}{\alpha_i}) + \frac{\beta_j}{\beta_i}^{\alpha_i} \Gamma(1 + \frac{\alpha_i}{\alpha_j}),$$

where γ is the Euler-Mascheroni constant. Figures 4.7 - 4.10 show the Bayesian clustering of the Kevlar dataset using the Kullback-Leibler I-divergence to implement the minimization algorithm described in section 2.5, when the hyperparameters of P_0 are $(a, b, c, d) = (1, 1, 1, 1)$ and $(a, b, c, d) = (0.5, 0.044, 2, 2)$. In particular, Figures 4.9 and 4.10 show the estimated clustering using $\log(1 + KL_{ij})$ to define the clusters, for two different values of ϵ. Once again, kernel density estimates and incidence matrixes are reported. We show results with fixed $N = 1$. Finally, in Figure 4.11, we report the estimate concerning the second set of hyperparameters of P_0, which gives always the same result, for all the different values of ϵ (still using the logarithm of the Kullback-Leibler I-divergence). This result is the same as in the case of the Euclidean distance, showing that the method is robust even with respect to the choice of the distance used to define the DBSCAN algorithm leading to the groups.

4.2.3 Final Considerations

From the results shown in this chapter, it is evident that the new method is robust with respect to the distance used to define the new equivalence relation. In fact, varying from Euclidean distance to Kullback-Leibler I-divergence does not seem to influence very much the corresponding clustering estimates.

Different considerations must be made concerning the choice of the hyperparameters of the mean distribution P_0. As usually, this choice strongly influences the resulting outcomes. In this particular model, we observed that the first set of hyperparameters, in the case of Eu-

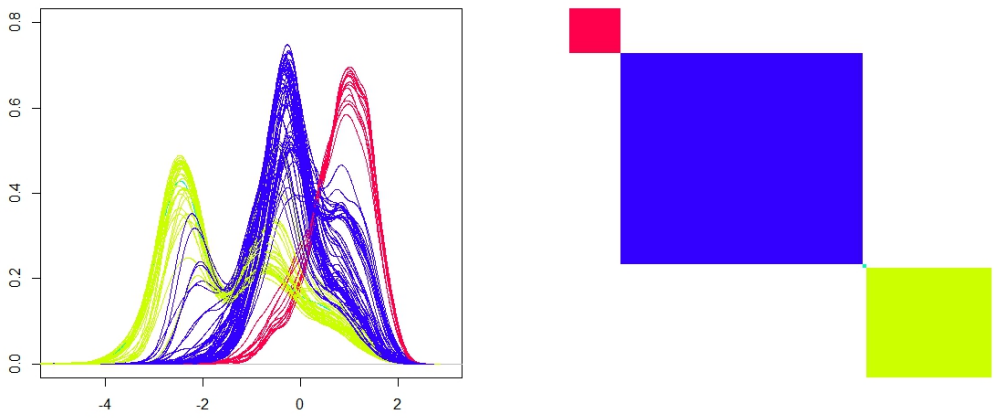


Figure 4.7: Estimate given by the new loss-function minimization method (KL I-divergence). $N = 1, \epsilon = 1$ and $(a, b, c, d) = (1, 1, 1, 1)$.

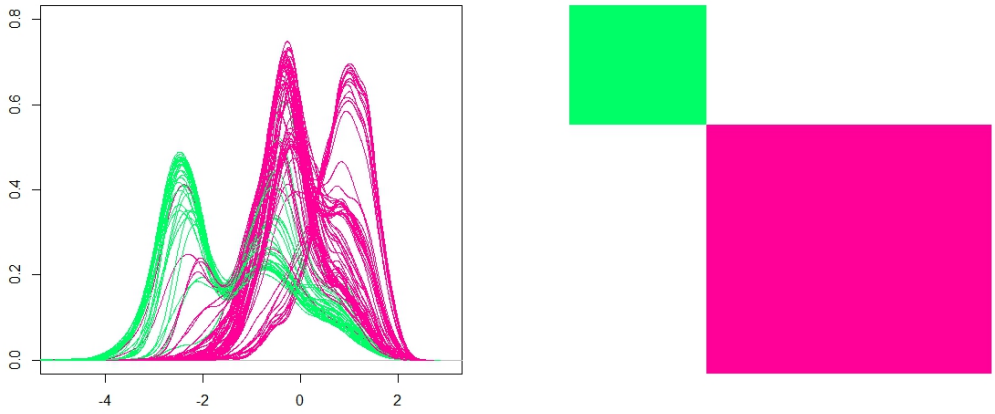


Figure 4.8: Estimate given by the new loss-function minimization method (KL I-divergence). $N = 1, \epsilon = 1.5$ and $(a, b, c, d) = (1, 1, 1, 1)$.

clidean distance, leads to a different estimate for the random partition, when compared with the Bayesian study carried out by Argiento et al. (2012) and the frequentist analysis of Crowder et al. (1991). Nevertheless, in this sense, the second set of hyperparameters gives better and more robust results. We point out that many other sets of such hyperparameters have been studied, but we choose to present only the two most significative, which carried different information and estimates.

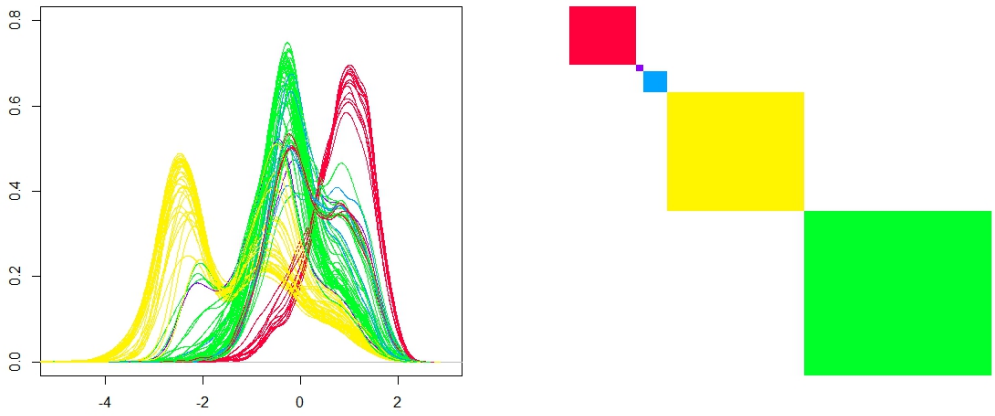


Figure 4.9: Estimate given by the new loss-function minimization method ($\log(1 + KL)$). $N = 1, \epsilon = 1$ and $(a, b, c, d) = (1, 1, 1, 1)$.

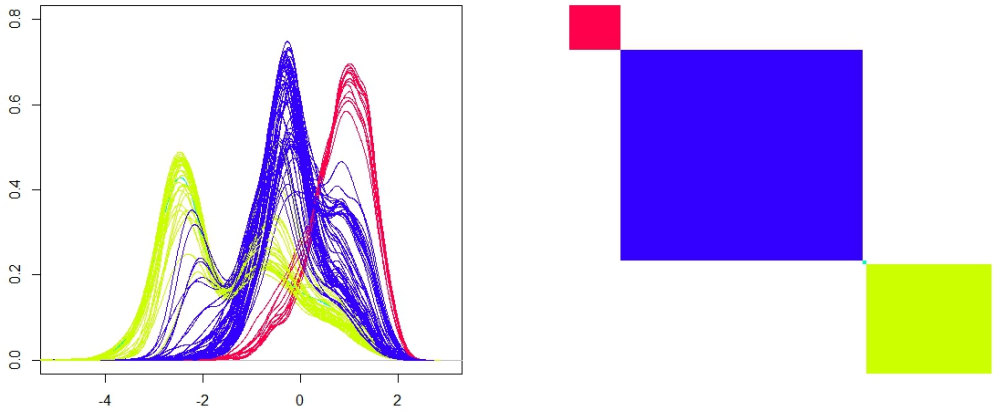


Figure 4.10: Estimate given by the new loss-function minimization method ($\log(1 + KL)$). $N = 1, \epsilon = 1.5$ and $(a, b, c, d) = (1, 1, 1, 1)$.

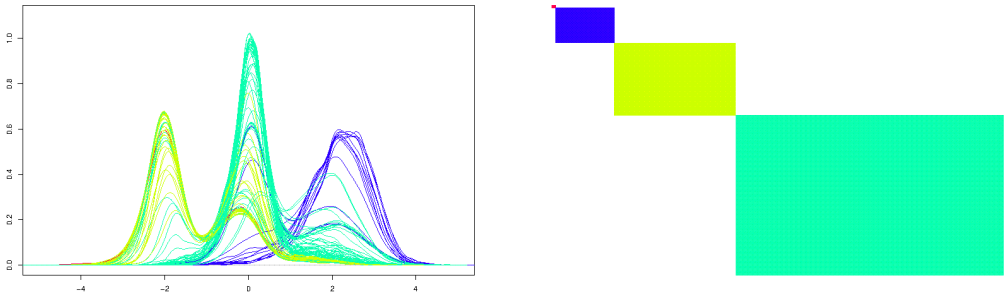


Figure 4.11: Estimate given by the new loss-function minimization method (KL I-divergence). $N = 1$ and $(a, b, c, d) = (0.5, 0.044, 2, 2)$.

Chapter 5

Simulated bivariate dataset having a non-convex support

In this chapter we present the clustering estimates given by the new loss-function minimization methods presented in Section 2.4.3 and 2.5, applied to two bivariate datasets of $n = 250$ and $n = 1000$ elements, respectively. In particular, we simulated the i.i.d. observations from a mixture of bivariate normal densities, imposing a distribution for the mean parameters. The density from which we sampled the two datasets is the following:

$$qN_2(\mathbf{0}, B) + (1-q)N_2(\boldsymbol{\beta}, B/5), \quad \text{where} \quad B = diag(0.1, 2),$$
$$\xi \sim U_{[-\pi/2,+\pi/2]}(d\xi), \quad \text{and} \quad \beta_1 = \cos\xi, \beta_2 = \sin\xi.$$

Here $N_2(\boldsymbol{\beta}, B)$ and $U_{[a,b]}(\cdot)$ stand for the bivariate normal distribution with mean vector $\boldsymbol{\beta}$ and covariance matrix B and the uniform distribution with mean $\frac{a+b}{2}$ and variance $\frac{(b-a)^2}{12}$, respectively. We fixed the weight of the mixture q at the value 0.5 for the first dataset ($n = 250$) and 0.35 for the second one ($n = 1000$); see Figure 5.2, 5.4 and 5.6. The resulting datasets are composed of two main groups of observations due to the two contributes of the mixture density: the first one has a sharp round shape and it is located around the value $\mathbf{0}$, while the second group lays on a semicircular region on the right of the first group. This peculiar disposition of the observations on a non-convex support is a popular choice when dealing with clustering algorithms (see, for example http://en.wikipedia.org/wiki/DBSCAN), in order evaluate how well they perform even in "unusual" situations.

In order to provide clustering estimation, we model the observations via a DPM model, as described in (2.7), using a Dirichlet process prior with random mass parameter $a \sim Gamma(\gamma_1, \gamma_2)$, where $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$, for $i = 1, ..., n$. In the case of dataset size $n = 250$ we choose two different sets of hyperparameters for the total mass parameter a, one such that $\mathbb{E}[a] = \frac{\gamma_1}{\gamma_2} = 15$ and $Var(a) = \frac{\gamma_1}{\gamma_2^2} = 2$, and the other one such that $(\gamma_1, \gamma_2) = (2, 0.01)$. When dealing with $n = 1000$ observations, we fixed only a set of values for this hyperparameters, such that $\mathbb{E}[a] = 11$ and $Var(a) = 4$. Notice that this choice leads to a large prior expected number of clusters and to quite non-informative prior distributions on a.

As far as the choice of the rest of the hyperparameters of the Dirichlet process concerns, that is the values of the set $(\boldsymbol{m}_0, k_0, \nu_1, \Psi_1)$, we fix them as in Chapter 3: we want a large prior variance for $\boldsymbol{\mu}$ and small prior mean and variance for Σ. Similarly to formulas in (3.2), we have:

$$Var(\boldsymbol{\mu}) = \frac{\Psi_1}{(\nu_1 - p - 1)k_0}, \quad \text{for } \nu_1 > p + 1;$$

$$\mathbb{E}[\Sigma] = \frac{\Psi_1}{(\nu_1 - p - 1)}, \quad \text{for } \nu_1 > p + 1;$$

$$Var(\Sigma)_{ij} = \frac{(\nu_1 - p + 1)\Psi_{1ij}^2 + (\nu_1 - p - 1)\Psi_{1ii}\Psi_{1jj}}{(\nu_1 - p)(\nu_1 - p - 1)^2(\nu_1 - p - 3)}, \quad \text{for } \nu_1 > p + 3.$$

where $p = 2$ represents the data dimension. Additionally, we imposed that Ψ_1 is a diagonal matrix. In order to obtain the fixed values of the hyperparameters we assumed:

$$\boldsymbol{m}_0 = \mathbf{0}, k_0 = 0.001, \nu_1 = 10, \Psi_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

We could have applied all the methods presented in Sections 2.4 and 2.5 to these two datasets, since the model we assumed is once again a conjugate DPM model, as in Chapter 3, but we decided to show only the estimates provided by the loss-function minimization methods, being the ones of particular interest in this work. In particular, we will present the loss-function minimization methods based on both the standard and the new similarity matrix (we recall that the standard method from Binder (1978) and Lau and Green (2007) can be seen as a particular case of the new proposed method, fixing the parameter ϵ of the DBSCAN algorithm

equal to zero).

5.1 Loss-function minimization methods

In order to apply the new clustering method, we need to choose a distance between the latent variables $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$, for $i = 1, ..., n$, of the DPM model, in order to define the new equivalent relation, as reported in Section 2.5. We decided to use the following distances in $\mathbb{R}^2$:

- Euclidean Distance between mean vectors,
- Symmetrized Kullback-Leibler I-divergence,
- Hellinger Distance,
- L^2 Distance,
- Symmetrized Mahalanobis Distance.

The clustering estimates provided through the first and the last distance will not be shown here, being very similar to the ones from the other distances. Therefore, here we present only their analytical expressions:

$$d_E(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \sqrt{(\mu_{i1} - \mu_{j1})^2 + (\mu_{i2} - \mu_{j2})^2},\ i, j = 1, ..., n;$$

$$d_{SymM}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = 0.5(d_M(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) + d_M(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)) = 0.5(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\Sigma_i^{-1} + \Sigma_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),\ i, j = 1, ..., n.$$

Before showing the clustering estimates, we recall that values for the parameter ϵ must be chosen. We choose to relate the value of ϵ to the prior distribution of the distances above introduced, estimated using samples from the prior distribution of the unique values $\boldsymbol{\phi}$ of the latent vectors. In fact, as functions of the latent variables $\boldsymbol{\theta}$, the distances that we use to define the new similarity matrix are random variables, as well. In particular, we let ϵ varies between 0 and the quantiles of order 0.01, 0.5 and 0.99, respectively, of the distances' prior distributions. These choices are related to different prior beliefs, reflecting our information about the prior distances of the latent vectors $\boldsymbol{\theta}$.

5.1.1 Kullback-Leibler I-divergence

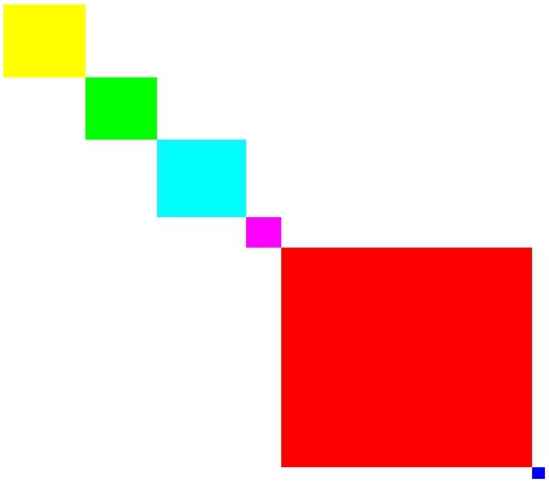
For the case of bivariate Gaussian kernel densities, the expression of the symmetrized Kullback-Leibler I-divergence, defined in (2.16), is:

$$KL_{ij} = 0.5(-2p + Tr(\Sigma_i\Sigma_j^{-1}) + Tr(\Sigma_j\Sigma_i^{-1}) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\Sigma_i^{-1} + \Sigma_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)),\ i, j = 1, ..., n$$

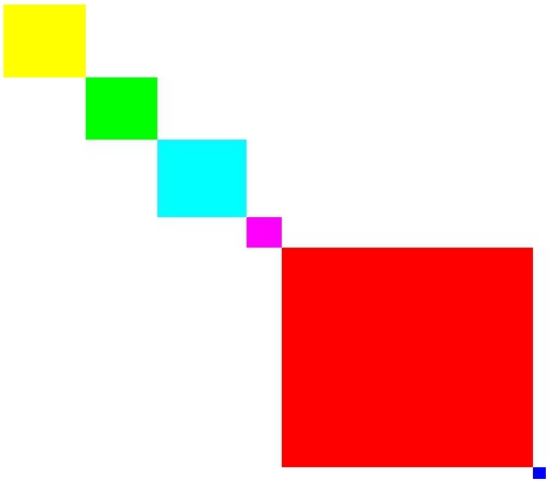
where $Tr(A)$ represents the trace of the square matrix A. Notice that the last term of this expression is the symmetrized Mahalanobis. This is the reason why using these two distances in the definition of the new equivalent relation leads to very similar clustering estimates.

Figures 5.1 to 5.6 show the estimates given by choosing ϵ equal to 0 or according to the values of the quantiles of order 0.01, 0.5 and 0.99, in the case of $n = 250$ and $n = 1000$. The images are grouped according to the dataset analyzed and the values chosen for $(\boldsymbol{m}_0, k_0, \nu_1, \Psi_1)$. In particular, the first of each group of figures shows the case of $\epsilon = 0$, representing the estimate provided by the method proposed by Binder (1978) and Lau and Green (2007). As before, incidence matrixes and scatterplots of the estimated partitions are displayed; each group has the same colour in the two graphs.

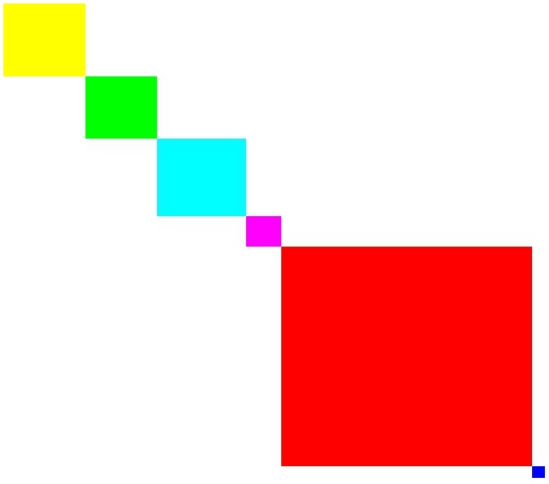
As can be easily inferred from the first three plots in Figures from 5.1 to 5.6, the estimates given by the new loss-function minimization method are the same for the first three values of ϵ, that is 0, $q_{0.01}$ and $q_{0.5}$. Furthermore, the different number of observations or set of hyperparameters of P_0 have little influence on the resulting estimates. As expected, the number of clusters reduces when ϵ increases. Differently, as indicated in the fourth plot of each group of figures, when $\epsilon = q_{0.99}$, we have estimated partitions composed of 2 or 3 clusters, which reflect in a very good way the original partition (we recall that the observations have been simulated from a mixture of two bivariate Gaussian densities).



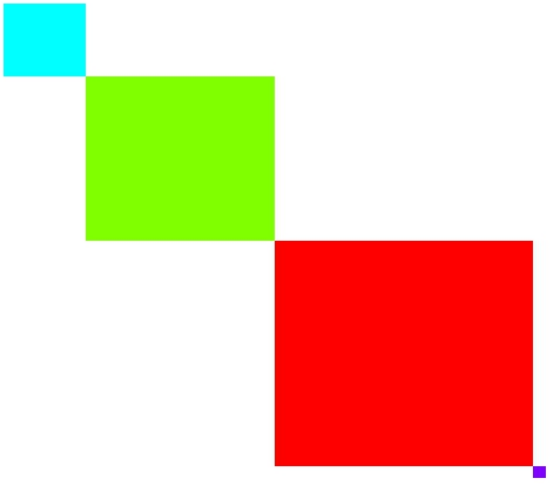
(a) $\epsilon = 0$.



(b) $\epsilon = q_{0.01}$

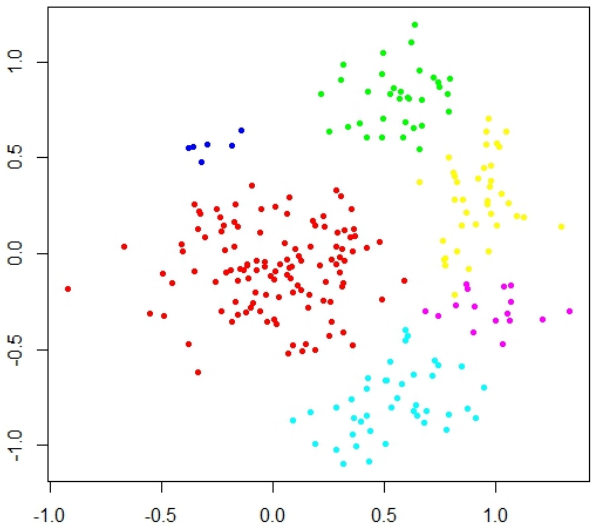


(c) $\epsilon = q_{0.5}$.

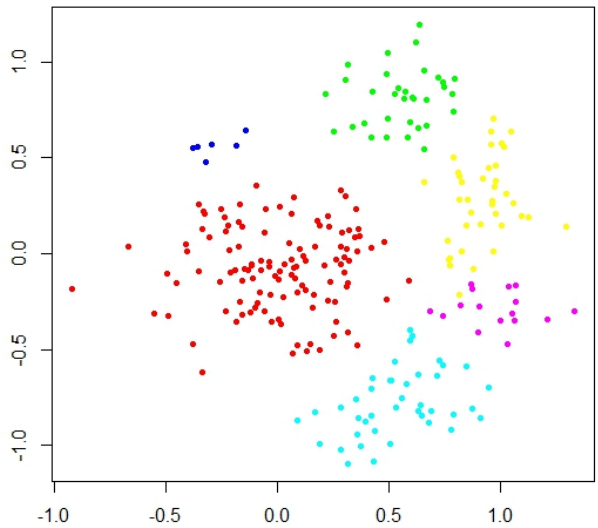


(d) $\epsilon = q_{0.99}$

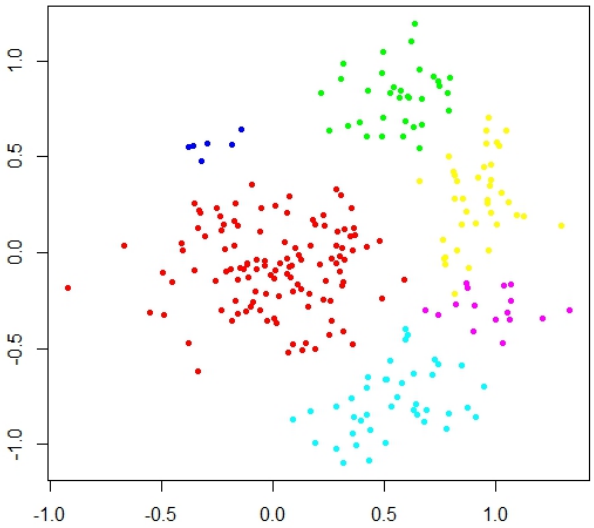
Figure 5.1: Incidence matrixes of the clustering estimates given by the new loss-function minimization method with KL I-divergence, for $N = 1$ and different values of ϵ. $n = 250$, first set of hyperparameters.



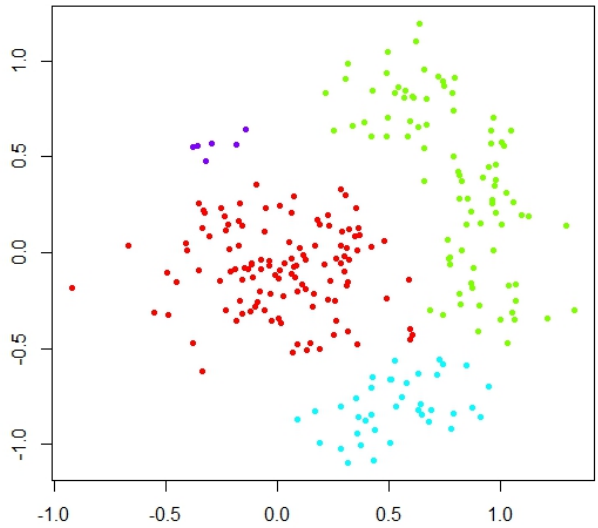
(a) $\epsilon = 0$.



(b) $\epsilon = q_{0.01}$

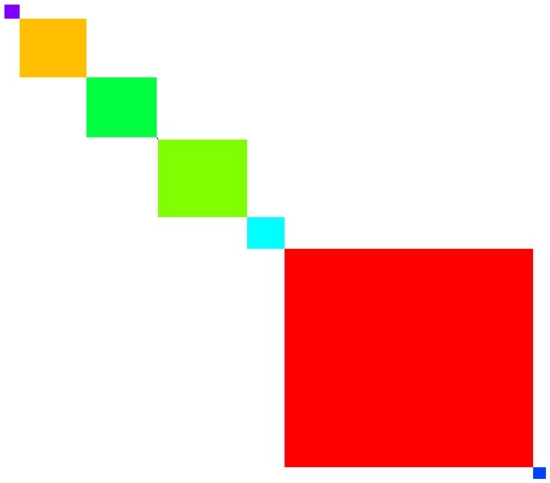


(c) $\epsilon = q_{0.5}$.

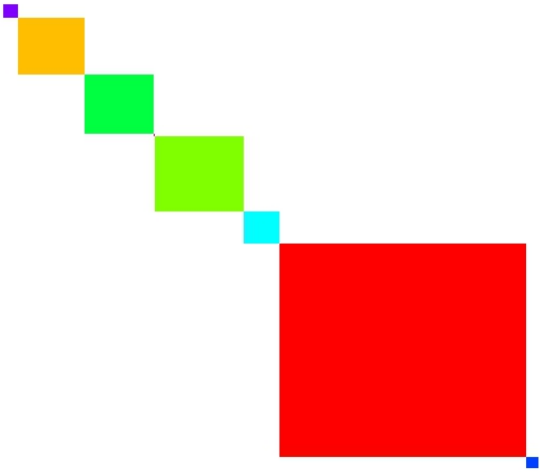


(d) $\epsilon = q_{0.99}$

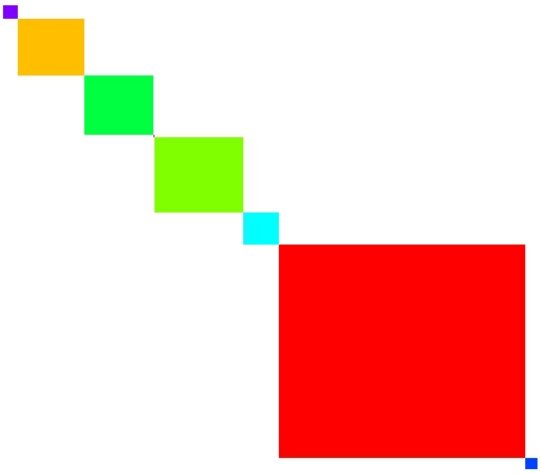
Figure 5.2: Scatterplots of the clustering estimates given by the new loss-function minimization method with KL I-divergence, for $N = 1$ and different values of ϵ. $n = 250$, first set of hyperparameters.



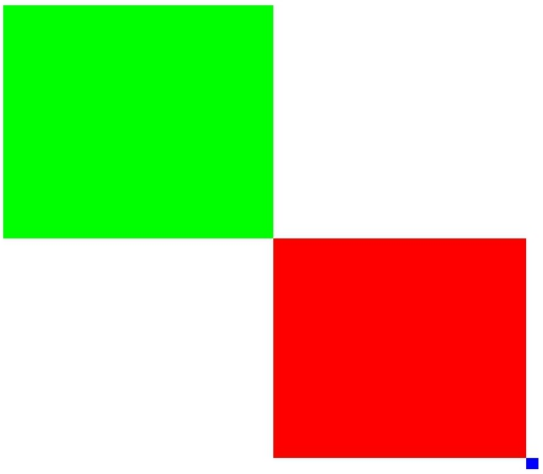
(a) $\epsilon = 0$.



(b) $\epsilon = q_{0.01}$

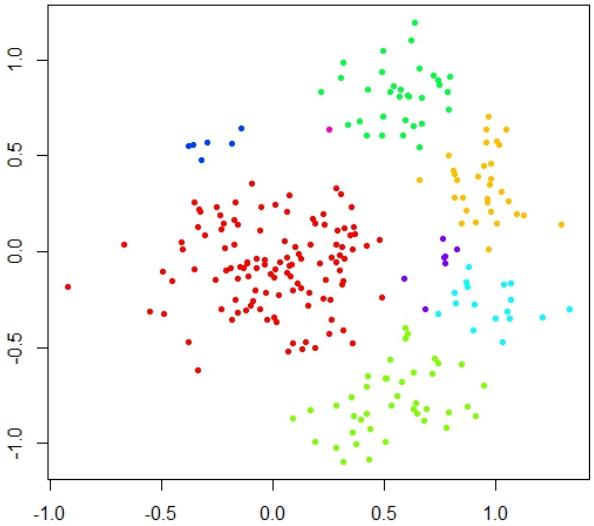


(c) $\epsilon = q_{0.5}$.

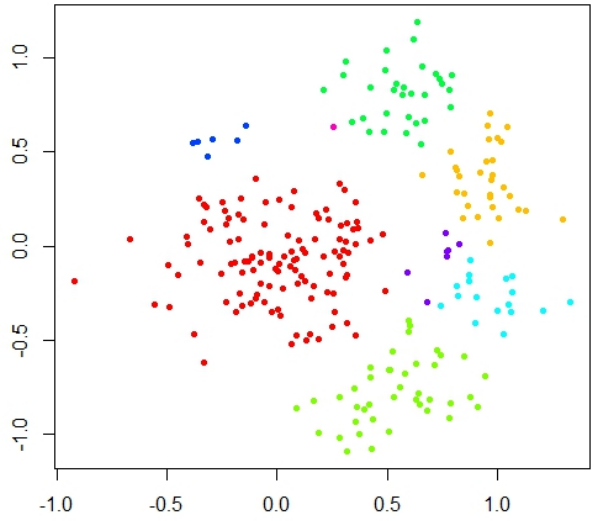


(d) $\epsilon = q_{0.99}$

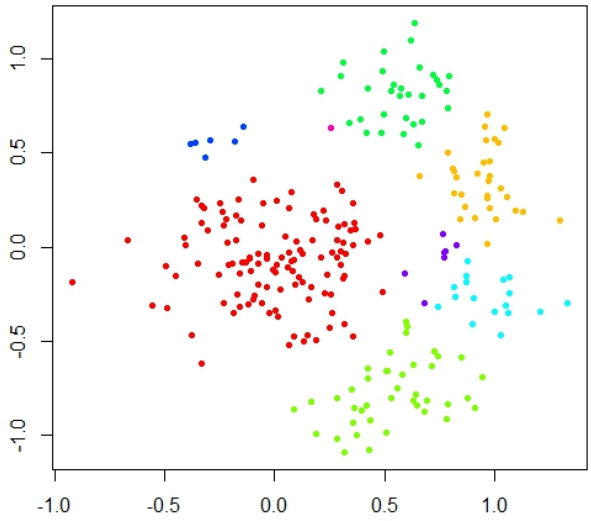
Figure 5.3: Incidence matrixes of the clustering estimates given by the new loss-function minimization method with KL I-divergence, for $N = 1$ and different values of ϵ. $n = 250$, second set of hyperparameters.



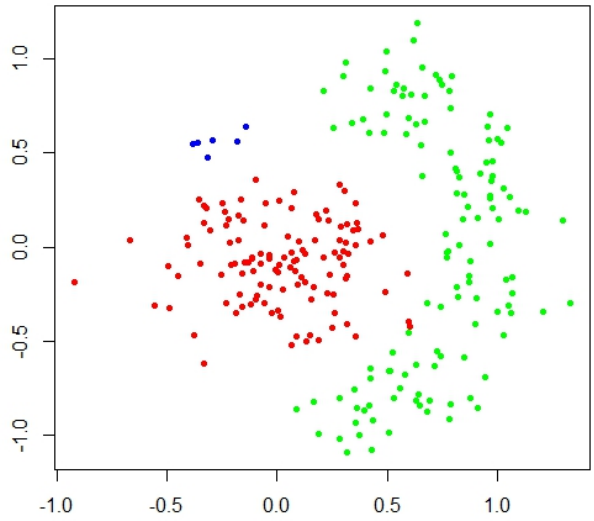
(a) $\epsilon = 0$.



(b) $\epsilon = q_{0.01}$

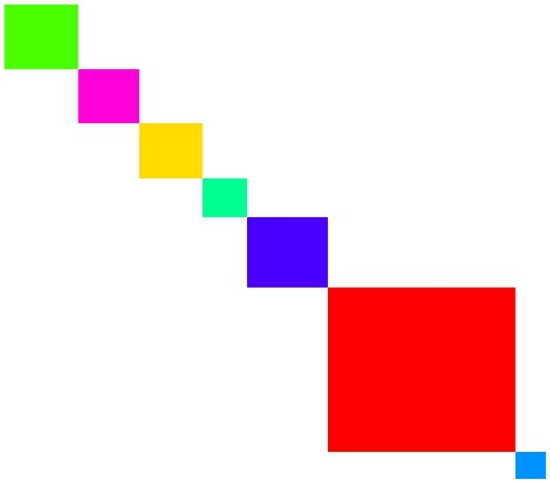


(c) $\epsilon = q_{0.5}$.

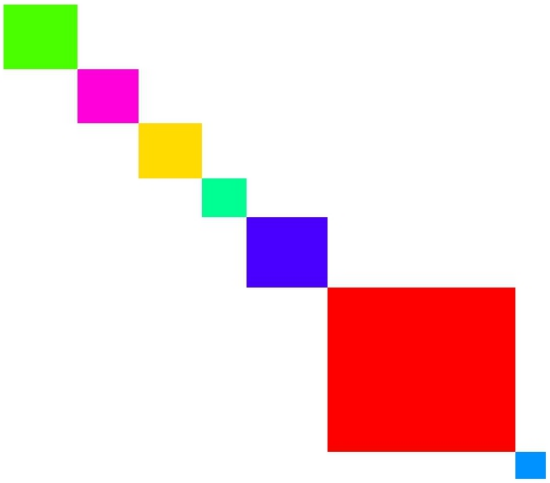


(d) $\epsilon = q_{0.99}$

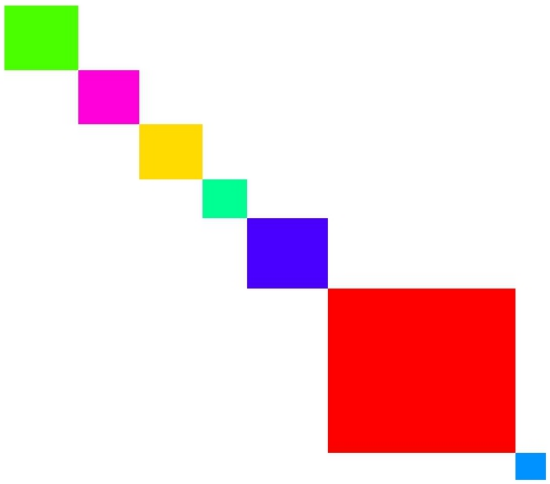
Figure 5.4: Scatterplots of the clustering estimates given by the new loss-function minimization method with KL I-divergence, for $N = 1$ and different values of ϵ. $n = 250$, second set of hyperparameters.



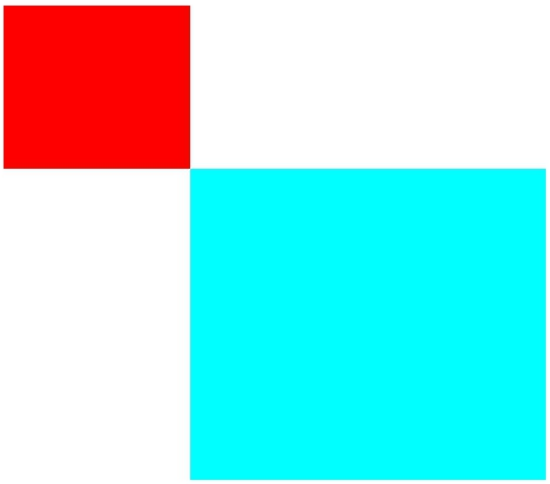
(a) $\epsilon = 0$.



(b) $\epsilon = q_{0.01}$

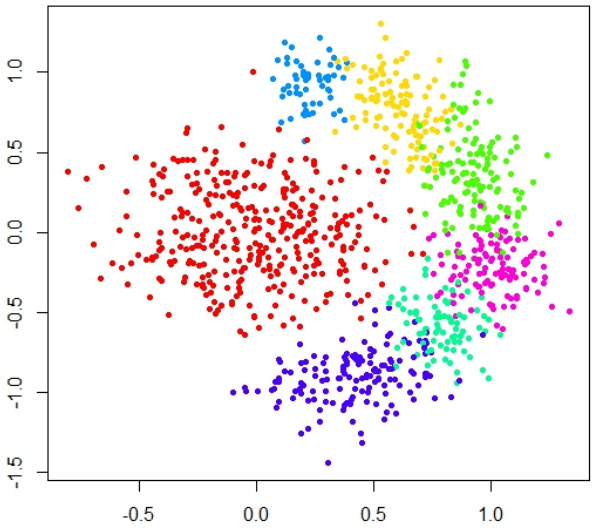


(c) $\epsilon = q_{0.5}$.

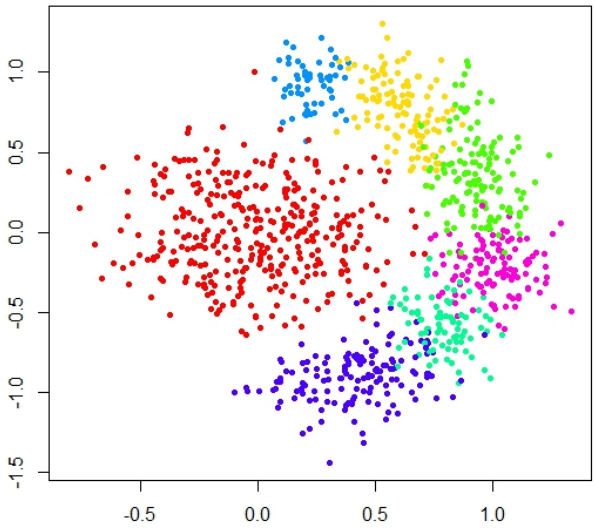


(d) $\epsilon = q_{0.99}$

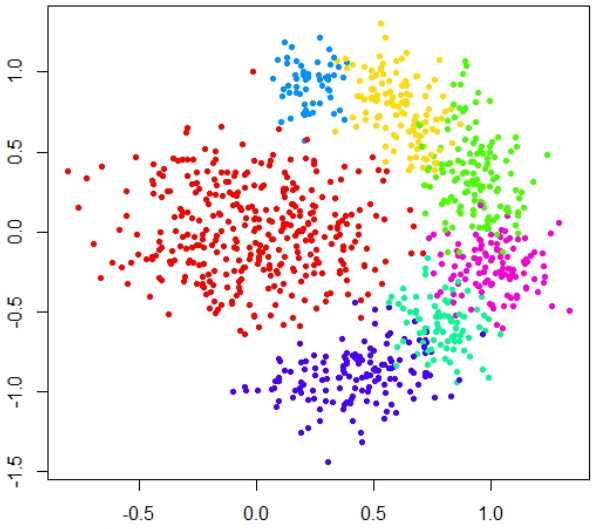
Figure 5.5: Incidence matrixes of the clustering estimates given by the new loss-function minimization method with KL I-divergence, when $n = 1000$, for $N = 1$ and different values of ϵ.



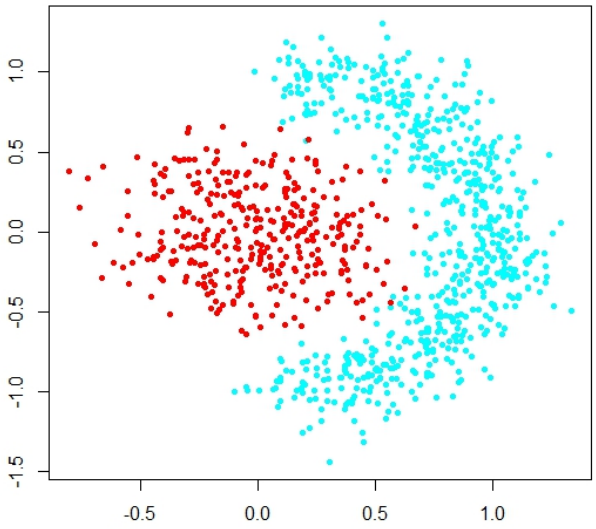
(a) $\epsilon = 0$.



(b) $\epsilon = q_{0.01}$



(c) $\epsilon = q_{0.5}$.



(d) $\epsilon = q_{0.99}$

Figure 5.6: Scatterplots of the clustering estimates given by the new loss-function minimization method with KL I-divergence, when $n = 1000$, for $N = 1$ and different values of ϵ.

5.1.2 L^2 and Hellinger distance

In this section, we will present the estimates obtained by applying the clustering methods when the distances used are the L^2 metric and the Hellinger distance. The L^2 distance between two bivariate Gaussian kernels is:

$$L^2_{ij} = \sqrt{2^{-p}\pi^{-\frac{p}{2}}(|\Sigma_i|^{-\frac{1}{2}} + |\Sigma_j|^{-\frac{1}{2}}) - 2(2\pi)^{-\frac{p}{2}}|\Sigma_i + \Sigma_j|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\Sigma_i + \Sigma_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\}} =$$
$$= \sqrt{2^{-p}\pi^{-\frac{p}{2}}(|\Sigma_i|^{-\frac{1}{2}} + |\Sigma_j|^{-\frac{1}{2}}) - 2N_p(\mathbf{0}|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j, \Sigma_i + \Sigma_j)}, \ i, j = 1, ..., n$$

where, once again, $p = 2$ represents the dimension of the data and N_p the bivariate Gaussian density. The Hellinger distance between two measurable functions is defined as the L^2 norm of the difference between their square roots. In the case of Gaussian densities we obtain:

$$H_{ij} = \sqrt{1. - |\Sigma_i|^{\frac{1}{4}}|\Sigma_j|^{\frac{1}{4}}2^{\frac{p}{2}}|\Sigma_i + \Sigma_j|^{-\frac{1}{2}} exp\{-\frac{1}{4}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\Sigma_i + \Sigma_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\}}, \ i, j = 1, ..., n$$

Due to the definition of the Hellinger distance and its relation to the L^2 norm, we choose to present the estimates provided by using these two distances together. In particular, for both these distances and for every choice of ϵ greater than 0 and $(\boldsymbol{m}_0, k_0, \nu_1, \Psi_1)$ among the ones discussed at the beginning of this chapter, we obtained the same result, which is reported in Figure 5.7, that is all the data are grouped into the same cluster. This situation is far different from the one depicted in the first plot of Figures from 5.1 to 5.6, referring to the standard loss-function minimization method ($\epsilon = 0$), in which a relevant number of clusters is identified (we recall that the standard similarity matrix is not influenced by the chosen distance). To better understand how the method works, we let ϵ vary from 0 to the value of $q_{0.01}$, on a grid of length 20. The most relevant estimates are presented in Figures 5.8 and 5.9 for the L^2 distance, and in Figures 5.10 and 5.11 for the Hellinger distance, when $n = 1000$. In particular, we can observe from Figures 5.8(a) and (b) and Figures 5.9(a) and (b) that, in the case of L^2 distance, there are no relevant changes until $\epsilon = 1.925$, which is rather close to the value of $q_{0.01}$ (which is $q_{0.01} = 2.75$). On the contrary, Figure 5.10 and 5.11 suggests that values of ϵ quite lower than $q_{0.01}$ for the Hellinger distance give more interesting estimates (in Figures 5.10(d) and 5.11(d) two clusters are found), this showing a significant difference between L^2 and Hellinger distance.

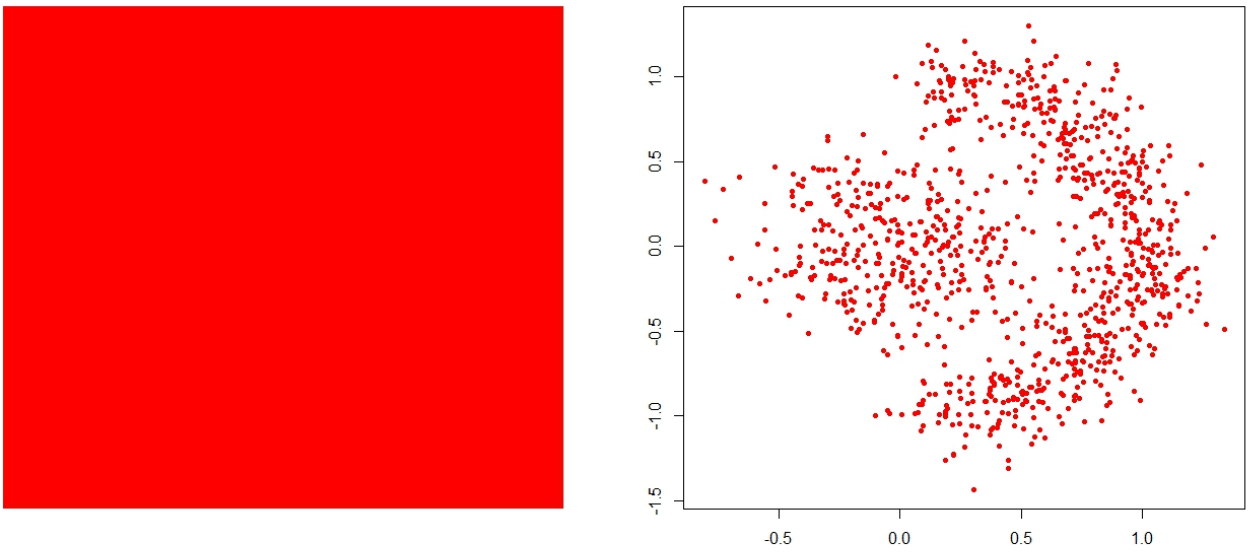
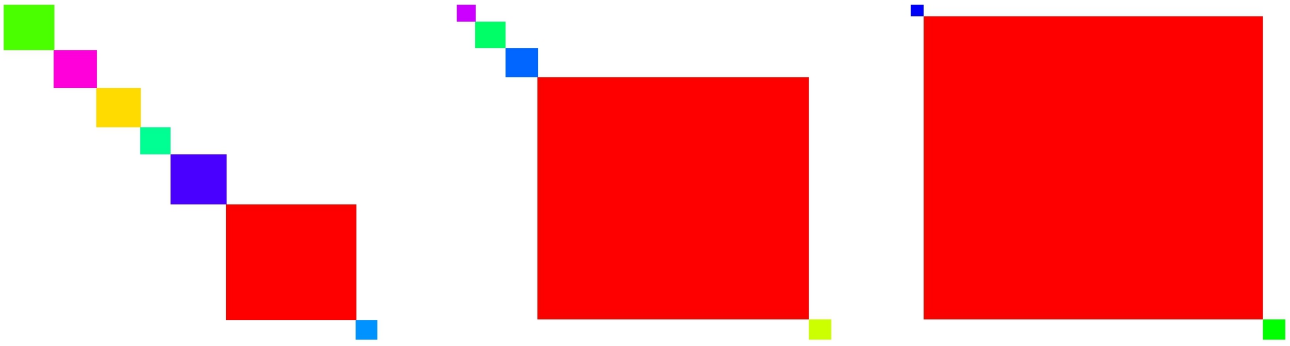
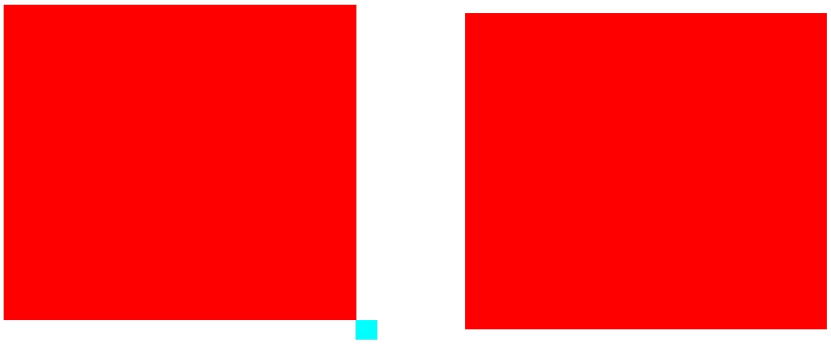


Figure 5.7: Clustering estimates provided by the new loss-function minimization method using L^2-norm or Hellinger distance.

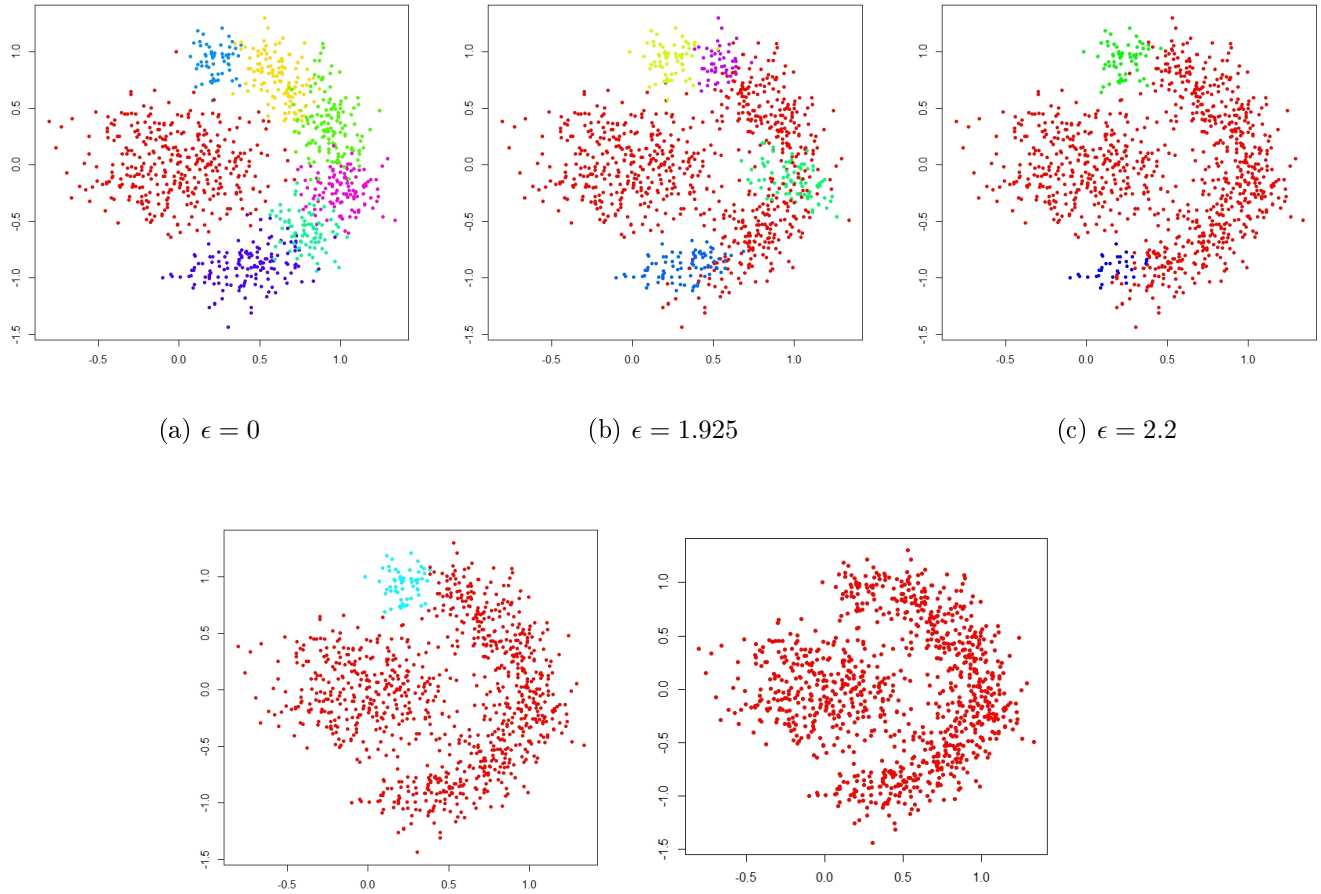


(a) $\epsilon = 0$ (b) $\epsilon = 1.925$ (c) $\epsilon = 2.2$



(d) $\epsilon = 2.475$ (e) $\epsilon = q_{0.01} = 2.75$

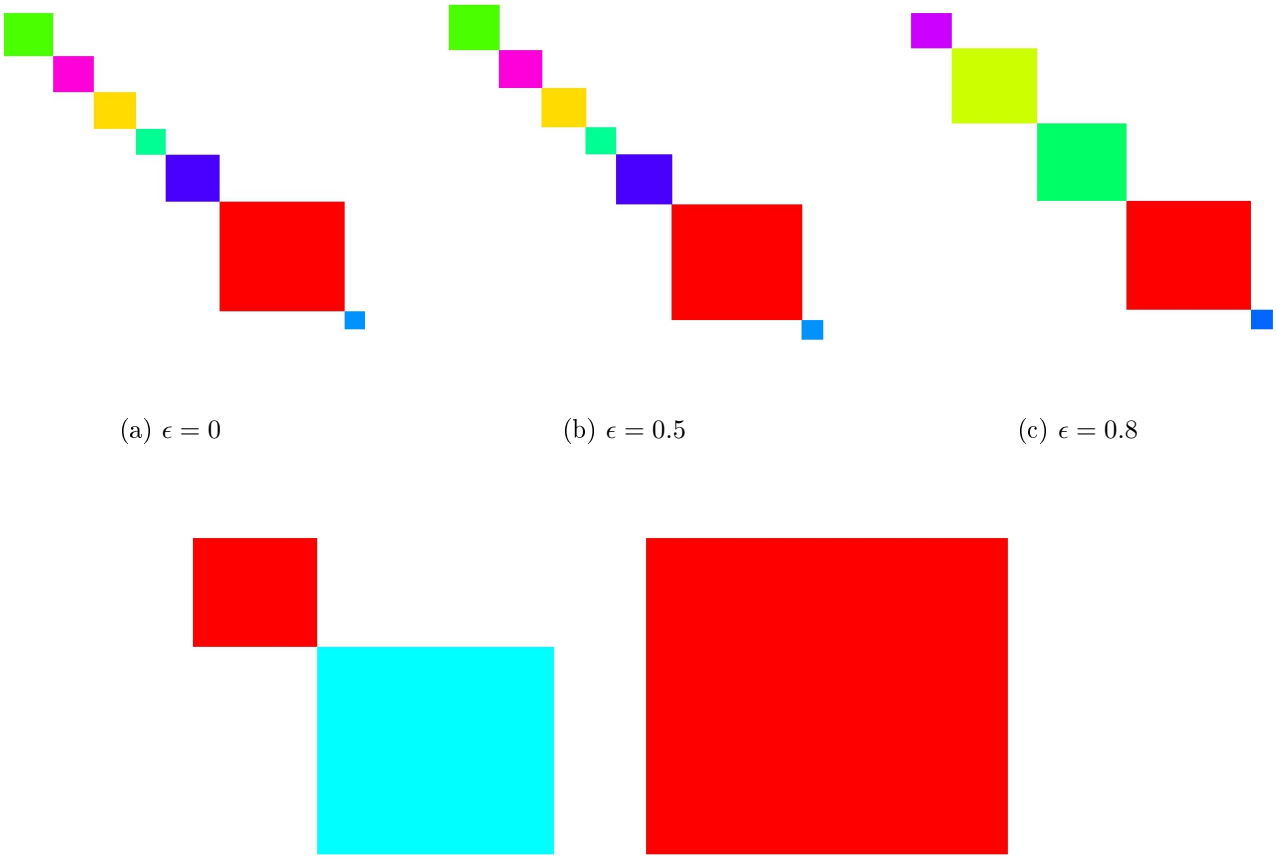
Figure 5.8: Incidence matrixes of the clustering estimates given by the new loss-function minimization method using L^2 distance, obtained for different values of ϵ.



(a) $\epsilon = 0$ (b) $\epsilon = 1.925$ (c) $\epsilon = 2.2$

(d) $\epsilon = 2.475$ (e) $\epsilon = q_{0.01} = 2.75$

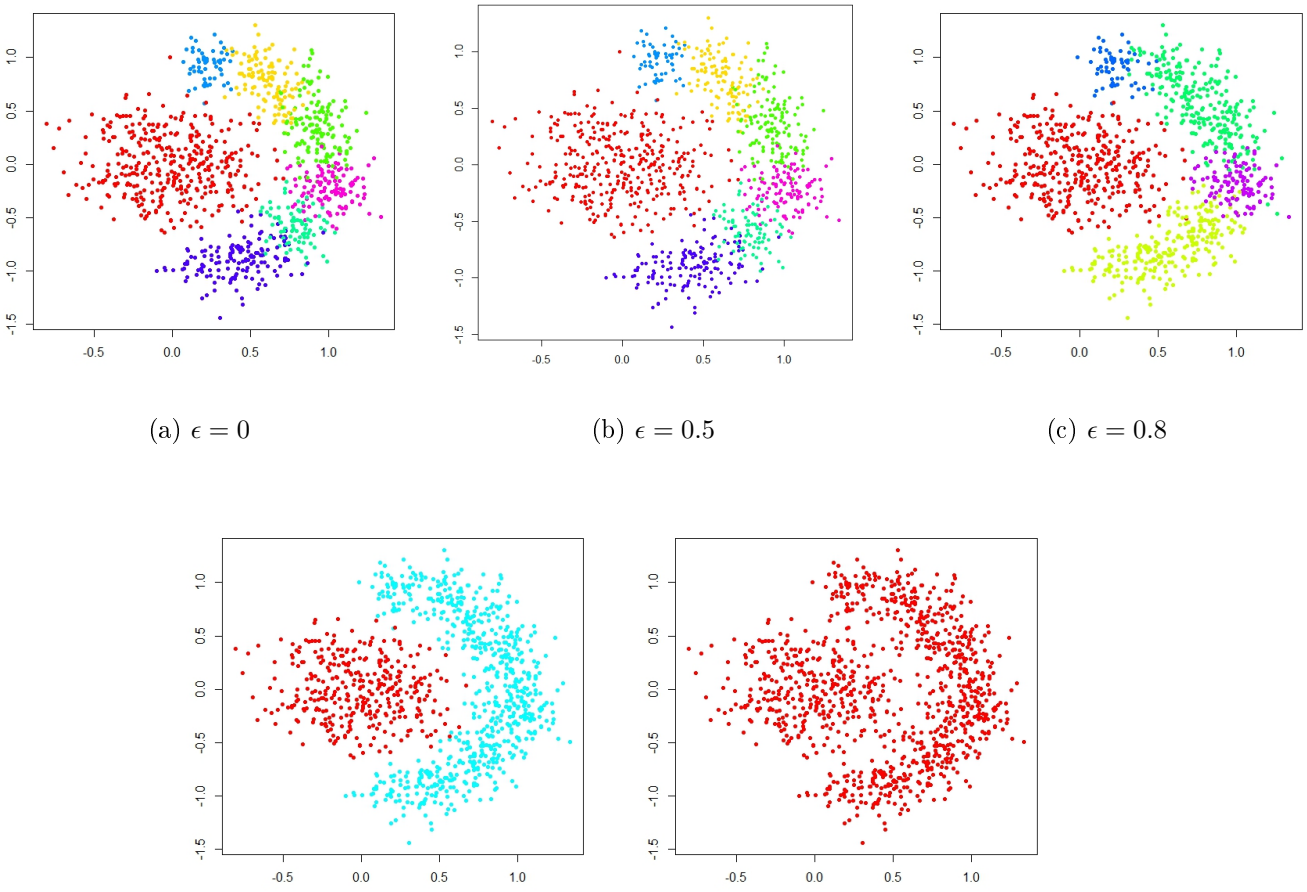
Figure 5.9: Scatterplots of the clustering estimates provided by the new loss-function minimization method using L^2 distance, obtained for different values of ϵ.



(a) $\epsilon = 0$ (b) $\epsilon = 0.5$ (c) $\epsilon = 0.8$

(d) $\epsilon = 0.9$ (e) $\epsilon = q_{0.01} = 0.99$

Figure 5.10: Incidence matrixes of the clustering estimates provided by the new loss-function minimization method using Hellinger distance, obtained for different values of ϵ.



(a) $\epsilon = 0$ (b) $\epsilon = 0.5$ (c) $\epsilon = 0.8$

(d) $\epsilon = 0.9$ (e) $\epsilon = q_{0.01} = 0.99$

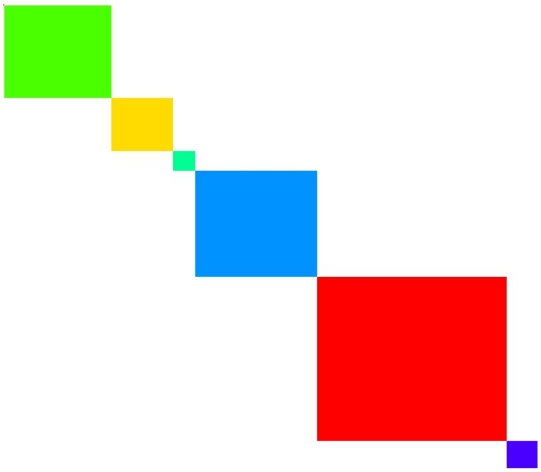
Figure 5.11: Scatterplots of the clustering estimates given by the new loss-function minimization method using Hellinger distance, obtained for different values of ϵ.

5.1.3 Varying the value of $\hat{K}$

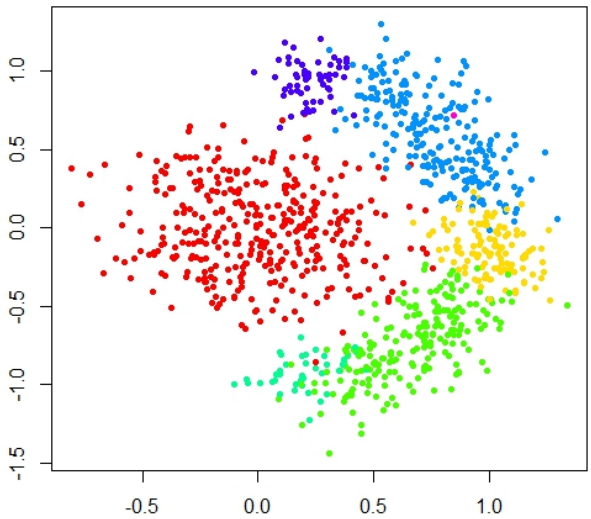
So far, we only illustrated loss-function minimization methods involving a particular choice of the misclassification costs. In particular, we fixed the parameter $\hat{K} = \frac{b}{a+b}$ equal to 0.5. This value corresponds to assign equal costs to the two kind of misclassification errors that may occur, and gives the same estimates provided by the quadratic loss-function method proposed by Dahl (2006). To recover a wider range of possible interpretations of the costs, we present in this section some estimates when the value of $\hat{K}$ is slightly different from 0.5, in particular fixed to 0.25 or 0.75. These new values assign different costs to the two misclassification errors, reflecting many practical situations.

In Figures 5.12 to 5.20 the estimates provided by the application of the loss-function minimization methods when $n = 1000$ are displayed. Plots are grouped according to the value of ϵ, which is fixed equal to 0, $q_{0.01}$ and $q_{0.99}$, for the three distances analyzed in this section. Figures 5.13 to 5.20 are also grouped according to the value of $\hat{K}$. We recall that, in the case of $\epsilon = 0$, the estimates are not influenced by the choice of the distance we use.

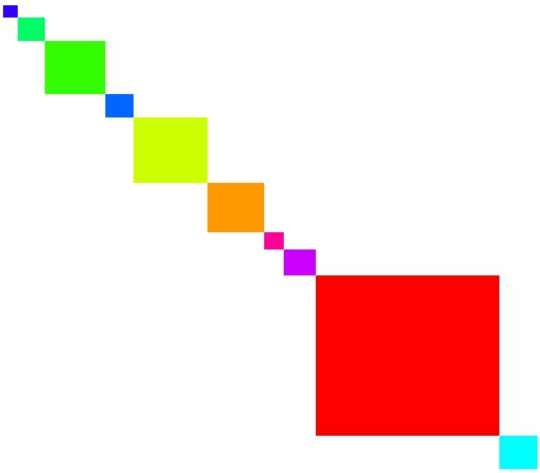
From the estimates presented in this section, we can argue that the method is very robust with respect to the value assigned to the parameter $\hat{K}$. In fact, all the estimates are very similar to the ones presented in the previous sections (as an example, compare Figure 5.12 with Figures 5.5(a) and 5.6(a)).



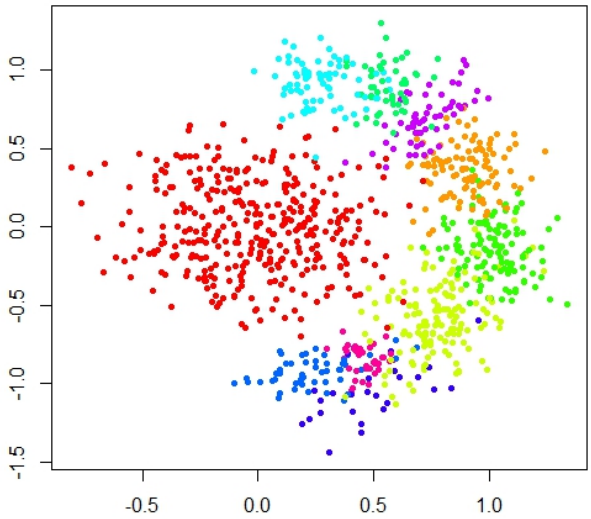
(a) $\hat{K} = 0.25$.



(b) $\hat{K} = 0.25$.



(c) $\hat{K} = 0.75$.

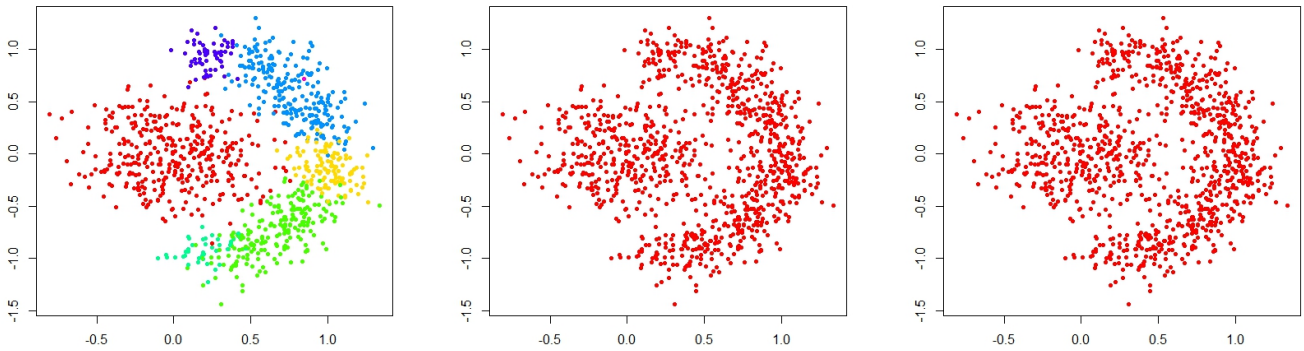


(d) $\hat{K} = 0.75$.

Figure 5.12: Clustering estimates provided by the new loss-function minimization method for different values of $\hat{K}$, obtained for $\epsilon = 0$ (standard similarity matrix) for the simulated dataset with $n = 1000$ observations.

(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.13: Incidence matrixes of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.25$, obtained for $\epsilon = q_{0.01}$ for the simulated dataset with $n = 1000$ observations.

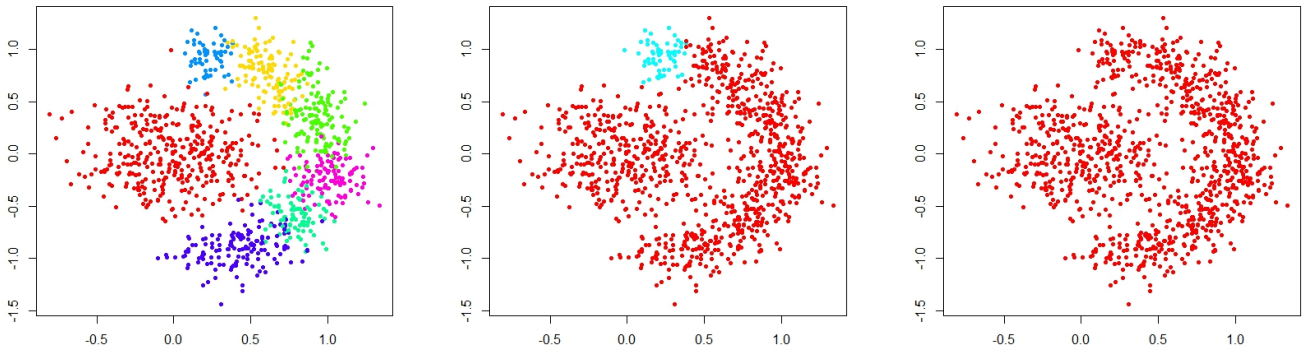


(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.14: Scatterplots of the clustering estimates given by the loss-function minimization method for different values of $\hat{K} =$, obtained for $\epsilon = q_{0.01}$ for the simulated dataset with $n = 1000$ observations.

(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.15: Incidence matrixes of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.75$, obtained for $\epsilon = q_{0.01}$ for the simulated dataset with $n = 1000$ observations.

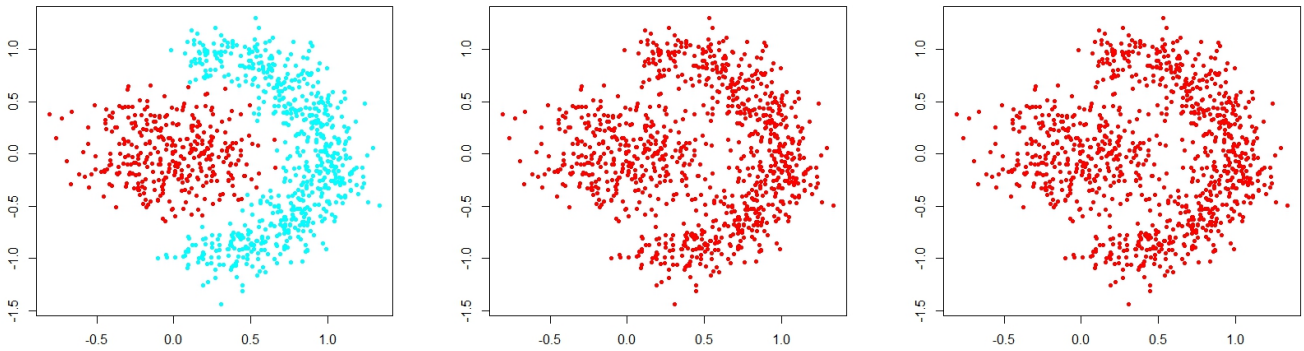


(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.16: Scatterplots of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.75$, obtained for $\epsilon = q_{0.01}$ for the simulated dataset with $n = 1000$ observations.

(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.17: Incidence matrixes of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.25$, obtained for $\epsilon = q_{0.99}$ for the simulated dataset with $n = 1000$ observations.

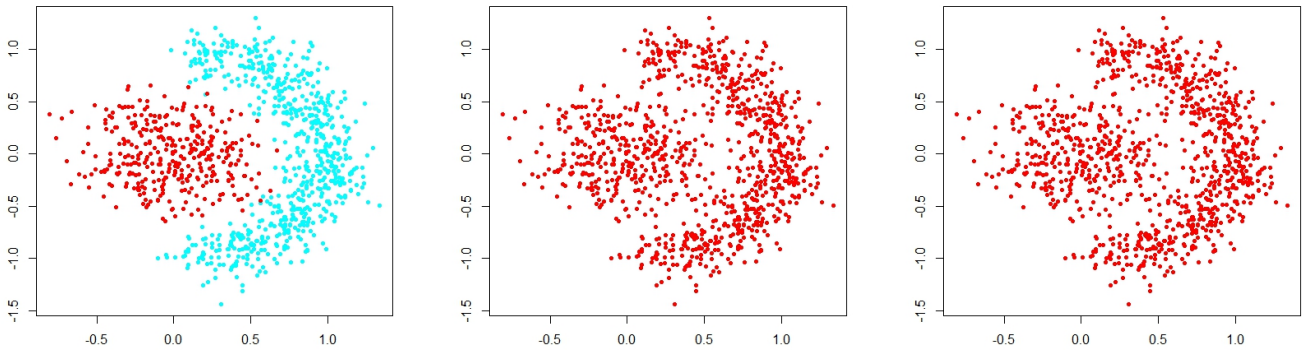


(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.18: Scatterplots of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.25$, obtained for $\epsilon = q_{0.99}$ for the simulated dataset with $n = 1000$ observations.

(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.19: Incidence matrixes of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.75$, obtained for $\epsilon = q_{0.99}$ for the simulated dataset with $n = 1000$ observations.



(a) KL I-divergence. (b) L^2 distance. (c) Hellinger distance.

Figure 5.20: Scatterplots of the clustering estimates provided by the loss-function minimization method for $\hat{K} = 0.75$, obtained for $\epsilon = q_{0.99}$ for the simulated dataset with $n = 1000$ observations.

5.2 Dealing with misclassifications

Since in this chapter we have considered simulated datasets, we know the true classification of the observations, i.e. corresponding to the data labelling from the data sampling scheme. The dataset was simulated from a mixture of two bivariate Gaussian kernels, that means the data come from two different clusters. In order to compare this information with the clustering estimates, we have to select those estimating the two clusters. In this sections we will consider the dataset with size $n = 1000$. As far as the choice of the distance and the value of ϵ is concerned, we refer to Figures 5.10(d) and 5.11(d) for the Hellinger distance ($\epsilon = 0.9$) and Figures 5.5(d) and 5.6(d) for the Kullback-Leibler I-divergence ($\epsilon = q_{0.99} = 3.25$).

We observe that the elements being misclassified by the estimate given by the new clustering method are the same in both the situations taken into account here, i.e. for the two different distances involved. In Table 5.1 a summary of the misclassification error is reported: we found that 337 points in cluster 1 and 644 points in cluster 2 were correctly classified; the misclassification rate is 1.9%. In Figure 5.21 the two classifications of the observations are displayed, one according to two different colours, and one described by different points.

	$Estimated1$	$Estimated2$
True 1	337	13
True 2	6	644

Table 5.1: Summary of the true and estimated clusterings.

As we can see, the misclassified elements lie in the middle of the two main groups. In order to better understand the relevance of the misclassified elements in the estimated partition, we should be able to evaluate the probability that these elements belong to the estimated cluster. Such probability is not well defined because of the label switching problem. In fact, labels continuously change during the Polya urn sampling algorithm, and therefore it is impossible to identify a cluster in relation to a clustering estimate. Nevertheless, we can evaluate the probability of a new observation falling into the same cluster of a given observation, say $\boldsymbol{x}_i$, such as a misclassified one, conditionally on the observations. In formulas, we have:

$$\frac{\mathbb{P}(\boldsymbol{X}_{n+1} \leftrightsquigarrow \boldsymbol{x}_i, \boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{x})}{\mathbb{P}(\boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{x})}, \tag{5.1}$$

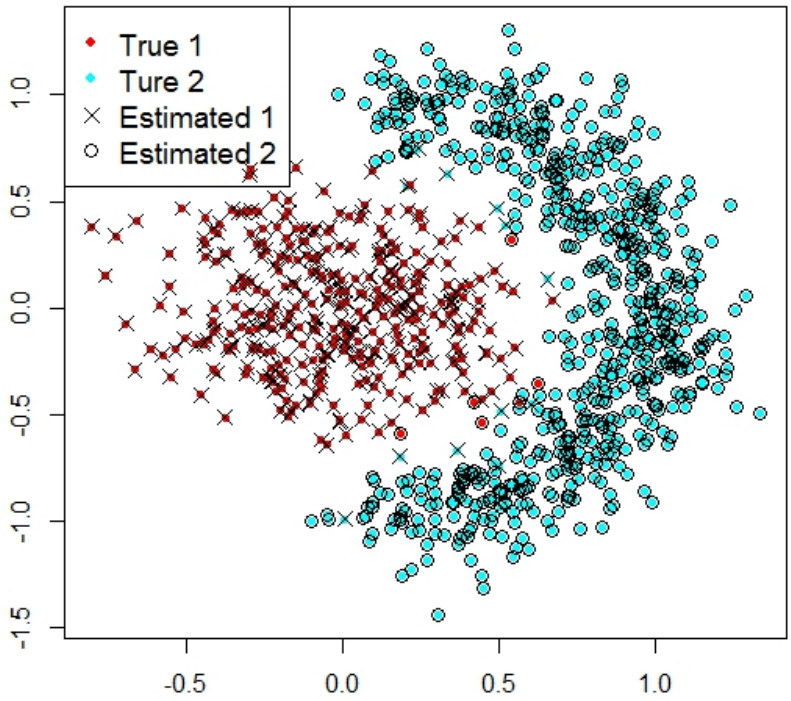


Figure 5.21: Misclassification.

where $\leftrightsquigarrow$ denotes that two elements are in the same cluster and $d\boldsymbol{y}$ stands for an infinitesimal interval of the sample space (in this case $\mathbb{R}^2$). The term $\boldsymbol{X}_{n+1} \in d\boldsymbol{y}$ gives more information and help in better understanding the probability that is going to be evaluated. Moreover, notice that the denominator is the predictive probability, and then, when $d\boldsymbol{y} \to 0$, we can write $\mathbb{P}(\boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{x}) \approx f_{\boldsymbol{X}_{n+1}}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y}$. As far as the numerator is concerned, we can write:

$$\mathbb{P}(\boldsymbol{X}_{n+1} \leftrightsquigarrow \boldsymbol{x}_i, \boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{x}) = \int_{\Theta\times\Theta^n} \mathbb{P}(\boldsymbol{X}_{n+1} \leftrightsquigarrow \boldsymbol{x}_i, \boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}, \boldsymbol{x})\mathcal{L}(d\boldsymbol{\theta}_{n+1}|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta}|\boldsymbol{x})$$

$$= \int_{\Theta\times\Theta^n} \underbrace{\mathbb{P}(\boldsymbol{X}_{n+1} \leftrightsquigarrow \boldsymbol{x}_i|\boldsymbol{X}_{n+1} \in d\boldsymbol{y}, \boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}, \boldsymbol{x})}_{P_1} \underbrace{\mathbb{P}(\boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}, \boldsymbol{x})}_{P_2} \mathcal{L}(d\boldsymbol{\theta}_{n+1}|\boldsymbol{\theta})\mathcal{L}(d\boldsymbol{\theta}|\boldsymbol{x}).$$

Thanks to the independence, conditionally on the observations, between the new latent variable and the vector $\boldsymbol{\theta}$, we have that the probability P_2 is approximately the kernel density associated with the new latent variable $\boldsymbol{\theta}_{n+1}$, in fact, when $d\boldsymbol{y} \to 0$, we have $P_2 = \mathbb{P}(\boldsymbol{X}_{n+1} \in d\boldsymbol{y}|\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}, \boldsymbol{x}) \approx k(\boldsymbol{y}|\boldsymbol{\theta}_{n+1})d\boldsymbol{y}$. Concerning the probability P_1, we have that it is equal to 1 in the case that the two latent variables are in the same class according to the new equivalence relation, (referring

to Section 2.5, $d_{n+1,i}(\epsilon) = 1$), and 0 otherwise. Thus, this probability can be written as $\mathbb{I}_{\{d_{n+1,i}(\epsilon)=1\}}(d\boldsymbol{\theta}_{n+1})$. Finally, to compute the ratio in (5.1), we used Monte Carlo simulation. In Table 5.2, the estimated probabilities for four selected misclassified elements are presented, whose position is shown in Figure 5.22. According to the true classification of the elements, points A and B are supposed to be in cluster 2 (light blue in Figure 5.21), while C and D should be in cluster 1 (red in Figure 5.21). All these points are erroneously clustered in the estimated partition. We recall that the misclassified elements are the same in the case of Hellinger distance or Kullback-Leibler I-divergence.

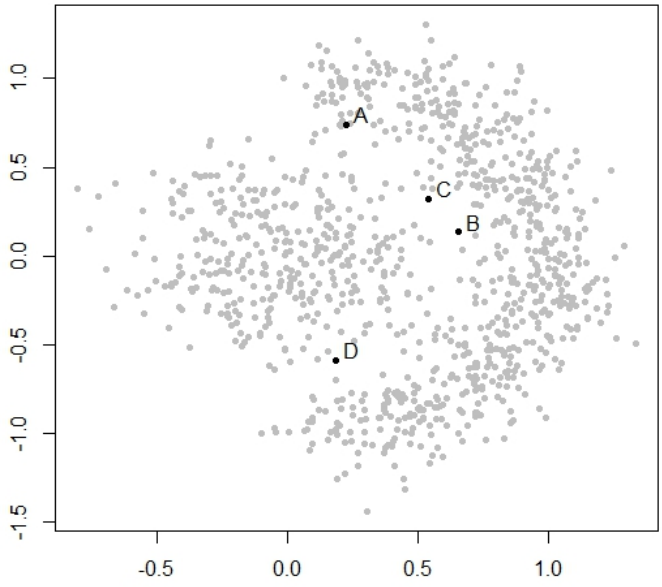


Figure 5.22: Location of the misclassified elements for which posterior probabilities are computed.

	A	B	C	D
Kullback-Leibler	0.8933	0.4844	0.5378	0.8415
Hellinger	0.9215	0.5613	0.5854	0.8662

Table 5.2: Posterior estimated expected values of the probability of being in the same cluster of $\boldsymbol{x}_i$, for four selected values of i.

The mean values of the probability we are interested in are quite large for some elements, indicating less uncertainty about their classification.

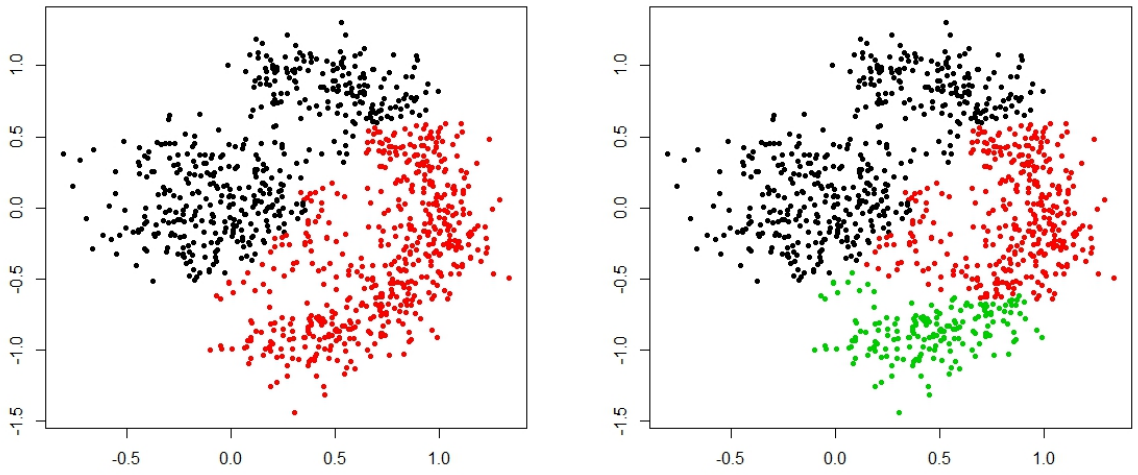
5.3 Application of some heuristic techniques for clustering

In this section we will show the cluster estimates provided by some of the heuristic clustering methods presented in Section 1.1, applied to the bivariate simulated dataset, when $n = 1000$, as the agglomerative hierarchical clustering, the K-means algorithm and the original formulation of the DBSCAN algorithm, as presented in Ester et al. (1996).

5.3.1 Agglomerative hierarchical clustering

As mentioned in Section 1.1, the hierarchical clustering technique is one of the most popular ones in the family of heuristic clustering. This method performs a sequential union of the observations following a decision rule based on a matrix of dissimilarities between the data, which in this case is represented by the matrix of Euclidean distances, and on a proper distance between clusters (called linkage). The most popular linkages are the single linkage, the complete linkage and the average linkage. After that a dendrogram is created (i.e., a scheme describing all the iterations of the algorithm), a partition is chosen, cutting the dendrogram in order to obtain a given number of clusters.

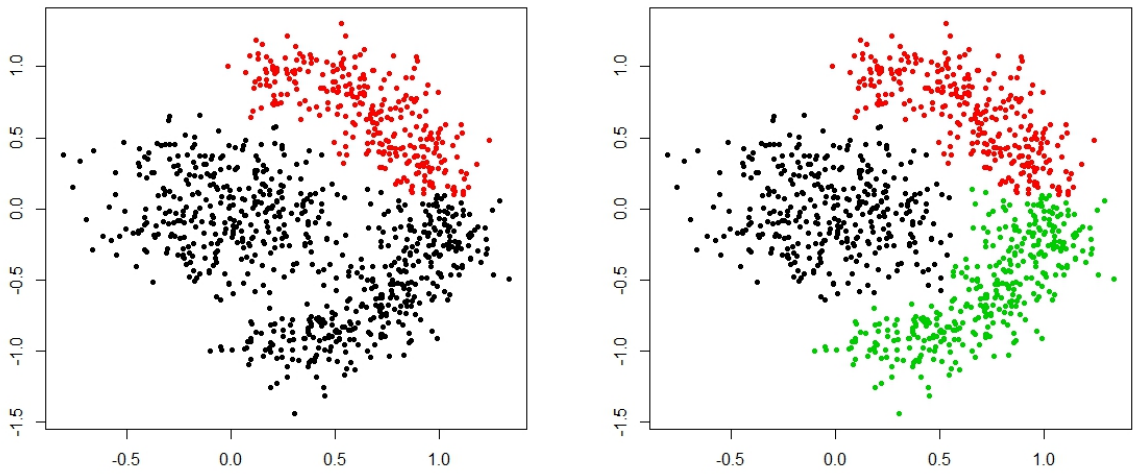
Figures 5.23 and 5.24 show the partitions resulting from cutting the dendrogram at 2 and 3 clusters, in the case of complete and average linkage. As we can see, the choice of the linkage strongly influences the outcomes of the algorithm. In particular, the average linkage is the one that better picks the subdivision between the central group and the semicircular region on the right of it.



(a) $K_n = 2$.

(b) $K_n = 3$.

Figure 5.23: Agglomerative hierarchical clustering applied to the dataset with $n = 1000$ observations (Complete Linkage).



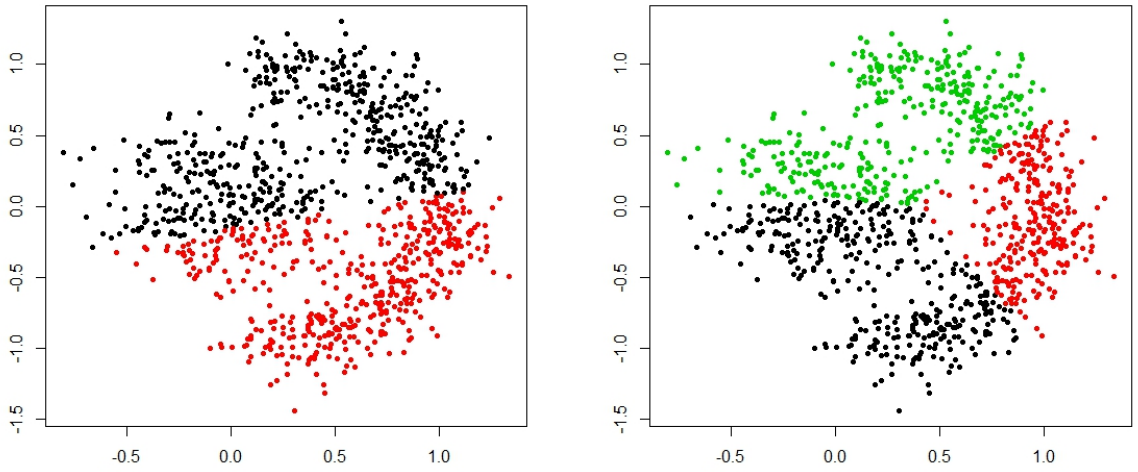
(a) $K_n = 2$.

(b) $K_n = 3$.

Figure 5.24: Agglomerative hierarchical clustering applied to the dataset with $n = 1000$ observations (Average Linkage).

5.3.2 K-means clustering

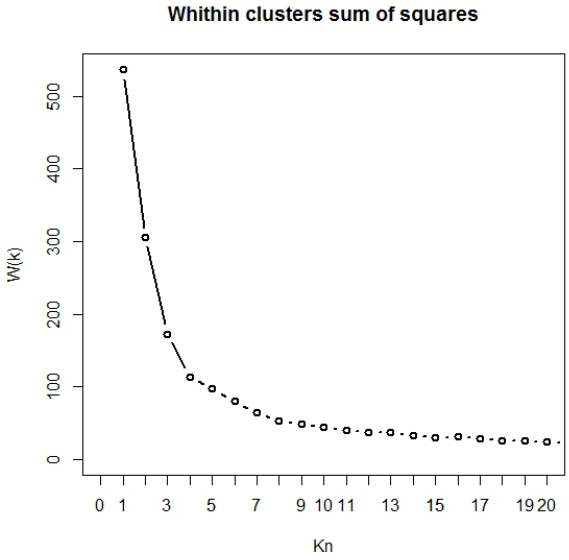
The K-means clustering technique aims at finding the partition that minimizes the sum of the squares between observations and a given number of means. So, this algorithm takes as input the number of clusters and the initial centers. We recall that, thanks to the convergence of the algorithm, the location of the initial centers does not affect the result. Concerning the number of clusters, we choose it according to a standard method based on the curve of the "within-cluster sum of squares", defined as $W_k = \sum_{j=1}^{k} \sum_{\boldsymbol{x}_i \in C_j} ||\boldsymbol{x}_i - \boldsymbol{o}_j||^2$, where k is the number of clusters, C_j, for $j = 1, ..., k$ is a subset of $\{1, ..., n\}$ such that $\bigcup_{j=1}^{k} C_j = \{1, ..., n\}$ and o_j, for $j = 1, ..., k$ is the centroid associated with the cluster C_j. Notice that this function is the same that is minimized during the K-Means algorithm, with respect to the possible partitions of $\{1, ..., n\}$. The function W_k is shown in Figure 5.26. The standard method for choosing the value of k proposes to identify an elbow in the graph of W_k, and to choose K_n as the corresponding number of clusters, representing the optimal partition. From Figure 5.26, we choose $K_n = 2$ and 3. Figure 5.25 displays the corresponding estimates, and it is clear that the moethod divides the region into K_n parts, described by three lines, which do not recover the original partition provided by the simulation of the data (for example, the central group of observations is always divided into two parts).



(a) $K_n = 2$.

(b) $K_n = 3$.

Figure 5.25: K-Means clustering applied to the dataset with $n = 1000$ observations.



(a)

Figure 5.26: Within clusters sum of squares for K-Means clustering. Dataset with $n = 1000$ observations.

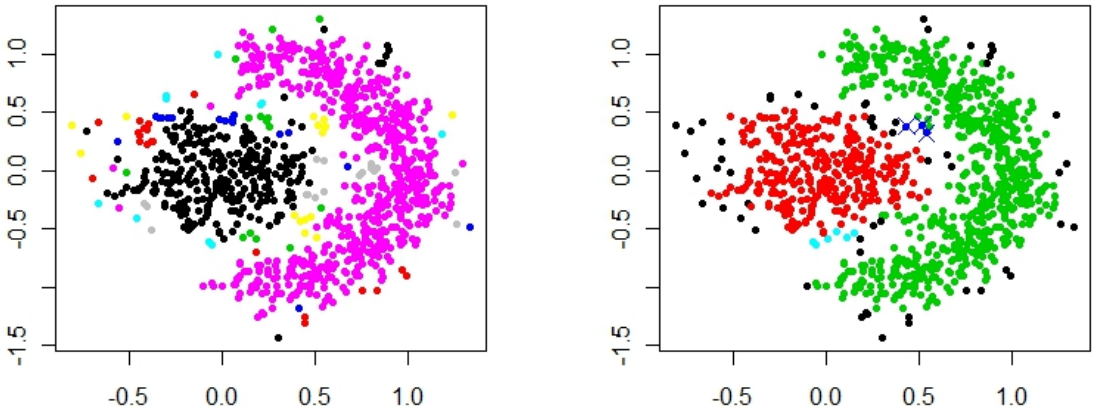
5.3.3 DBSCAN algorithm

As described in Section 1.1, the DBSCAN algorithm is a heuristic clustering method that unifies elements close to each other, and is able to locate dense group of observations. This algorithm is the key point to define the new loss-function minimization method proposed in this work (see Section 2.5). In this section, we apply its original definition to the simulated bivariate dataset, in order to compare the outcomes. Differently from the previous sections, here we fix the value of N not only to 1, but also to 6. We recall that, when $N > 1$, the partitions are not uniquely determined, because the definition used do not define an equivalent relation. Furthermore, when $N > 1$, noise elements can be identified by the algorithm. In Figure 5.27 we report clustering estimates for different values of N and ϵ. We tuned the value of ϵ in order to find the subdivision of the data which better preserves the two main groups.

When $N = 1$ (Figure 5.27(a)), noise elements are not allowed, and every singleton could represent a cluster. This is the reason why so many different clusters are identified by the method (colours are repeated). Of course, this partition does not seem to be satisfying, if compared with previous results.

Differently, when $N = 6$, less clusters are found, but many noise elements are located (black points in Figure 5.27(b)). The main reason why this happens is that no model is adopted in order to define the DBSCAN algorithm, s points generated from the tails of the true distribution

are not included into the clusters. This limitation is a serious drawback of the class of "density-based" algorithms, which can exclude some elements that are "far" from a cluster in term of distance. As an example of the non-uniqueness of the partition found by the method when $N > 1$, takes the crossed blue cluster just above the red central group, which is composed of only 3 elements. We recall that N is the minimum number of elements required to call a group cluster, so this is clearly not a cluster. The ambiguity arises since, when $N > 1$, the symmetry of the relation defined by the DBSCAN algorithm does not hold, not leading to an equivalence relation, and therefore uncertainty situations can occur.



(a) $N = 1$, $\epsilon = 0.075$.

(b) $N = 6$, $\epsilon = 0.1$.

Figure 5.27: DBSCAN algorithm applied to the dataset with $n = 1000$ observations.

5.3.4 Final Considerations

The cluster estimates presented in this chapter show that the method based on the new similarity matrix is strongly influenced by the choice of the distance. In fact, when ϵ is fixed, Kullback-Leibler I-divergence and L^2 distance (or Hellinger distance) give very different estimates. Nevertheless, similar clustering estimates are given if we consider different combinations of ϵ values and distances (compare, for instance, Figures from 5.8 to 5.11).

Despite the strong influence of the choice of the distance, robust features of the method holds with respect to the choice of the hyperparameters of P_0 (above all, γ_1 and γ_2 in the prior distribution of a) and of the parameter $\hat{K}$ (the costs of misclassification). In fact, varying these

hyperparameters, clustering estimates turned out to be very similar.

As far as the choice of the value of ϵ concerns, we tried to fix it according to our prior beliefs, that is to assign it according to the values of some quantiles of the prior distributions of the distances between the parametric kernel densities associated with the prior latent vectors. Once again, we obtained very different estimates when different distances are taken into account. In particular, for the L^2 and the Hellinger distances, we found that prior quantiles correspond to very high values of the prior distances (for example, in the case of the Hellinger distance we have $q_{0.01} = 0.99$, when its maximum is 1), and so the method provides estimates with only one cluster. For this reason, we carried out further analysis, exploring a more vast range of values for ϵ (see Figures 5.8 to 5.11). Differently, when dealing with the Kullback-Leibler I-divergence, different values of ϵ lead to estimates very different from each other, and some of them seem to explain fairly well the particular shape of the data disposition (see, for instance, Figures 5.5(d) and 5.6(d)).

Finally, hierarchical and K-Means clustering do not provide estimates close to the true partition (see Figures 5.23, 5.24 and 5.25). Furthermore, the standard DBSCAN method is not able to determine a clear partition, even if no noise is allowed (see Figure 5.27(a)) since, when $N = 1$, too many clusters are found. On the other hand, when $N = 6$, the DBSCAN partition is not uniquely determined, as we discussed in Section 2.5. In contrast, the proposed new method often gives the true partition, or at least a reasonable one.

Chapter 6

Posterior sampling and density estimation

To apply the clustering methods presented in this work, we need to provide samples from the posterior distribution of the random partition. To do so, exploiting the definition of DPM models, proper algorithms to sample from the distribution of the latent variables $p(\boldsymbol{\theta}|\boldsymbol{x})$ are used. In the case of conjugate models (such as for Galaxy data in Chapter 3 and for the bivariate normal simulated dataset in Chapter 5), a Polya Urn sampling scheme has been adopted (see Neal, 2000), while, when non-conjugate priors are chosen (as in the Kevlar data), the sampling algorithm is that proposed in Argiento et al. (2010).

In this chapter, we proceed with an introduction to the general Polya Urn scheme algorithm, used to sample from the posterior distribution of the latent variables when the model is conjugate. We recall that all the algorithm for sampling have been coded using R (for the univariate case, with Dirichlet or NGG process prior) and C (for the bivariate case, with Dirichlet process prior having random mass parameter). After the first part concerning the sampling algorithm, density estimates will be shown, for both the case of univariate (Galaxy dataset in Chapter 3) and bivariate (simulated dataset in Chapter 5) Gaussian kernels.

6.1 Polya Urn scheme

As mentioned in (2.1), a sample from a Dirichlet process prior can be characterized as a generalized Polya urn; this result can be easily extended to the case of NGG process prior, the details will be reported in this chapter. Therefore, the same representation can be used to describe the predictive laws of the latent variables in a Dirichlet or a NGG process mixtures model.

Recalling the notation of the DPM model in (2.7), if $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ is a sample from a Dirichlet process prior, the joint law of $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$ can be described as:

$$\begin{aligned} \boldsymbol{\theta}_1 &\sim P_0 \\ \boldsymbol{\theta}_i|\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{i-1} &\sim \frac{a}{a+i-1}P_0 + \frac{1}{a+i-1}\sum_{j=1}^{i-1}\delta_{\boldsymbol{\theta}_j} \text{ for } i = 1, ..., n, \end{aligned} \tag{6.1}$$

where, as usual, $P_0(\cdot)$ represents the mean distribution of the nonparametric prior, and a is the total mass parameter (see Chapter 2 or Blackwell and MacQueen, 1973). This means that a sample from a Dirichlet process can be described as a sequential extraction from a Polya Urn, where the possible outcomes range in "a continuum of colours".

What we need in order to set up a Gibbs sampler algorithm is the expressions of the full conditionals of the interested variables, that is, the distributions of $\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{x}$, where $-i$ stand for the vector of the latent variables excluding the i-th. Thanks to Bayes theorem and the independence of each observation $\boldsymbol{x}_i$ from $\boldsymbol{\theta}_{-i}$, we obtain:

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{x}) \propto p(\boldsymbol{x}_i|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}) = k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)(\frac{a}{a+i-1}P_0 + \frac{1}{a+i-1}\sum_{j=1}^{i-1}\delta_{\boldsymbol{\theta}_j}).$$

The first term of the last expression is well known from the definition of the DPM model (see Section 2.3.2), representing the kernel density associated with the i-th observation, while the second part goes back to formula (6.1).

Finally, setting:

$$\begin{aligned} q_{0i} &= \int_\Theta k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)P_0(d\boldsymbol{\theta}_i) \\ q_i &= \int_\Theta \sum_{j\neq i} k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)\delta_j(d\boldsymbol{\theta}_i) = \sum_{j\neq i} k(\boldsymbol{x}_i|\boldsymbol{\theta}_j) \end{aligned} \tag{6.2}$$

we obtain:

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{x}_i) = \omega_0\frac{k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)P_0(d\boldsymbol{\theta}_i)}{q_{0i}} + \omega_1\frac{\sum_{j\neq i}k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)\delta_{\boldsymbol{\theta}_j}(d\boldsymbol{\theta}_i)}{q_i}, \tag{6.3}$$

where ω_0 and ω_1 represent the weights of the normalized mixture, that is:

$$\omega_0 = \frac{aq_{0i}}{aq_{0i} + q_i}, \omega_1 = \frac{q_i}{aq_{0i} + q_i}$$

In the case of Gaussian kernel densities with normal inverse-gamma mean distribution P_0, the absolutely continuous part of the full conditional turns out to be a normal inverse-gamma distribution as well. At the same time, the discrete part of the distribution represents the possibility to assume values already sampled in the $i-1$-th previous extractions from the Polya urn.

So, the Gibbs sampling algorithm proceeds as follows:

Step 1 for (i in 1:n) do:

Sampling from the full conditional $\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{x}$, using formula (6.3);

Step 2 for (iter in 1:number of iterations) do:

Sampling of additional parameters (such as random mass parameter a), conditionally on $\boldsymbol{\theta}$ and $\boldsymbol{x}$;

This sampling scheme is of fundamental importance, not only to provide samples from the posterior distribution of the latent variables $\boldsymbol{\theta}$, necessary to apply the proposed clustering methods, but also to deal with density estimation. Of a particular interest is the computation of the predictive density of a new observation, conditionally to the others, which can be written as:

$$f_{\boldsymbol{X}_{n+1}|\boldsymbol{X}}(\boldsymbol{x}_{n+1}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{x}}[f(\boldsymbol{x}|\boldsymbol{\theta})] = \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}_{n+1}|\boldsymbol{\theta}) p(d\boldsymbol{\theta}|\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \int_{\Theta} f(\boldsymbol{x}_{n+1}|\boldsymbol{\theta}_{n+1}) p(d\boldsymbol{\theta}_{n+1}|\boldsymbol{\theta}) p(d\boldsymbol{\theta}|\boldsymbol{x})$$

where Θ and $\boldsymbol{\Theta}$ represent the sampling spaces of the latent vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{n+1}$, respectively. The outer integral can be computed through a Monte Carlo simulation, using the samples generated from the Gibbs sampling algorithm above described. As far as the inner integral concerns, it is of the same form of (6.2). Concluding, we can write:

$$f_{\boldsymbol{X}_{n+1}|\boldsymbol{X}}(\boldsymbol{x}_{n+1}|\boldsymbol{x}) \approx \frac{1}{G}(\sum_{g=1}^{G} \frac{a}{a+n} q_{0n+1}^{(g)} + \frac{1}{a+n} q_{n+1}^{(g)})$$

where $q_{0n+1}^{(g)}$ and $q_{n+1}^{(g)}$ stands for the same integrals as in (6.2), but referring to a sample of size n rather than $i-1$, and $g = 1, ..., G$ represent the iterations of the Gibbs sampling algorithm. This sampling scheme has been firstly proposed by Escobar and West (1995). For all the choices of nonparametric prior and hyperparameters we ran the Gibbs sampler with a burn-in of 50.000 iterations, saving samples having length of 5.000 iterations, with a thinning of 15. Concerning the convergence analysis of the MCMC samples, we obtained very good results.

In the next sections, details about the different sampling scheme and density estimations resulting from choosing a different nonparametric prior will be given.

6.2 Dirichlet Process

In order to implement the Gibbs sampling algorithm, we need to compute the values ω_0 and ω_1, which directly depend on q_{0i} and q_i defined in (6.2). In this section, characterizations of these quantities will be given, when the nonparametric prior is a Dirichlet process. To give a general result, the p-variate case will be analyzed.

As far as the quantity q_i concerns, it is very easy to compute, being the sum of $(n-1)$ evaluation of normal multivariate densities. Dealing with q_{0i} is slightly more difficult. In fact, we have to compute the following multivariate integral:

$$q_{0i} = \int_{\Theta} k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)P_0(d\boldsymbol{\theta}_i) = \int_{\mathbb{R}^p \times \mathcal{M}^p} N_p(\boldsymbol{x}_i|\boldsymbol{\mu_i}, \Sigma_i)N_p(d\boldsymbol{\mu_i}|\boldsymbol{m}_0, \frac{\Sigma_i}{k_0})IW(d\Sigma_i|\nu_1, \Psi_1)$$

where $N_p(\boldsymbol{\mu}, \Sigma)$ represents the p-variate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, and $IW(\nu_1, \Psi_1)$ represents the inverse-Wishart distribution with mean $\frac{\Psi_1}{\nu_1 - p - 1}$, for $(\nu_1 > p+1)$. We choose these densities to generalize the univariate model presented in (3.1). This integral leads to a multivariate Student's t-density. According to the locate-scale representation, this distribution has location $\boldsymbol{m}_0$, scale matrix $\frac{1+k_0}{k0(\nu_1-1)}\Psi_1$ and $\nu_1 - 1$ degrees of freedom.

A final remark is about the step 2 of the sampling algorithm, when the random mass parameter a is sampled. In fact, before entering the cycle that samples the vector of latent variables, we need to sample a new value of the mass parameter, conditionally on the observations and the latent variables observed in the previous iteration. In other words, we are interested in sampling from $p(a|\boldsymbol{x}, \boldsymbol{\theta})$. Following Escobar and West (1995), we can exclude the dependency

of this distribution from the observations, as from the latent variables. So, the only dependency remaining is the one related to the number of observations n and the number of clusters K_n. We can write:

$$p(da|\boldsymbol{x}, \boldsymbol{\theta}) = p(da|K_n = k) \propto p(K_n = k|a)p(da)$$

where we used the variable K_n to summarize the dependencies stated above and the expression on the right side of the proportionality is obtained thanks to the Bayes theorem. All the terms of this equation are completely known, and go back to formula (3.3), used to compute the prior distribution of the number of clusters (in the case of random mass parameter). We can write the formula above in the following way:

$$p(da|K_n = k) \propto a^k \frac{\Gamma(a)}{\Gamma(a+n)} p(da)$$

recalling that $\frac{\Gamma(a)}{\Gamma(a+n)} = \frac{\beta(a,n)}{\Gamma(n)}$, where $\beta(a, n)$ represents the normalization constant of a $Beta$ distribution with parameters a and n, with mean $\frac{a}{a+n}$ and variance $\frac{an}{(a+n)^2(a+n+1)}$. So, we are able to perform an augmentation technique for sampling, including the new $Beta$ random variable, and we can see the formula above as the distribution resulting by integrating out the random variable $\eta \sim Beta(a, n)$ from the joint distribution with the random mass parameter a, conditionally on the number of clusters K_n.

Using the Bayes theorem, we can write the full conditionals of the two variables a and η, that allow us to perform a separated Gibbs sampling algorithm. Notice that, in order to avoid computational problems, in the following formulas, we will get the normalizing constant $\beta(a + 1, n)$ rather than $\beta(a, n)$, leading to an additional term $\frac{a+n}{a}$ in the equation (see Escobar and West (1995) for further details).

$$\eta|a, K_n = k \sim Beta(a + 1, n)$$

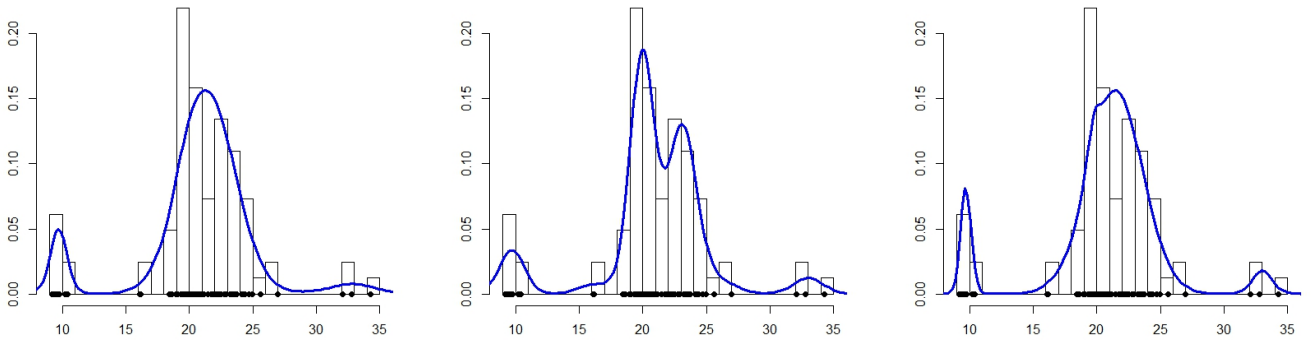
$$a|\eta, K_n = k \sim p_\eta \Gamma(\gamma_1 + k, \gamma_2 - \log \eta) + (1 - p_\eta)\Gamma(\gamma_1 + k - 1, \gamma_2 - \log \eta)$$

where $p_\eta = \frac{\gamma_1 + k - 1}{n(\gamma_2 - \log \eta)}$. The presence of a mixture of two Gamma densities in the full conditional of a is due to the presence of the term $(a + n)$.

6.2.1 Galaxy Data

In this section, we tackle the density estimation problem for the Galaxy dataset, performed using the sampling algorithm just described. All the MCMC samples used are of length 5.000 iterations, provided with a burn-in of 50.000 iterations and a thinning of 15.

In Figure 6.1 the case of Dirichlet process prior with random mass parameter is reported, varying the set of hyperparameters of P_0. In particular, we assume, as in Escobar and West (1995), $a \sim Gamma(2, 4)$, where $Gamma(\alpha, \beta)$ represents the univariate gamma distribution. In Figure 6.2, posterior distributions of the number of clusters are presented (blue), compared with the correspondent prior distributions (green). Due to the fact that we are dealing with a Dirichlet process prior with random mass parameter, it is important to show the estimated posterior distribution of such mass parameter, displayed in Figure 6.3.



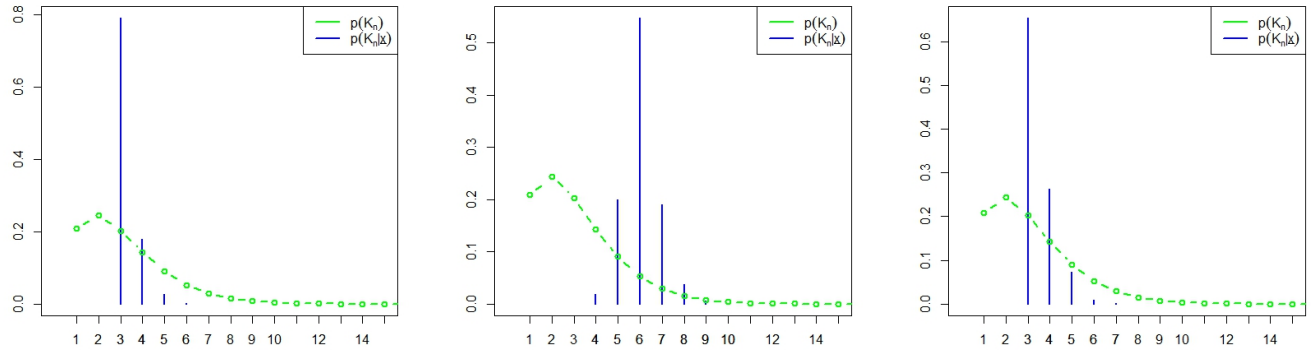
(a) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$ (b) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 20, 20)$ (c) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$

Figure 6.1: Density estimation for Galaxy data. Dirichlet process with random mass parameters. $\mathbb{E}[K_n] = 3$ and $(\gamma_1, \gamma_2) = (2, 4)$.

6.2.2 Bivariate Dataset with non-convex support

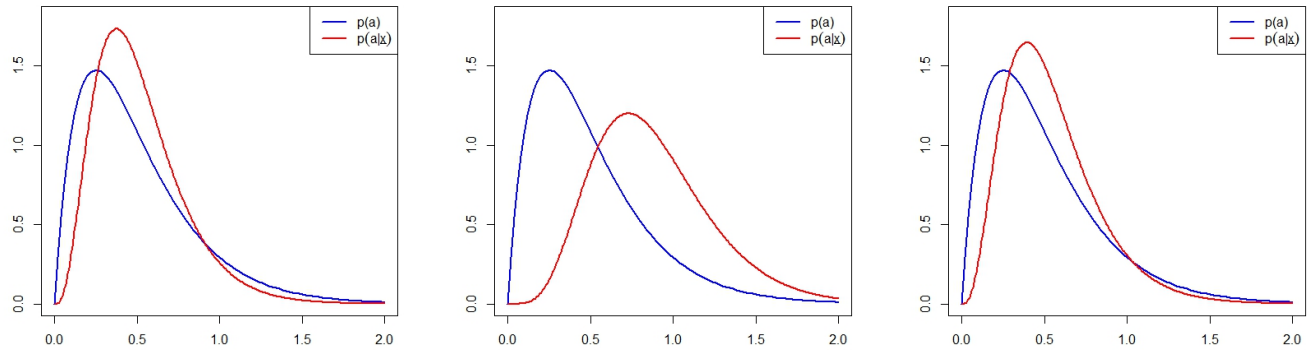
We analyzed a simulated bivariate dataset. We sampled $n = 250$ and $n = 1000$ points from a mixture of bi-variate normal densities in order to obtain points laying on a non-convex region. In particular, we obtained a sharp cloud of points in the middle, and a semi-circular group of elements on the right side (see Figure 6.4). The two sets of hyperparameters we used are described at the beginning of Chapter 5.

To provide the Bayesian estimates, we sampled each MCMC chain for a total number of itera-



(a) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$ (b) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 20, 20)$ (c) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$

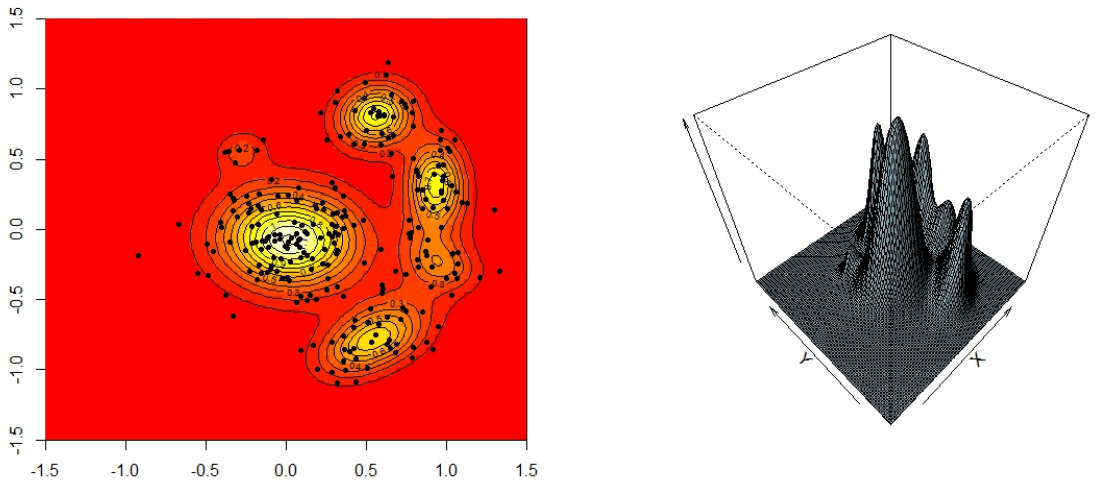
Figure 6.2: Prior (green) and estimated posterior (blue) number of clusters. Dirichlet process with random mass parameters. $\mathbb{E}[K_n] = 3$ and $(\gamma_1, \gamma_2) = (2, 4)$.



(a) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$ (b) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 20, 20)$ (c) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$

Figure 6.3: Prior (blue) and estimated posterior (red) distributions for the mass parameter a. Dirichlet process with random mass parameters. $\mathbb{E}[K_n] = 3$ and $(\gamma_1, \gamma_2) = (2, 4)$.

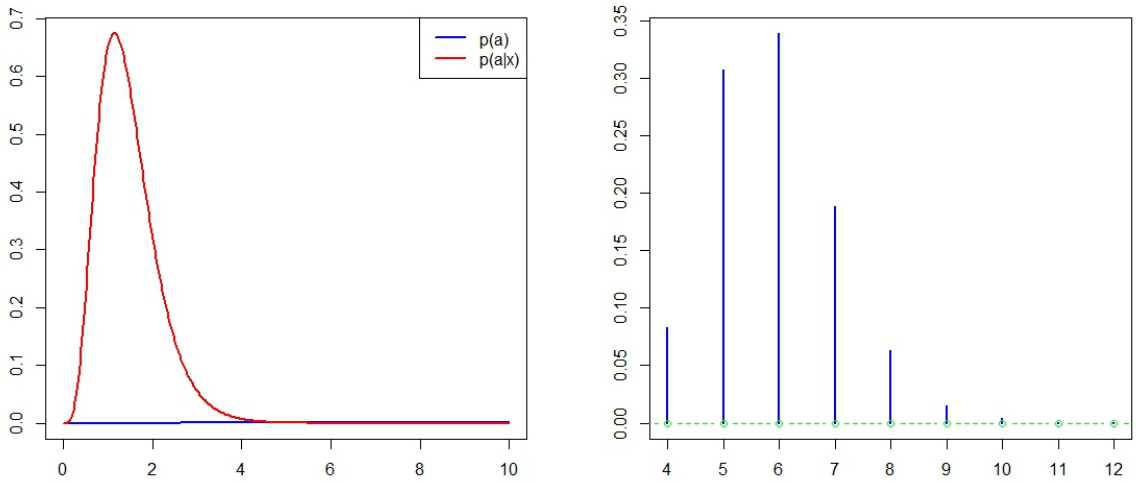
tions equal to 5.000, with a burn-in of 50.000 and a thinning of 15. In Figure 6.4 we show the density estimation for this dataset, given by the multivariate Polya urn sampling algorithm, together with the estimation of the posterior density of the random mass parameter a and of the number of clusters K_n in Figure 6.5. As expected, the density estimation provided by the Polya urn sampling scheme tends to locate a dominant mode on the central group of observations, while several more little modes disposed on the semi-circular region. This is due to the fact that we choose hyperparameters for the Dirichlet process prior such that a large prior expected number of clusters is obtained, leading to an estimation of the generating density composed of multiple distinct contributes.



(a) Contour plot.

(b) Perspective plot

Figure 6.4: Density estimation for the simulated dataset with $n = 250$ observations, for the first set of hyperprameters.



(a) Prior (blue - $Gamma(2, 0.01)$) and estimated posterior (blue) distributions of the total mass parameter a.

(b) Prior (green) and estimated posterior (blue) distributions of the number of clusters K_n.

Figure 6.5: Dataset with $n = 250$ observations, first set of hyperparameters.

6.3 NGG Process

In the case of NGG process prior, the predictive density for the latent variable, following the Polya urn described in (6.1), is of the form:

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_{i-1} \sim \omega_0^{(n)} P_0 + \omega_1^{(n)} \sum_{j=1}^{K_n} (n_j - \sigma)\delta_{\boldsymbol{\phi}_j} \text{ for } i = 1,...,n,$$

where K_n is the observed number of clusters, before the i-th extraction, and:

$$\omega_0^{(n)} = \frac{\sigma}{n} \frac{\sum_{i=0}^{n} \binom{n}{i} (-1)^i \beta^{\frac{i}{\sigma}} \Gamma(K_n + 1 - i/\sigma; \beta)}{\sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{\frac{i}{\sigma}} \Gamma(K_n - i/\sigma; \beta)}$$
$$\omega_1^{(n)} = \frac{1}{n} \frac{\sum_{i=0}^{n} \binom{n-1}{i} (-1)^i \beta^{\frac{i}{\sigma}} \Gamma(K_n - i/\sigma; \beta)}{\sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{\frac{i}{\sigma}} \Gamma(K_n - i/\sigma; \beta)}$$

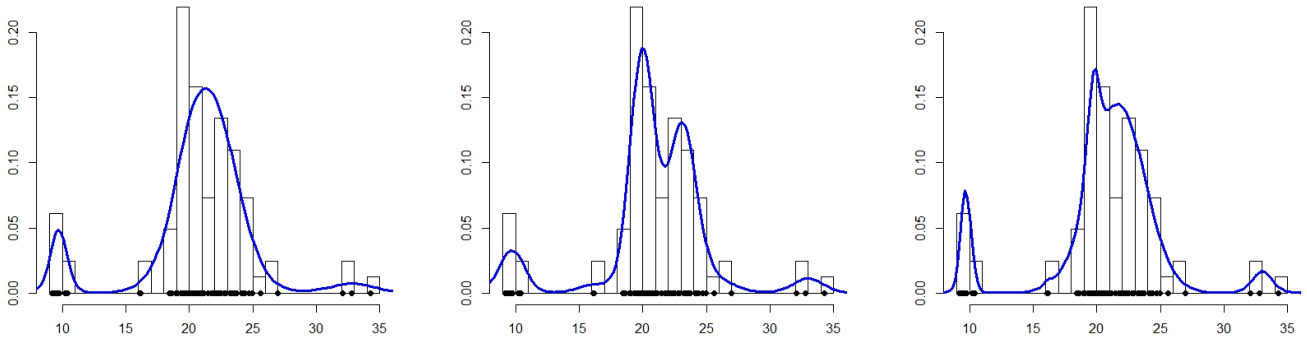
where $\Gamma(a;x) = \int_x^{+\infty} s^{a-1} exp(-s) ds$ is the incomplete gamma function. Such weights are very complex to compute, for further details see Argiento et al. (2010). To present the connection between the DPM model involving the NGG process prior and the corresponding PPM, we write the predictive density in the following way (see Lijoi et al., 2007):

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_{i-1} \sim \frac{V_{n+1,K_n+1}}{V_{n,K_n}} P_0 + \frac{V_{n+1,K_n}}{V_{n,K_n}} \sum_{j=1}^{K_n} (n_j - \sigma)\delta_{\boldsymbol{\phi}_j} \text{ for } i = 1,...,n,$$

Provided this form of the predictive density of the latent variables, together with the formula (3.7), we can set up the Gibbs sampling algorithm, similarly to what has been done in the case of the Dirichlet process prior.

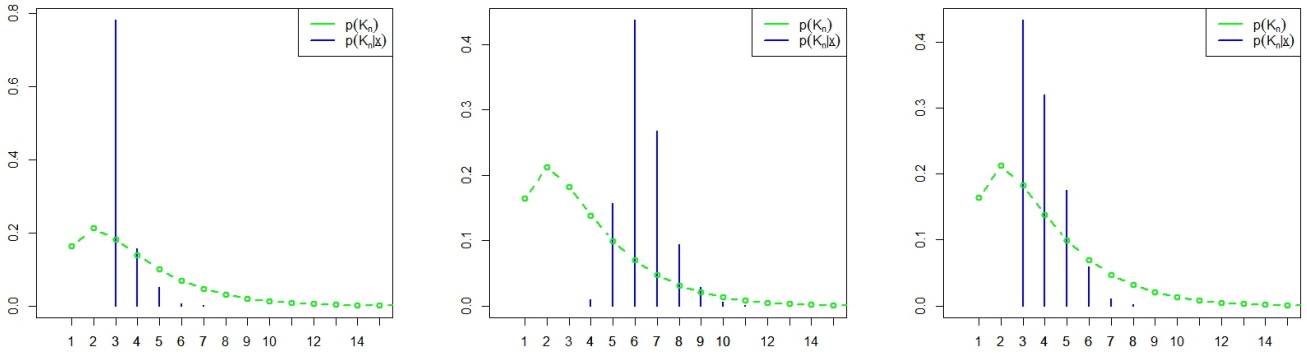
6.3.1 Galaxy Data

We present in Figures 6.6 and 6.7 the density estimations and posterior number of clusters obtained with the sampling from an NGG process prior, applied to the Galaxy dataset. The case is the one with $\mathbb{E}[K_n] = 3$. The samples are of 5.000 iterations, with a burn-in of 50.000 iterations and a thinning of 15.



(a) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$ (b) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 20, 20)$ (c) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$

Figure 6.6: Density estimation for Galaxy dataset, using the NGG process prior. $\mathbb{E}[K_n] = 3$ and $(\sigma, \kappa) = (0.25, 0.05)$.



(a) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 2, 1)$ (b) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.01, 20, 20)$ (c) $(m_0, k_0, \nu_1, \Psi_1) = (0, 0.001, 3, 0.2)$

Figure 6.7: Prior (green) and estimated posterior (blue) number of clusters. NGG process prior. $\mathbb{E}[K_n] = 3$ and $(\sigma, \kappa) = (0.25, 0.05)$.

Bibliography

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

[2] Argiento, R., Guglielmi, A. and Pievatolo, A. (2010) *A Mixed-effects modelling of Kevlar fibre failure times through Bayesian nonparametrics.* Complex data modeling and computationally intensive statistical methods, eds. Mantovan, P. and Secchi, P., Springer, New York, 2010a.

[3] Argiento, R., Guglielmi, A. and Pievatolo, A. (2010) *Bayesian density esitmation and model selection using nonparametric hierarchical mixtures.* Computational Statistics and Data Analysis, Vol. 139, pp 3989-4005.

[4] Argiento, R., Guglielmi, A. and Soriano, J. (2012) *A semiparametric Bayesian generalized linear mixed model for the reliability of Kevlar fibres.* Technical report.

[5] Bell, E. T. (1934). *Exponential Numbers.* : The American Mathematical Monthly, Vol. 41, No. 7, pp. 411-419.

[6] Binder, D.A. (1978) *Bayesian Cluster Analysis*, Biometrika, Vol.65, No.1, pp. 31-38.

[7] Blackwell, D. and MacQueen, J. B. (1973). *Ferguson Distributions Via Polya Urn Schemes.* The Annals of Statistics, Vol. 1, No. 2, pp 353-355.

[8] Brix, A. (1999) *Generalized gamma measures and shot-noise Cox processes.* Adv. Appl. Probab., 31, pp. 929-953.

[9] Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991) *Statistical analysis of reliability data.* Chapman and Hall, London, 1991.

[10] Csiszar, I. (1975) *I-Divergence Geometry of Probability Distributions and Minimization Problems*. The Annals of Probability, Vol. 3, No. 1, pp. 146-158.

[11] Dahl, D. B. (2006). *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*. Gene Expression and Proteomics, Pages 201-218. Cambridge University Press.

[12] Dahl, D. B. (2009). *Modal Clustering in a Class of Product Partition Models*. International Society for Bayesian Analysis.

[13] Ester M., Kriegel H.-P., Xu X. (1996). *Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification*, Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 1995, in: Lecture Notes in Computer Science, Vol. 951, Springer, pp. 67-82.

[14] Escobar, M. D. and West, M. (1995) *Bayesian density estimation and inference using mixtures*, Journal of the American Statistical Association, Vol. 90, N0. 430, pp. 577-588.

[15] Ferguson, T.S. (1973) *A bayesian analysis of some nonparametric problems*, The Annals of Statistics.

[16] Fraley, C. and Raftery, A. E. (1998). *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis*, THE COMPUTER JOURNAL.

[17] Fraley, C. and Raftery, A. E. (2002). *Model-Based Clustering, Discriminant Analysis, and Density Estimation*, Journal of the American Statistical Association.

[18] Fritsch, A. and Ickstadt, K. (2009). *Improved Criteria for Clustering Based on the Posterior Similarity Matrix*. International Society for Bayesian Analysis, Vol. 4, Number 2, Pages 367-392.

[19] Gnedin, A. and Pitman, J. (2005). *Exchangeable Gibbs partitions and Stirling triangles*. Zap. Nauchn. Sem. St Petersburg. Otdel. Mat. Inst. Steklov., Vol. 325, pp. 83-102.

[20] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. Springer, New York.

[21] Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006). *A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves.* Journal of the American Statistical Association, Vol. 101, No. 473.

[22] Jain, A. K., Dubes, R. C. (1988). *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, New Jersey.

[23] James, L.F., Lijoi, A. and Prünster, I. (2006): *Conjugacy as a distinctive feature of the Dirichlet process.* Scandinavian Journal of Statistics, Vol. 33, pp. 105-120.

[24] Johnson, S. C. (1967): *Hierarchical Clustering Schemes.* Psychometrika, Vol. 2, N0. 3, pp. 241-254.

[25] Kingman, J. F. C. (1975): *Random Discrete Distributions.* Journal of the Royal Statistical Society B 37, 1-22.

[26] Lau, J. W. and Green, P. J. (2007). *Bayesian model based clustering procedures.* Journal of Computational and Graphical Statistics.

[27] Lijoi, A., Mena, R.H. and Prünster, I. (2007). *Controlling the reinforcement in Bayesian non-parametric mixture models.* Journal of the Royal Statistical Society, Vol. 69, Part 4, pp. 715-740.

[28] Lo, A.J. (1984). *On a Class of Bayesian Nonparametric Estimates: I. Density Estimates.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics.

[29] Mac Queen, J. (1967). *Some methods for classification and analysis of multivariate observations.* The Annals of Statistics, Vol. 12, No. 1, pp 351-357.

[30] Neal, R. M. (2000). *Markov Chain Sampling Methods fro Dirichlet Process Mixture Models.* Journal of Computational and Graphical Statistics, Vol. 9, No. 2, pp 249-256.

[31] Pitman, J. (2003) *Poisson-Kingman partitions*Lecture Notes-Monograph Series, Vol. 40, Statistics and Science: A Festschrift for Terry Speed, pp. 1-34. Institue of Mathematical Statistics Hayward, California.

[32] Pitman, J. (2006) *Combinatorial Stochastic Processes*. Springer, New York.

[33] Quintana, F. A. and Iglesias, P. L. (2003). *Bayesian Clustering and Product Partition Models*. Journal of the Royal Statistical Society, Series B, Methodological.

[34] Regazzini, E. Lijoi, A. and Prünster, I. (2003). *Distributional results for means of random measures with independent increments*. Annals of Statistics, Vol. 31, No. 2, pp. 560-585.

[35] Roeder, K. (1990). *Density estimation with confidence sets exemplified by superclusters and voids in the galaxies*. Journal of the American Statistical Association, Vol. 85, No. 411, pp. 617-624.