UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXII

Climbing modes and exploring mixtures: a journey in density-based clustering

Coordinatore del Corso: Prof. Massimiliano Caporin

Supervisore: Prof.ssa Giovanna Menardi

Co-supervisore: Prof. José E. Chacón

Dottorando: Alessandro Casa

30th September 2019

Abstract

The density-based formulation aims at recasting the clustering problem to a mathematically sound framework, by linking the groups to some features of the density assumed to underlie the data. Even if early probabilistic approaches to cluster analysis can be traced back to fifty years ago, the topic has recently found a renewed and vibrant interest in the scientific community. This may be motivated both by the computational advancements witnessed in the last years and by more conceptual and substantial reasons. The increased availability of mixed type and complex structured data has indeed required a rigorous formalization of the clustering problem.

Stemming from the same roots, density-based clustering has been developed following two distinct paths. In its parametric formulation, a connection among groups and unimodal components of a mixture model is drawn. On the other hand, according to the nonparametric paradigm, clusters are seen as the domains of attraction of the modes of the density. Revolving around this approach to clustering, the thesis explores both the formulations by highlighting at the same time contact points and dissimilarities. Moreover differently structured data are considered ranging from the unidimensional setting to more complex *three-way* structure.

Three main contributions can be highlighted. In the first one we derive some asymptotic results to address nonparametric density estimation from a clustering-oriented perspective. In the second contribution we propose an ensemble approach for density-based clustering, which inherits the strengths from both the parametric and the nonparametric formulation, and possibly enhances the robustness and the stability of the partitions. The third contribution addresses the problem of clustering complex multivariate time-dependent data by adopting a parametric approach and proposing a flexible modification of the Latent Block Model.

Sommario

Nell'ambito dell'analisi di raggruppamento l'approccio basato su densità mira ad ottenere una formalizzazione matematica del problema di *clustering* associando il concetto di gruppo ad alcune caratteristiche della funzione di densità sottostante i dati. Sebbene i primi approcci probabilistici al *clustering* risalgano a cinquant'anni fa, recentemente si è potuto notare un rinnovato interesse verso questo argomento. Alcune possibili motivazioni possono essere rintracciate negli sviluppi computazionali a cui si è assistito negli ultimi anni o nella crescente complessità dei dati a disposizione che ha richiesto una formalizzazione più rigorosa del problema di raggruppamento.

Il *clustering* basato su densità, pur condividendo concettualmente lo stesso punto di partenza, è stato sviluppato secondo due paradigmi differenti. Nella sua formulazione parametrica vi è una relazione biunivoca tra i gruppi e le componenti unimodali di un modello di mistura. D'altra parte l'approccio non parametrico associa i gruppi ai domini di attrazione delle mode della funzione di densità. Questa tesi mira ad esplorare entrambe le formulazioni, evidenziandone punti di contatto e differenze. Allo stesso tempo vengono analizzati dati aventi strutture radicalmente differenti, che spaziano da scenari unidimensionali fino a più complessi dati *three-way*.

In questo lavoro possono essere evidenziati tre contributi principali. Innanzitutto sono stati sviluppati alcuni risultati asintotici per la stima di densità non parametrica quando considerata come accessoria al *clustering*. In secondo luogo si è proposto un approccio *ensemble* in contesto di *clustering* basato su densità il quale, ereditando i punti di forza da entrambe le formulazioni, mira a migliorare la qualità delle partizioni ottenute in termini di robustezza e stabilità. Infine, ci si è posti l'obiettivo di modellare in maniera flessibile dati multivariati tempo-dipendenti mediante l'identificazione di gruppi di curve aventi un comportamento omogeneo, tramite un'opportuna modifica del modello a blocchi latenti.

"Not all those who wander are lost"

- J.R.R. Tolkien -

Acknowledgements

I would like to thank the people I worked with during these years.
First of all I would like to express my gratitude to Prof. Giovanna Menardi who has shared with me a long trip in the clustering world starting from my Bachelor Thesis. Her constant support, patience, insights, suggestions and passion have had an invaluable impact on my growth as a researcher and on the realization of this thesis.
I am also extremely grateful to Prof. José Chacón, Prof. Luca Scrucca and Prof. Charles Bouveyron for the chances they have given me to work on such interesting topics, for their help and suggestions along the road and for the fruitful conversations.

Contents

List of Figures

List of Tables

Introduction

Overview

All the attempts to partition a set of data into some homogeneous groups may be gathered under the unified heading of *cluster analysis*. Cluster analysis has been pervasively pursued in many different fields both as a preliminary step and as the main focus of the data analysis, for classification or data compression. Examples of popular frameworks where clustering has found fruitful applications are market segmentation, classification of species in biology or, more recently, recommendation systems in information technology, the automation of diagnostic processes in the analysis of medical images, and anomaly detection, among many others.

Tons of different tools have been proposed over the years, most of them based on some measure of distance or dissimilarity. The soundness of these techniques is often questionable, as they pursue some vague and heuristic notion of cluster. The search for homogeneous and unknown patterns in data, while intuitively clear, lacks, in fact, a specification of the target of the analysis or, in other words, of what we are precisely searching for.

An effort to overcome the ill-posedness of the clustering problem has been made via the so called *density-based* approach. Here, a probability density function is assumed to underlie the observed data, properly describing the corresponding generative mechanism. A precise notion of cluster is provided by drawing a correspondence between the groups and the features of the density itself, thus allowing to frame the clustering problem into a standard inferential context. Therefore, proper tools are available in the estimation process and to evaluate the obtained partitions with respect to a targeted and well-defined ground truth.

The density-based framework has been developed taking two distinct but related paths. The parametric, or *model-based*, formulation assumes, as an underlying probability distribution, a mixture model, and identifies a partition by exploiting the correspondence between groups and components of the mixture itself. On the other hand

in the nonparametric, or *modal*, counterpart the density is estimated nonparametrically and its modes are seen as the archetypal points of the clusters, in turn corresponding to their domains of attraction. Albeit related, the two paradigms enjoy different strengths and reasons of attractiveness, which make the one or the other overall preferable, depending on data features and subject-matter considerations.

Revolving around the density-based framework, this thesis aims at proposing some alternative solutions to clustering-related problems, faced via distinct approaches and involving different data specificities. From a conceptual point of view it may be seen as a journey in the density-based clustering realm, where different keys to reading can be highlighted.

According to a first perspective, the thesis presents three contributions which explore different formulations of density-based clustering, and focus on different estimation approaches and notions of cluster. A first contribution is based on a fully nonparametric approach, letting the data driving us in the modality exploration, without specific assumptions on the clusters shape. A further contribution focuses on the model-based paradigm where the resort to some parametric assumptions allows handling highly complex data structures and unveiling parsimonious patterns. The last contribution of the thesis may be placed somewhere in between the previously mentioned ones. An ensemble clustering approach is proposed, working as a bridge between the two different density-based formulations. Both modal and model-based ideas are employed to blend together the different perspectives and to show how this blending may be beneficial in terms of cluster characterization.

As for the second perspective, a different interpretative point of view may be offered, mainly focusing on the involved data structures. Despite the strong contact points among parametric and nonparametric formulations, their differences reflect on their capabilities to handle specific settings; dimensionality and data complexity turn out to be among the factors having more impact on the clustering performances and consequently on the choice of a specific paradigm. Nonparametric techniques, strongly relying on the concept of neighborhood, focus on local structures in order to get a sense out of the data. As a consequence, when dealing with high dimensional spaces, where much of the probability mass flows to the tails of the underlying density, the resulting sparsity takes it toll by deteriorating the quality of their estimates. Therefore, modal clustering is usually more effective with low dimensional data. Consistently, the approach has been explored in the thesis mostly in the unidimensional (*one-way*) setting, where a mathematical formalization is also easier to be derived. Conversely, parametric methods are less sensitive to the data dimensionality, hence the model-based formulation

has been employed to both deal with a wider range of dimensionality scenarios in the common subject-variables framework (*two-way* data), and to handle the complexity of multivariate time-dependent (*three-way*) data.

Main contributions of the thesis

The previous section revealed the silver threads that flow underneath this thesis, simultaneously motivating it and linking together its different parts. Nevertheless a pretty neat distinction among the chapters is not only possible but it may also come to an aid when highlighting the main contributions of the work. After a brief introduction to the density-based clustering world in Chapter 1, the contributions of this thesis can be summarized as follows:

- Chapter 2 delves into the nonparametric density estimation process when considered as an instrumental step for the further identification of a clustering structure. When resorting to nonparametric estimators, a fine tuning of the amount of smoothing, which governs the density shape and hence its modal structure, is required. While thoroughly analyzed in the context of density estimation, this issue has been scarcely studied for clustering purposes. In this work we address the problem mainly in the unidimensional setting and from an asymptotic perspective. Stemming from Chacón (2015) we introduce an appropriate metric measuring the discrepancy among the partitions induced by the true and the estimated density functions. Afterwards we derive an asymptotic approximation of the considered metric that allows introducing new automatic bandwidth selectors specifically tailored for modal cluster analysis.

- In Chapter 3 we propose an ensemble clustering approach aiming to overcome the strong reliance on the so called *single best model paradigm*. In the model-based formulation, usually a set of different mixture models is estimated based on a different number of clusters and/or different parametrizations and only the best one is selected according to an information criterion. Arguing that such approach could be sub-optimal from clustering and inferential points of view (see e.g. McNicholas, 2016), the idea of mixing together different models has already been explored by Wei and McNicholas (2015) and Russell *et al.* (2015) where Bayesian model averaging approaches are proposed. In the thesis we take a different path by introducing an estimator defined as a convex linear combination of the density estimates obtained from the models in the ensemble. The estimation process is

practically carried out via likelihood penalization, and some alternative penalties are proposed. The resulting estimate is operationally exploited to obtain partitions by resorting to the modal concept of cluster; therefore we propose an hybrid approach blending together parametric and nonparametric approaches to clustering thus enjoying their pertaining advantages.

- Chapter 4 focuses on the problem of clustering multivariate time-dependent data. A model-based co-clustering strategy (Govaert and Nadif, 2013) is examined, aimed at simultaneously partitioning individuals and variables, as it appears particularly suitable when parsimonious summaries of complex structured data are needed. Specifically we extend a widely used co-clustering model to incorporate random effects in the specification of the data generative mechanism where the *Shape Invariant Model* (Lindstrom, 1995) is considered. As a consequence an high flexibility is introduced which allows encompassing arbitrarily shaped concepts of cluster. As an estimation tool we introduce a *Marginalized SEM-Gibbs algorithm* being a modification of the widely used *SEM-Gibbs* with an added step specifically designed to handle the random effects.

Chapter 1

An overview on density-based clustering

1.1 Introduction

The goal of partitioning a set of data into some groups, diffusely known as clustering, has been extensively studied in the last decades, and proved its usefulness in a wide range of fields of application both as an exploratory step and as the focus of the data analysis. Over the years a plethora of different techniques, based on different rationales, have been proposed, most of them relying on the notion of cluster as a group of loosely similar objects; standard methods include, for example, hierarchical and k-means clustering (see e.g Hennig *et al.*, 2016, for an overview). Despite their appealing interpretability and conceptual simplicity, the soundness of these techniques is questionable, since they build on a vague and heuristic definition of cluster mainly based on concepts as distance and dissimilarity. This lack of a "ground truth" usually prevents the possibility to resort to formal inferential techniques in order to evaluate and compare alternative partitions or to select the number of groups in the data. As a further evidence of the ill-posedness of the clustering problem, several authors along the years have advocated for a mathematical formalization. As examples Aitkin *et al.* (1981) pointed out that "when clustering samples from a population, no cluster analysis is a priori believable without a statistical model" while Meilă (2007) noted how clustering remains a domain "where rigorous methodology is still striving to emerge" and where theoretical developments could be proven to be effective in addressing some of the most commonly arising criticalities.

An attempt to formalize the clustering problem by framing it into a statistically rigorous context has been pursued via the so called *density-based approach*. Here the concept of cluster is linked to some specific feature of the probability density function

assumed to underlie the observed data.

The focus on the underlying density and the correspondence drawn between its characteristics and the groups introduces some appealing advantages in the clustering process. First of all, an *ideal population goal*, defined as the partition induced by the true density, is introduced and can be exploited to evaluate and compare different data-based clusterings. Furthermore the density-based formulation enables to obtain partitions of the whole sample space and not only of the observed data, allowing the chance to classify also incoming, not already seen, observations. Lastly, in this framework the choice of the number of groups can be easily recasted to a model selection problem or, generally speaking, included in the modelling process.

The idea on which the density-based formulation is built has been explored following two distinct paths. It is possible to highlight a parametric, or *model-based*, approach and a nonparametric, or *modal*, one with the first one being unarguably more widespread and better established. Despite sharing the same rationale, the two approaches present some relevant differences, not only from a practical point of view. Operationally different density estimators are considered and, from a more conceptual point of view, different notions of cluster and different ideal population goals are aimed at.

The next two sections will be devoted to a more comprehensive introduction to the two paradigms outlined above focusing on their basic formulations, on the estimation procedures and on the practical identification of the groups. Furthermore a glimpse of their strengths and weaknesses will be given as well as their contact points.

1.2 Parametric formulation

1.2.1 Cluster notion and model specification

The model-based approach to clustering (Banfield and Raftery, 1993; Fraley and Raftery, 2002) represents undoubtedly the most studied and known formulation in the density-based family. A first, tentative, definition of the concept of group in this framework could be dated back to Wolfe (1963) who defined it as "a distribution which is one of the components of a mixture of distributions". A more recent and comprehensive definition is given in McNicholas (2016) where a cluster is seen as "a unimodal component within an appropriate finite mixture model". Roughly speaking, the parametric formulation to cluster analysis draws a one-to-one correspondence among the groups and the components of a parametric mixture model. More specifically, denoted by $f : \mathbb{R}^d \to \mathbb{R}$ the probability density function assumed to underlie the observed data

$\mathbb{X} = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$, f is specified as follows

$$f(x|\Theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k) \ , \tag{1.1}$$

where K is the number of mixture components, $f_k(\cdot)$ the *kth* component density, while $\Theta = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ is the full parameter vector, with π_ks representing the mixing proportions, $\pi_k > 0$, $\forall k = 1, \ldots, K$ and $\sum_k \pi_k = 1$.

Given the formulation (1.1) it is straightforward to define the *ideal population goal* in this framework. Indeed the population clustering $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, induced by the true density $f(\cdot|\Theta)$, has ideal clusters defined as

$$\mathcal{C}_k = \{x \in \mathbb{R}^d : \pi_k f_k(x|\theta_k) \geq \pi_j f_j(x|\theta_j), \ \forall j \neq k\} \tag{1.2}$$

with $k = 1, \ldots, K$.

In applications usually Gaussian densities are considered as the component ones; therefore $f_k(\cdot|\theta_k) = \phi_k(\cdot|\theta_k)$ with $\theta_k = (\mu_k, \Sigma_k)$. Nonetheless, since Gaussian mixture models are restricted to the detection of elliptically shaped clusters, other parametric distributions have been studied and exploited in a clustering framework: mixtures of multivariate t-distribution has been used in McLachlan and Peel (1998), while Lin (2009, 2010) proposed mixtures of skew-normal and skew-t, only to mention a few works in this direction. Further references can be found in Bouveyron *et al.* (2019, Ch.9)

Note that when the dimensionality d of the data increases, mixture models turn out to be highly overparametrized. As an example, in the Gaussian setting, the covariance matrices Σ_k has an exploding number of parameters. In order to alleviate this issue, several solutions have been proposed. The most common one, proposed by Banfield and Raftery (1993), consists in considering an eigendecomposition of the component covariance matrices as $\Sigma_k = \lambda_k D_k A_k D_k'$ where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix with elements being proportional to the eigenvalues and λ_k is a constant of proportionality. This decomposition entails an appealing interpretation from a geometric point of view; D_k determines the orientation of the *k-th* component of the mixture while A_k governs its shape and λ_k its volume. By imposing different constraints on the elements involved in the decomposition of the covariance matrices, Celeux and Govaert (1995) obtain a family of Gaussian parsimonious clustering models (see Table 1.1)

Type	Model	Volume	Shape	Orientation	Σ_k
Spherical	EII	Equal	Spherical	-	λI
	VII	Variable	Spherical	-	$\lambda_k I$
Diagonal	EEI	Equal	Equal	Axis-aligned	λA
	VEI	Variable	Equal	Axis-aligned	$\lambda_k A$
	EVI	Equal	Variable	Axis-aligned	λA_k
	VVI	Variable	Variable	Axis-aligned	$\lambda_k A_k$
General	EEE	Equal	Equal	Equal	$\lambda DAD'$
	VEE	Variable	Equal	Equal	$\lambda_k DAD'$
	EVE	Equal	Variable	Equal	$\lambda DA_k D'$
	EEV	Equal	Equal	Variable	$\lambda D_k A D'_k$
	VVE	Variable	Variable	Equal	$\lambda_k DA_k D'$
	VEV	Variable	Equal	Variable	$\lambda_k D_k A D'_k$
	EVV	Equal	Variable	Variable	$\lambda D_k A_k D'_k$
	VVV	Variable	Variable	Variable	$\lambda_k D_k A_k D'_k$

TABLE 1.1: Nomenclature and covariance structure of the 14 models in the Gaussian parsimonious clustering models family.

1.2.2 Estimation and allocation procedures

In order to practically obtain a partition of the observed data according to the model-based formulation, an estimate $\hat{\Theta}$ of the full parameter vector is required. In this framework the most commonly adopted choice consists in maximizing the likelihood of model (1.1) defined as

$$\mathcal{L}(\Theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k) \tag{1.3}$$

with corresponding log-likelihood being

$$\ell(\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k). \tag{1.4}$$

Maximization of (1.3) is carried out by means of the *Expectation-Maximization algorithm* (EM, Dempster *et al.*, 1977). This algorithm offers a general approach to maximum likelihood estimation in a variety of incomplete-data situations.

In the mixture model framework this propriety comes in handy since it is easy to recast the problem to a missing data one. Thus the observed data x_i are referred to as *incomplete*. In turn the *complete data* are defined as the couples $(x_i, z_i)_{1<i<n}$, where

$\mathbf{z} = \{z_i\}_{1 \leq i \leq n}$ with $z_i = (z_{i1}, \ldots, z_{iK})$ represents the unobserved component membership. More specifically

$$z_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to group } k \\ 0 & \text{otherwise.} \end{cases}$$

The EM algorithm aims at maximizing the *complete data likelihood* $\mathcal{L}_c(\Theta)$ by means of an iterative procedure which alternates between two steps: an Expectation step (*E-step*), where the conditional expectation of the *complete data log-likelihood* is computed given the current estimates of the parameters and the observed data, and a Maximization step (*M-step*) in which the parameter estimates are updated by maximizing the expected log-likelihood obtained in the previous step. Under some regularity conditions (see McLachlan and Krishnan, 2007), the EM algorithm converges to a local maximum of $\mathcal{L}_c(\Theta)$. Nevertheless, even when these conditions are not met, the algorithm has shown good performances in practical applications.

In the considered framework, assuming that z_i is drawn from a multinomial distribution with probabilities $\pi_1, \ldots, \pi_K$, the *complete data likelihood* and the associated log-likelihood are defined as

$$\mathcal{L}_c(\Theta, \mathbf{z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f_k(x_i|\theta_k)]^{z_{ik}} \tag{1.5}$$

$$\ell_c(\Theta, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[\pi_k f_k(x_i|\theta_k)]. \tag{1.6}$$

In the E-step determining the expected complete data log-likelihood corresponds to replace the z_{ik}s in (1.6) with their expectation

$$\hat{z}_{ik} = \frac{\pi_k^{(q)} f_k(x_i|\theta_k^{(q)})}{\sum_{k'}^{K} \pi_{k'}^{(q)} f_{k'}(x_i|\theta_{k'}^{(q)})}, \tag{1.7}$$

where $\pi_k^{(q)}$ and $\theta_k^{(q)}$ are the parameter estimates at *q-th* iteration of the algorithm. In the M-step the availability of closed form solutions depends on the parametric family chosen to model the component densities.

The two steps are iterated until a convergence criterion, usually on the values assumed by $\ell_c(\Theta)$ on subsequent iterations, is met.

Despite being successfully exploited in a variety of different real data applications, the EM algorithm has shown a certain number of limitations when used in the mixture modelling framework. Firstly, the rate of convergence can be slow and the initialization

turns out to be particularly relevant both for the convergence speed and to avoid spurious solutions corresponding to local maxima of $\mathcal{L}_c(\Theta)$. Secondly, when considering Gaussian mixtures, singularity or nearly singularity of covariance matrices leads the algorithm to a breakdown. In order to overcome some of the limitations of the standard EM algorithm Ingrassia and Rocci (2007, 2011) show how working on a constrained space, where constraints are imposed on the eigenvalues of the component covariance matrices, solves the singularity issues and decreases the number of local maxima. Furthermore some variants of the algorithm have been proposed, as for example the *classification* EM (CEM, Celeux and Govaert, 1992) or the *stochastic* EM (SEM, Celeux and Diebolt, 1985), being less sensitive to the problems mentioned above.

Once the parameter estimates $\hat{\Theta} = (\hat{\pi}_1, \ldots, \hat{\pi}_K, \hat{\theta}_1, \ldots, \hat{\theta}_K)$ are obtained, the allocation of the observations to the clusters is derived straightforwardly via maximum a posteriori (MAP) classification. More specifically the *ith* observation is assigned to cluster k^* if $k^* = \arg\max_k \hat{z}_{ik}$ where $\hat{z}_{ik}$ is defined as in (1.7) with $\pi_k^{(q)}$ and $\theta_k^{(q)}$ respectively replaced by $\hat{\pi}_k$ and $\hat{\theta}_k$.

From a practical point of view, the usual working routine in model-based clustering follows the so called *single best model paradigm*. It consists firstly in estimating a set of models corresponding to different number of mixture components, different specifications for the component densities or, in the Gaussian case, distinct parametrizations of the component covariance matrices. Afterwards, the best model among the estimated ones is selected and used to obtain a partition and for the subsequent analysis steps. Model selection is usually carried out according to an information criterion with the Bayesian Information Criterion (BIC, Schwarz, 1978) being the most popular one. It is defined as

$$\text{BIC} = 2\ell(\hat{\Theta}) - \gamma \log n \ , \tag{1.8}$$

where $\ell(\cdot)$ is defined in (1.4) and γ is the number of free parameters in the model and can be seen as a proxy of the complexity of the model. Therefore the second term in (1.8) represents a sample-size dependent penalty leading to the selection of more parsimonious models. Despite having shown remarkable results in a plethora of different applications, the BIC is neither specifically conceived for the clustering setting nor it assures the selection of the model with the best classification performance. For this reason some other options have been studied as, for example, the Integrated Complete Likelihood

(ICL, Biernacki *et al.*, 2000) usually approximated by

$$\text{ICL} \simeq \text{BIC} + 2\sum_{i=1}^{n}\sum_{k=1}^{K}\text{MAP}(\hat{z}_{ik})\log\hat{z}_{ik} \tag{1.9}$$

where $\text{MAP}(\hat{z}_{ik}) = 1$ if observation i belongs to cluster k. The second term in (1.9) aims at reflecting the uncertainty in the final partition therefore the ICL tends to select models where separation among clusters is more clear.

1.3 Nonparametric formulation

1.3.1 Cluster notion and model specification

The modal formulation of density-based clustering is unarguably less widespread than its parametric counterpart. Research in this field have been conducted in a considerably more scattered way and it is hence harder to give a comprehensive view of the developments and of their recent directions. A notable attempt to systematically review the state of the art in nonparametric clustering has been made in Menardi (2016) which the reader can refer to for more details.

A first rough definition of the concept of cluster, coherent with the modal formulation, can be traced back to Carmichael *et al.* (1968) where groups are seen as "continuous, relatively densely populated regions of the space, surrounded by continuous, relatively empty regions". Further developments of the definition are given by Wishart (1969) asserting that clusters should be "distinct data modes, independently of their shapes and variance" and by Hartigan (1975) who stated that "clusters may be thought of as regions of high density separated from other such regions by regions of low density". This early attempts to connect clustering to the modal structure of the density are somewhat heuristic. A more structured definition of the concept of group was then given by Stuetzle (2003) who drew a correspondence between clusters and the "domains of attraction" of the density modes. Even if still rather vague this definition has relieved the ill-posedness of the clustering problem by linking the groups to some features of the underlying density. Roughly speaking it is then possible to highlight a one-to-one correspondence among clusters and modal regions of the density, with the modes being the archetypes of the clusters themselves.

It is only recently, however, that the concept of cluster could find a rigorous formalization. To this aim, Chacón (2015) has resorted to the the aid of Morse Theory, a branch

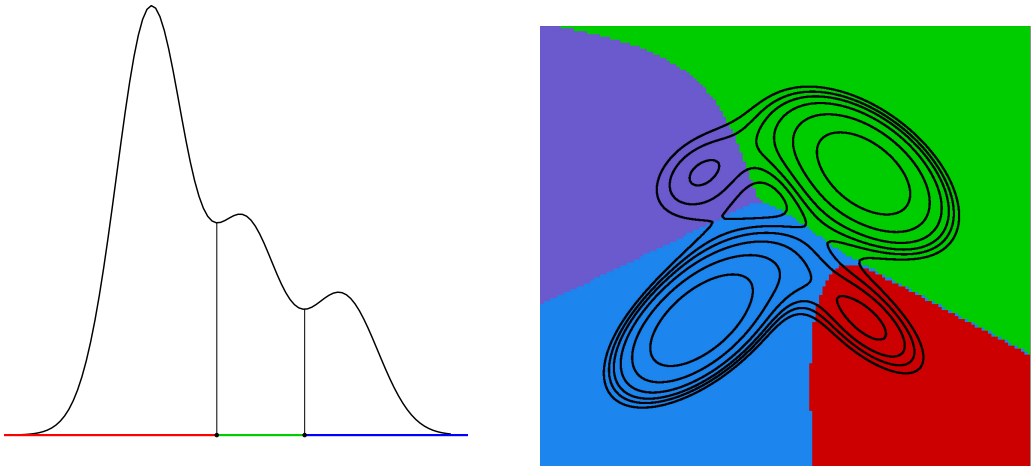


FIGURE 1.1: Ideal population clusters according to the modal approach in one dimension (left plot) and in two dimensions (right plot).

of differential topology focusing on the large scale structure of an object via the analysis of the critical points of a function (see e.g Matsumoto, 2002, for an introduction).

More specifically, assume that the observed data $\mathbb{X} = \{x_1, \ldots, x_n\}$ are i.i.d. realizations of a continuous random variable X, with probability density function $f\colon \mathbb{R}^d \to \mathbb{R}$. Assume that f is a Morse function, i.e. a smooth enough function having nondegenerate critical points, and denote by $M_1, \ldots, M_K$ the modes of f (i.e. its local maxima). For a given initial value $x \in \mathbb{R}^d$, an *integral curve* of the negative density gradient $-\nabla f$ is defined as the path $\nu_x : \mathbb{R} \to \mathbb{R}^d$ such that

$$\nu_x'(t) = -\nabla f(\nu_x(t)), \qquad \nu_x(0) = x.$$

The set of points whose integral curve starts at a critical point x_0 (as $t \to -\infty$) goes under the name of *unstable manifold* of x_0 and is defined as

$$W_-^u(x_0) = \{x \in \mathbb{R}^d : \lim_{t \to -\infty} \nu_x(t) = x_0\}.$$

It has been showed (Thom, 1949) that the class of the unstable manifolds of every critical point of a Morse function yields a partition of the whole space. With these notions at hand, the modal ideal population clustering $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ associated to a density function f is then defined as the set of the unstable manifolds $\{W_-^u(M_1), \ldots W_-^u(M_K)\}$ of the modes of f. Equivalently, if the integral curves associated to the positive density gradient are considered, then a modal cluster is defined as the set of points whose integral curves converge (as $t \to +\infty$) at the same mode. By borrowing concepts from terrain analysis, the underlying intuition is that, if f is figured as a mountainous landscape

where the modes are the peaks, a modal cluster is the region that would be flooded by a fountain emanating from a peak. When $d = 1$, clusters are then unequivocally defined by the locations of the minima points of f, which represent the cluster boundaries. The concept of modal clusters as the domains of attraction of the density modes stems naturally from this definition. For a visual interpretation of the population clusters in one and two dimensions see Figure 1.1.
The outlined notion of cluster claims several reasons of attractiveness. It is not bound to a particular shape since, in contrast to the parametric counterpart, it is linked to features of the density without requiring assumptions on the true data generative mechanism. Moreover it is naturally complying with the geometric intuition of dense sets making it close to an intuitive grouping of data. Also, the number of clusters is an intrinsic property of the data generator mechanism, thereby well defined, at least conceptually, and estimable within the process of clustering itself.

1.3.2 Estimation and allocation procedures

All of the above-mentioned definitions of modal population clusters emphasize the crucial role of the density in this framework. Since f is practically unknown, from an operational point of view a density estimate is needed to determine the high density regions which govern the final clustering. Which specific estimator is employed depends on either conceptual or operational convenience, but the selection falls usually within a nonparametric formulation, to guarantee the flexibility of possibly identifying groups of arbitrary shape. In this framework the kernel density estimator represents the most common choice (see, for a recent account, Chacón and Duong, 2018) and is defined as

$$\hat{f}_{\mathbf{H}}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(x - x_i) \tag{1.10}$$

where $\mathbf{H}$ is a symmetric positive definite *bandwidth* matrix, $K_{\mathbf{H}}(x) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2} x)$ and $K(\cdot)$ is the *kernel*, usually a symmetric, smooth, non-negative function which integrates to 1.

While the choice of the function K has been proven not to have a strong impact on the resulting density estimate, an appropriate selection of $\mathbf{H}$ is critical. For the sake of simplicity, we assume in the following a diagonal structure for the bandwidth matrix, i.e. $\mathbf{H} = h^2 I$. If a too small value for h is chosen, the density estimate will be undersmoothed and possibly characterized by spurious modes. On the other hand large values of h will result in oversmoothed estimates possibly covering relevant features of

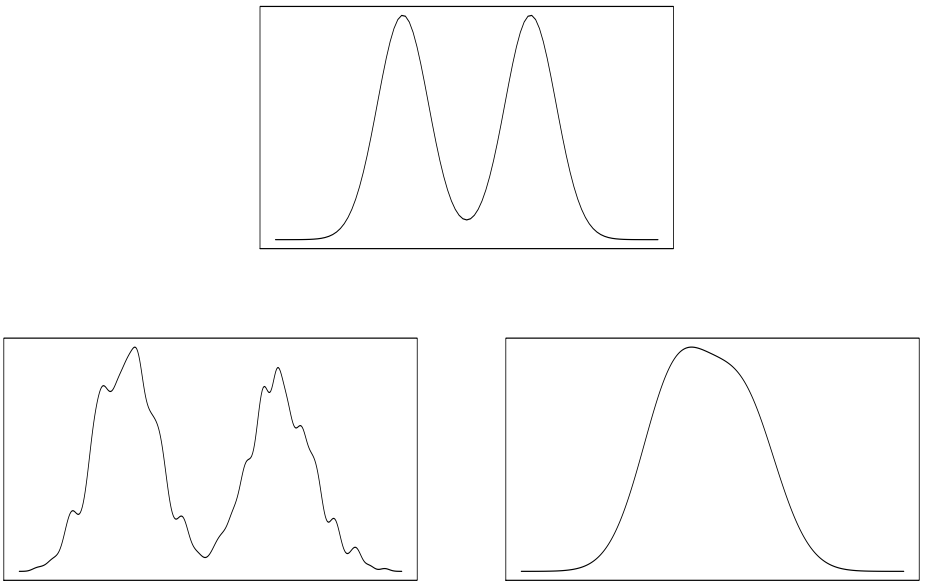


FIGURE 1.2: Top panel: two-component normal mixture density. Bottom panel: examples of under and oversmoothing density estimates, respectively on the left and on the right, using the kernel density estimator.

the density. A one-dimensional illustration of the key role played by the bandwidth in this framework is given in Figure 1.2.

Due to the pivotal role of the bandwidth matrix, several efforts have been made in order to address its selection. The usual way forward consists in selecting the smoothing parameter which minimizes some measure of distance between the estimated and the true density. A common choice is then represented by the *Integrated Squared Error*, defined as

$$\mathrm{ISE}(h) = \int_{\mathbb{R}^d} \{\hat{f}_h(x) - f(x)\}^2 dx. \tag{1.11}$$

Since the (1.11) depends on the observed data usually its expected value is considered, not being subject to random variability that could hinder the bandwidth selection problem (see Hall and Marron, 1991). The *Mean Integrated Squared Error* is then defined as

$$\mathrm{MISE}(h) = \mathbb{E}\left[\mathrm{ISE}(h)\right] \tag{1.12}$$

and commonly considered as a non-stochastic error distance. The optimal bandwidth, according to the MISE, is subsequently obtained as $h_{\mathrm{MISE}} = \mathrm{argmin}_{h>0}\,\mathrm{MISE}(h)$. Note that the minimization process does not lend itself to closed form solution, therefore often the MISE asymptotic counterpart – the AMISE – is considered instead.

Since all of the above discrepancy measures are depending on the true and unknown density function f, some approaches to estimate them is needed. In the last decades several proposals have been made as, for example, the ones based on *least squares cross validation*, *biased cross validation* or *plug-in bandwidth selectors*. A comprehensive review of these methods is beyond the scopes of this thesis and can be found in Silverman (1986), Wand and Jones (1995), and more recently in Chacón and Duong (2018).

Once that an estimate of the density has been obtained, the following question to be addressed is concerned with the need of operationally identifying its modal regions to partition the observed data.

A first strand of methods aims at detecting directly the modes of the density and then associates each data point to the pertaining mode coherently with the definition given in the Section 1.3.1. Most of the contributions moving in this direction turn out to be numerical optimization methods mainly based on the *mean-shift algorithm*. The algorithm has been proposed by Fukunaga and Hostetler (1975) but it has been brought to new life only more recently, with the computational advancements, by the work of Cheng (1995) and the variants proposed by Comaniciu and Meer (2002) and Carreira-Perpinán (2008).

The mean-shift transforms an initial point $x^{(0)}$ recursively, and identifies a sequence $(x^{(0)}, x^{(1)}, x^{(2)}, \dots)$ according to an updating mechanism defined as

$$x^{(l+1)} = x^{(l)} + A\frac{\nabla \hat{f}(x^{(l)})}{\hat{f}(x^{(l)})}\,, \tag{1.13}$$

where A is a $d \times d$ positive definite matrix chosen to guarantee the convergence to a local maximum of f and $\nabla \hat{f}$ is the gradient of $\hat{f}$. In practice, at each iteration the algorithm moves a generic data point along the steepest ascent path of the gradient of a kernel estimate, until converging to a mode. The final partition is then straightforwardly obtained by applying the mean-shift algorithm to each observed data point and by grouping them in the same cluster if they ascend to the same mode.

A similar approach to the identification of the modal regions has been proposed by Li *et al.* (2007) under the name of *Modal EM algorithm* (MEM). This technique aims at seeking the local maxima of the density by exploiting the peculiar mixture structure of the kernel density estimator. Apart from being built on a mixture construction, the algorithm shares some some other connections with the parametric formulation of cluster analysis since it alternates between two iterative steps in the guise of the EM algorithm introduced in Section 1.2.2. Despite the contact points among the two algorithms it should be noted that they aim at completely different scopes: while the EM searches

for the maximum of the likelihood in order to provide parameters estimates, the MEM seeks directly for the local maxima of the density function which work as the archetypes of the clusters.

A second strand of nonparametric clustering methods finds a direct inspiration from the definition of modal cluster given in Hartigan (1975). This approach, instead of associating clusters directly to the modes, links them to the connected components of the density level sets. Specifically, a section of f at a given level λ singles out the (upper) level set

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\} \;\; 0 \leq \lambda \leq \max f \tag{1.14}$$

which may be connected or disconnected. In the latter case, it consists of a number of connected components, each of them associated with a cluster at level λ.

While there may not exist a single λ which catches all the modal regions, any connected component of $L(\lambda)$ includes at least one mode of the density and, on the other hand, for each mode there exists some λ for which one of the connected components of the associated $L(\lambda)$ includes this mode at most. Hence, not only it is not necessary to define a specific level λ to identify the groups, which would be difficult and often not effective in providing the overall number of modes, but conversely, all the modal regions may be detected by identifying the connected components of $L(\lambda)$ for different λs. Varying λ along its range gives rise to a hierarchical structure of the high-density sets, known as the *cluster tree*. For each λ, it provides the number of connected components of $L(\lambda)$, and each one of its leaves corresponds to a *cluster core*, i.e. the largest connected component of $L(\lambda)$ including one mode only.
Operationally $L(\lambda)$ will be estimated by substituting f in (1.14) with its nonparametric estimate $\hat{f}$. Afterwards the detection of its connected components is required in order to practically obtained a partition. In the multivariate domain this turns out to be an awkward issue that, in the past, has seriously limited the use of this approach to cluster analysis. Only quite recently the question has been addressed with the aid of the graph theory. More specifically, let $\mathcal{G}$ be a graph with vertices given by $x_1, \ldots, x_n$. A suitable subgraph $\mathcal{G}_\lambda$, induced by the sample level set

$$\mathcal{S}(\lambda) = \{x_i \in (x_1, \ldots, x_n) : \hat{f}(x_i) \geq \lambda\} \tag{1.15}$$

is built, by removing from $\mathcal{G}$ the vertices not in (1.15) and all the edges involving at least one of these vertices. The connected components of $\mathcal{G}_\lambda$ are therefore straightforwardly determined by those observations connected through an edge in the graph. A crucial

aspect then revolves around choosing how to practically construct the graph and several efforts have been made along this direction. A suitable choice is represented by the nearest-neighbour graph proposed in Cuevas *et al.* (2000, 2001) and Stuetzle (2003) while Azzalini and Torelli (2007) exploit the Delaunay triangulation in order to build $\mathcal{G}$. Representing advancements of the two latter works, it is worth to mention also the proposals in Stuetzle and Nugent (2010) and Menardi and Azzalini (2014) where the graph is built according to a more "density informative" formulation; an edge between two vertices is indeed drawn if $\hat{f}$ does not show any valley in the segment joining them.

Obtaining operationally a partition turns out to be more tricky when considering level sets-based methods with respect to mode hunting approaches. While observations belonging to one of the cluster cores are naturally assigned to the group associated to the pertaining mode, data points not falling in any of the cores, referred to as *fluff* in Stuetzle and Nugent (2010), remain unallocated thus requiring some sort of classification tool. A tentative solution is provided in Azzalini and Torelli (2007) where each fluff point is assigned to the most likely cluster in terms of the density.

Chapter 2

On the selection of an appropriate bandwidth for modal clustering

2.1 Introduction

Especially in the initial stages of the analysis of a set of data, one wishes to gain insight about the nature of the phenomenon of interest, without imposing preconceived notions or models. This applies in particular when data exhibit non-Gaussian features, since the possible identification of such behaviour may aid to decide how to subsequently approach the analysis the most fruitfully. Often due to the unavailability of some relevant variable, either unobserved or unobservable, data exhibit some unlabeled heterogeneity, which typically arises in multimodal structures. In such situations, and a fortiori when the heterogeneity arises along with asymmetry and heavy tails, a suitable approach for group identification is the modal formulation of density-based clustering.

As highlighted in Chapter 1, modal clustering associates groups to the domains of attraction of the modes of the density supposed to underlie the data. The reasons for pursuing such approach, rather than its parametric counterpart, especially lie in the opportunity of finding groups without specific, predetermined shapes. Moreover, the number of clusters is an intrinsic property of the data generator mechanism, thereby its determination is itself an integral part of the estimation procedure.

The existence of a formalized notion of cluster, based on the features of the density, leads to the concept of *ideal population clustering*, i.e. clusters are defined in terms of the true distribution. By serving as a reference "ground truth" to aim at, this concept introduces a benchmark to evaluate the performance of data-based partitions. Additionally, the purpose of the analysis is not limited to simply produce a partition of the observed data; instead, a *whole-space clustering* can be obtained, that is a partition

of the whole sample space (Ben-David *et al.*, 2006; Chacón, 2015).

Despite the attractiveness of building clustering on the density underlying the data, such density is, in practice unknown, and its estimation assumes a key role in order to approximate the ideal population goal. While the modal formulation does not preclude using a parametric density estimate as a first step to perform a data-based modal clustering (Scrucca, 2016; Chacón, 2019), a long-standing practice resorts to nonparametric estimators. Precisely, in this chapter the focus lies on those estimators based on kernel smoothing (see e.g. Wand and Jones, 1995; Chacón and Duong, 2018).

Under- or over-smoothed estimates may lead to deceiving indications about the modal structure of the underlying density function, and this problem is usually quantified through some measure of the discrepancy between the estimate and the target density. In contrast, the aim of this work is to consider nonparametric density estimation as a tool for the final purpose of modal clustering, focusing on an appropriate metric to compare the partitions induced by the true and the estimated distribution.

Our main result provides an asymptotic approximation for the considered metric, which allows introducing new automatic bandwidth selection procedures specifically designed for nonparametric modal clustering. The accuracy of this approximation and the performance of the new methods in practice, with respect to the proposed error criterion, is extensively studied via simulations, and compared with some plausible competitors.

The rest of the Chapter is structured as follows. In Section 2.2 the distance criterion to target density estimation for modal clustering is presented, along with the main asymptotic result and its consequences. Section 2.3 contains the setup and results of the numerical experiments. A generalization to the multidimensional setting is discussed in Section 2.4. Finally, some concluding remarks are stated in Section 2.5.

2.2 Density estimation for modal clustering

2.2.1 Asymptotic bandwidth selection for modal clustering

As introduced in Section 1.3.2 nonparametric clustering is conducted via exploration of the modality structure of the probability density function f. Since f is usually unknown, an estimator is needed in order to provide an estimate eventually inducing a partition. In this work we focus our attention on the kernel density estimator (1.10) in a univariate setting, to ease the mathematical formalization. As a consequence the selection of a single bandwidth $h > 0$ is needed, instead of a complete bandwidth matrix.

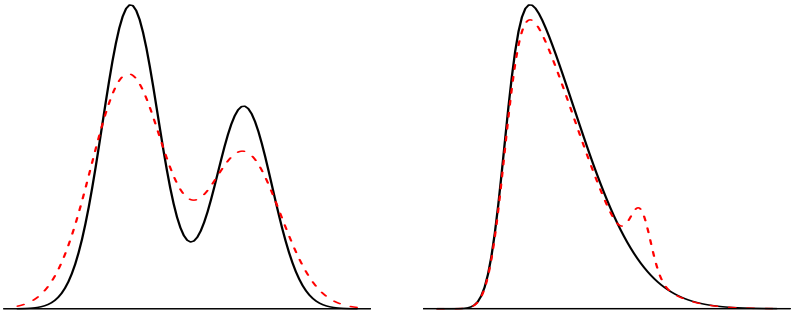


FIGURE 2.1: Left picture: two quite different densities, from an ISE perspective, inducing the same partition of the space. Right picture: two closer densities having different number of clusters.

In the previous Chapter we have seen how the smoothing parameter is usually selected by minimizing a suitable measure of distance among the true density and its kernel estimate, with a commonly considered measure being the ISE defined in (1.11). Bandwidth selectors based on the ISE or akin distances pursue the aim of obtaining an appropriate estimate of the density. However, the goal of modal clustering is markedly different from that of density estimation (see e.g. Cuevas *et al.*, 2001). In fact, two densities that are close with respect to the ISE may result in quite different clusterings while, on the other hand, densities far away from an ISE point of view could lead to the same partition of the space. A graphical illustration of this idea is provided in Figure 2.1. The inappropriateness of the ISE, or related distances, depends on its focus on the global characteristics of the density, while modal clustering strongly builds on specific and local features, more closely related to the density gradient or the high-density regions (see also Chen *et al.*, 2017). Therefore, the choice of the amount of smoothing makes sense to be tailored specifically for clustering purposes.

So far, the aim of choosing an amount of smoothing for the specific task of highlighting clustering structures has been scarcely pursued in literature. A related idea, although without particular reference to cluster analysis, has been developed by Samworth and Wand (2010), who propose a plug-in type bandwidth selector appropriate for estimation of highest density regions (see also Qiao, 2018; Doss and Weng, 2018). Another related work, more focused on the clustering problem, is the one by Einbeck (2011), where the author suggests considering the self-coverage measure as a criterion for bandwidth selection. Alternatively, the potential adequacy of a bandwidth selected to properly estimate the density gradient has been pointed out informally by Chacón and Duong (2013) and explored numerically by Chacón and Monfort (2006). The theoretical motivation of this suggestion lies on the strong dependence of both the population

modal clustering and the mean shift updating mechanism on the density gradient. The suggestion in Chen *et al.* (2016) follows the same rationale and the bandwidth is proposed to be selected as a modification of the normal reference rule for density gradient estimation.

To address the problem of bandwidth selection for modal clustering, an appropriate measure of distance should compare the data-based clustering induced by a kernel density estimate with the ideal population one. Stemming from Chacón (2015), a natural choice is the *distance in measure*, where the considered measure here is the probability $\mathbb{P}$ induced by the density f. Formally, let $\mathcal{C} = \{C_1, \ldots, C_r\}$ and $\mathcal{D} = \{D_1, \ldots, D_s\}$ be two partitions with $r \leq s$ (i.e. possibly different number of groups). The distance in measure between $\mathcal{C}$ and $\mathcal{D}$ is defined as

$$d(\mathcal{C}, \mathcal{D}) = \frac{1}{2} \min_{\sigma \in \mathcal{P}_s} \left\{ \sum_{i=1}^{r} \mathbb{P}(C_i \Delta D_{\sigma(i)}) + \sum_{i=r+1}^{s} \mathbb{P}(D_{\sigma(i)}) \right\}, \tag{2.1}$$

where $C \Delta D = (C \cap D^c) \cup (C^c \cap D)$ is the symmetric difference between any two sets C and D and $\mathcal{P}_s$ denotes the set of permutations of $\{1, 2, \ldots, s\}$. When $r > s$ we can easily define the distance in measure between $\mathcal{C}$ and $\mathcal{D}$ as $d(\mathcal{D}, \mathcal{C})$.

This distance finds an interpretation as the minimal probability mass that would need to be re-labeled to transform one clustering into the other (see Figure 2.2 for a graphical illustration). In this sense, the second term in (2.1) serves as a penalization for unmatched clusters in one of the clusterings. Practically, this distance conveys the idea that two partitions are similar not when they are physically close, but when the differently-labeled points do not represent a significant portion of the distribution.

It should be noted that the choice of this distance to evaluate the performance of a data-based clustering is not arbitrary. Indeed, many other possibilities are described in Meila (2016), but the conclusion of that study is that the distance in measure (called misclassification error there) is "the distance that comes closest to satisfying everyone". Furthermore, in Von Luxburg (2010) the distance in measure is considered as "the most convenient choice from a theoretical point of view".

As with the ISE-MISE duality, the distance in measure is a stochastic error distance, so for the purpose of bandwidth selection it seems more convenient to focus on the *Expected Distance in Measure*

$$\mathrm{EDM}(h) = \mathbb{E}\big[d(\hat{\mathcal{C}}_h, \mathcal{C}_0)\big], \tag{2.2}$$

where $\hat{\mathcal{C}}_h$ is the data-based partition induced by $\hat{f}_h$ and $\mathcal{C}_0$ represents the ideal population

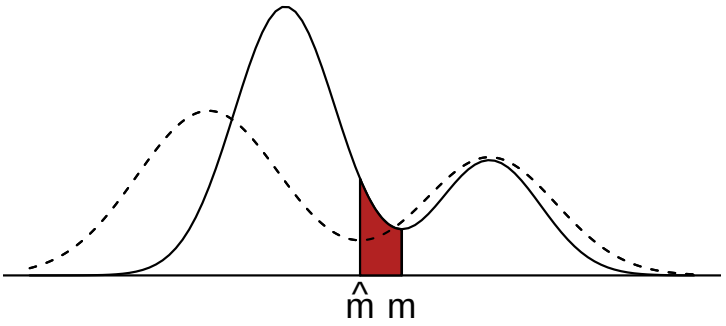


FIGURE 2.2: Graphical interpretation of the distance in measure: the shaded area represents the probability mass that would need to be re-labeled to transform one induced clustering into the other.

clustering. Once the appropriate error distance is defined, the optimal bandwidth h is given by $h_{\mathrm{EDM}} = \mathrm{argmin}_{h>0} \mathrm{EDM}(h)$.

As it happened with h_{MISE}, it does not seem possible to find an explicit expression for h_{EDM}. Hence, our goal will be to obtain an asymptotic form for the EDM that allows deriving a simple approximation to h_{EDM}.

To this aim, consider a standard normal random variable Z, and denote by $\psi(\mu, \sigma^2) = \mathbb{E}|\mu + \sigma Z|$ for $\mu \in \mathbb{R}$ and $\sigma > 0$. Since $|\mu + \sigma Z|$ has a folded normal distribution (Leone *et al.*, 1961), it follows that $\psi(\mu, \sigma^2)$ can be explicitly expressed as

$$\begin{aligned} \psi(\mu, \sigma^2) &= (2/\pi)^{1/2} \sigma e^{-\mu^2/(2\sigma^2)} + \mu\big\{1 - 2\Phi(-\mu/\sigma)\big\} \\ &= (2/\pi)^{1/2} \Big\{ \sigma e^{-\mu^2/(2\sigma^2)} + |\mu| \int_0^{|\mu|/\sigma} e^{-z^2/2} dz \Big\}, \end{aligned} \tag{2.3}$$

where Φ denotes the distribution function of Z. This function ψ plays a key role in the asymptotic behavior of the expected distance in measure, as the next result shows.

Theorem 2.1. *Assume that f is a bounded Morse function with compact support, $r \geq 2$ modes and local minima $m_1 < \cdots < m_{r-1}$, three-times continuously differentiable around each m_j, that $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$, and that the kernel K is supported on $(-1, 1)$, has four bounded derivatives and satisfies $\int_{-\infty}^{\infty} K(x) dx = 1$, $\int_{-\infty}^{\infty} xK(x) dx = 0$ and $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x) dx < \infty$. Define $R(K^{(1)}) = \int_{-\infty}^{\infty} K^{(1)}(x)^2 dx$ and suppose also that $h \equiv h_n$ is such that $h \to 0$, $nh^5 / \log n \to \infty$ and $(nh^7)^{-1}$ is bounded. Then,* $\mathrm{EDM}(h)$ *is*

asymptotically equivalent to

$$\mathrm{AEDM}(h) = \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} \psi\Big(\tfrac{1}{2}\mu_2(K) f^{(3)}(m_j) h^2, R(K^{(1)}) f(m_j)(nh^3)^{-1}\Big), \tag{2.4}$$

where $g^{(k)}$ refers to the k-th derivative of a function $g(\cdot)$.

Proof. From Theorem 4.1 in Chacón (2015) it follows that, with probability one, there exists $n_0 \in \mathbb{N}$ such that the kernel density estimator $\widehat{f}_h$ has the same number of local minima as f for all $n \geq n_0$. Let us denote by $\widehat{m}_{h,1} < \cdots < \widehat{m}_{h,r-1}$ the local minima of $\widehat{f}_h$. Then, the expected distance in measure between the data-based clustering $\widehat{\mathcal{C}}_h$ and the population clustering $\mathcal{C}_0$ can be written as

$$\mathrm{EDM}(h) = \sum_{j=1}^{r-1} \mathbb{E}|F(\widehat{m}_{h,j}) - F(m_j)|. \tag{2.5}$$

Write, generically, $\widehat{m}$ and m for any of the estimated and true local minima. A Taylor expansion with integral remainder allows writing

$$F(\widehat{m}) - F(m) = (\widehat{m} - m) \int_0^1 f\big(m + t(\widehat{m} - m)\big) dt.$$

The assumptions imply that $\widehat{m} \to m$ almost surely (see, for instance, Romano, 1988) and, since f is bounded and continuous, this readily yields $\int_0^1 f\big(m+t(\widehat{m}-m)\big)dt \to f(m)$ almost surely, which entails that $\mathbb{E}|F(\widehat{m}) - F(m)| \sim f(m)\mathbb{E}|\widehat{m} - m|$. The result then follows from Equation (2.6) in Grund and Hall (1995), where the asymptotic form of $\mathbb{E}|\widehat{m} - m|$ is given. □

The asymptotically optimal bandwidth h_{AEDM} is then defined as the value of $h > 0$ that minimizes $\mathrm{AEDM}(h)$. Due to the structure of $\psi(\cdot, 1)$, minimization of (2.4) is closely related to the problem of minimizing the L_1 distance in kernel density estimation and, in fact, reasoning as in Hall and Wand (1988) it is possible to show that h_{AEDM} is of order $n^{-1/7}$. Unfortunately, as it happened with h_{EDM}, it seems that neither h_{AEDM} admits an explicit representation. Hence, to get further insight into the problem of optimal bandwidth selection for density clustering, it appears necessary to rely on a tight upper bound for $\mathrm{AEDM}(h)$.

To find such a bound it is useful to note that many properties of $\psi(u, 1)$ are given in Devroye and Györfi (1985, Ch. 5), and can be translated to our function of interest by taking into account that $\psi(\mu, \sigma^2) = \sigma\psi(\mu/\sigma, 1)$. It follows that $\psi(\mu, \sigma^2)$ is symmetric

with respect to μ, nondecreasing for $\mu > 0$ and convex, attaining its minimum at $\mu = 0$ so that $\psi(\mu, \sigma^2) \geq \psi(0, \sigma^2) = (2/\pi)^{1/2}\sigma$ for all $\mu \in \mathbb{R}, \sigma > 0$.

By taking into account that $e^{-\mu^2/(2\sigma^2)}$ and $|1 - 2\Phi(-\mu/\sigma)|$ are both bounded by 1, Devroye and Györfi (1985) also noted that

$$\psi(\mu, \sigma^2) \leq (2/\pi)^{1/2}\sigma + |\mu| \tag{2.6}$$

for all $\mu \in \mathbb{R}, \sigma > 0$. However, a tighter bound for small values of μ is given in the next lemma.

Lemma 2.2. *The bound* $\psi(\mu, \sigma^2) \leq (2/\pi)^{1/2}\sigma + (2\pi)^{-1/2}\mu^2/\sigma$ *holds for all* $\mu \in \mathbb{R}$ *and* $\sigma > 0$.

Proof. From $\psi(\mu, \sigma^2) = \sigma\psi(\mu/\sigma, 1)$, it suffices to show that $\psi(u, 1) \leq (2/\pi)^{1/2} + (2\pi)^{-1/2}u^2$ for $u \geq 0$. From the definition of ψ, this is equivalent to proving that $\alpha(u) \leq 1$, where $\alpha(u) = e^{-u^2/2} + u\int_0^u e^{-z^2/2}dz - u^2/2$. Since $\alpha(0) = 1$, it is enough to show that α is nonincreasing, but this immediately follows from the fact that $\alpha'(u) = \int_0^u e^{-z^2/2}dz - u$. □

The bound in Lemma 2.2 is tighter than (2.6) whenever $|\mu| \leq (2\pi)^{1/2}\sigma$, but the situation reverses for bigger values of $|\mu|$, so that none of the two bounds is uniformly better (see Figure 2.3) hence we should keep track of both of them. They lead to upper bounds for the asymptotic EDM.

Corollary 2.3. *Under the conditions of Theorem 2.1, the asymptotic EDM satisfies* $\mathrm{AEDM}(h) \leq \min\{\mathrm{AB1}(h), \mathrm{AB2}(h)\}$ *for all* $h > 0$, *where*

$$\begin{aligned}
\mathrm{AB1}(h) &= (2/\pi)^{1/2}R(K^{(1)})^{1/2}bn^{-1/2}h^{-3/2} + \tfrac{1}{2}\mu_2(K)a_1h^2, \\
\mathrm{AB2}(h) &= (2/\pi)^{1/2}R(K^{(1)})^{1/2}bn^{-1/2}h^{-3/2} \\
&\quad + (32\pi)^{-1/2}\mu_2(K)^2R(K^{(1)})^{-1/2}a_2n^{1/2}h^{11/2}.
\end{aligned}$$

Here, $b = \sum_{j=1}^{r-1} b_j$ *and* $a_\ell = \sum_{j=1}^{r-1} a_{j\ell}$ *for* $\ell = 1, 2$, *where*

$$\begin{aligned}
a_{j1} &= f(m_j)|f^{(3)}(m_j)|/f^{(2)}(m_j), & b_j &= f(m_j)^{3/2}/f^{(2)}(m_j), \\
a_{j2} &= f(m_j)^{1/2}f^{(3)}(m_j)^2/f^{(2)}(m_j).
\end{aligned}$$

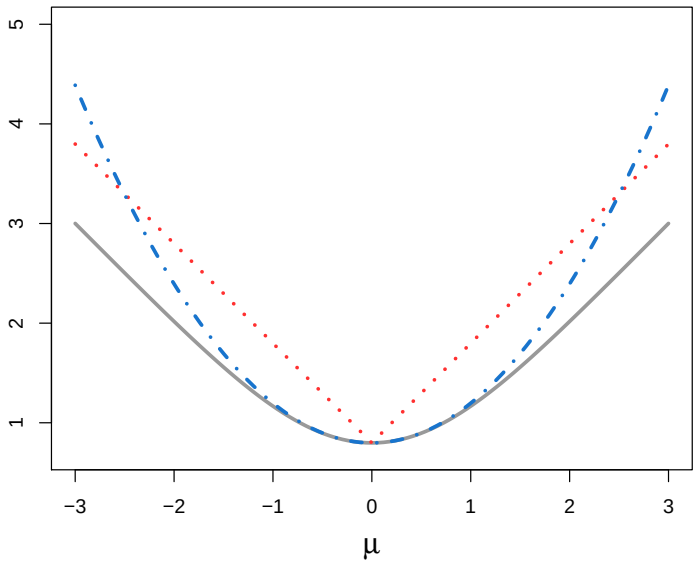


FIGURE 2.3: Graph of $\psi(\mu, 1)$ as a function of μ (grey solid curve), together with the bound (2.6) (red dotted line) and the bound from Lemma 2.2 (blue dot-dashed curve).

The minimizers of AB1(h) *and* AB2(h) *can be computed explicitly, and are given by*

$$h_{\mathrm{AB1}} = \left(\frac{9R(K^{(1)})b^2}{2\pi\mu_2(K)^2 a_1^2} \right)^{1/7} n^{-1/7} \tag{2.7}$$

$$h_{\mathrm{AB2}} = \left(\frac{24R(K^{(1)})b}{11\mu_2(K)^2 a_2} \right)^{1/7} n^{-1/7} . \tag{2.8}$$

2.2.2 Some remarks

In this section we discuss in more depth some of the results derived in Section 2.2.1. The aim is to provide insights on the behavior of the approximations and bandwidth selectors and to discuss possible competitors.

Remark 2.1 Theorem 2.1 provides an asymptotic expression for the EDM that is valid as long as the true density has two or more modes. When the true density is unimodal ($r = 1$), expression (2.4) is not well-defined. However, under the assumptions of the theorem the kernel estimator is also unimodal with probability one for big enough n. Thus, asymptotically the distance in measure would be identically zero, hence the AEDM formula would remain valid under the usual convention setting $\sum_{j=1}^{0} = 0$.

Moreover, for unimodal densities the numerical work in Section 2.3 suggests that there exists $h_0 > 0$ such that $\mathrm{EDM}(h) = 0$ for all $h \geq h_0$. Hence, in that case it seems sensible to define $h_{\mathrm{EDM}} = \inf\{h > 0 \colon \mathrm{EDM}(h) = 0\}$.

Remark 2.2 A natural estimator of the density first derivative is the first derivative of

the kernel density estimator. For this estimator it is possible to define the MISE as in (1.12), and to consider its minimizer $h_{\text{MISE},1}$ and its asymptotic approximation $h_{\text{AMISE},1}$ (see Singh, 1987; Chacón *et al.*, 2011). The bandwidths (2.7) and (2.8) share the same order as $h_{\text{AMISE},1}$, whose expression is given by

$$h_{\text{AMISE},1} = \left(\frac{3R(K^{(1)})}{\mu_2(K)^2 R(f^{(3)})} \right)^{1/7} n^{-1/7}, \tag{2.9}$$

with $R(f^{(3)}) = \int_{-\infty}^{\infty} f^{(3)}(x)^2 dx$. This consideration strengthens the intuition, outlined in Section 2.2.1, that (2.9) could be an adequate bandwidth choice for modal clustering purposes.

Remark 2.3 By explicitly plugging expression (2.3) for ψ into (2.4), it is easily seen that the AEDM can be decomposed into two summands. Studying their behavior, as a function of h, it can be checked that when $h \to 0$ the first term decreases while the second one tends to increase. Vice versa, when h increases, the opposite behaviour is witnessed. A similar trade-off occurs with the decomposition of the AMISE into the *Asymptotic Integrated Squared Bias* and the *Asymptotic Integrated Variance*, which are minimized for diverging values of h.

Remark 2.4 If the true density is locally symmetric around its minima, the considerations in the previous item do not hold anymore. Symmetry around a minimum m implies $f^{(k)}(m) = 0$, for any odd value of k. Therefore the first summand of the AEDM expression, related to the bias, vanishes, leading to a monotonically decreasing behavior of the AEDM itself. This would represent in principle a serious issue as in principle it prevents us from using the proposed bandwidth selector. However, such situation is highly unlikely to occur in practice, as motivated in Remark 2.5. A similar anomaly was observed in the related problem of mode estimation in Chernoff (1964): if the true density is symmetric around its mode, then Chernoff's mode estimator is unbiased. Hence, in some special cases symmetry plays a certain role in the performance of these smoothing methodologies.

Remark 2.5 The derived bandwidths depend on some unknown quantities such as the true density, its local minima and its second and third derivatives. In order to be of practical use we shall resort to plug-in strategies, that is, data-based bandwidth selectors will be proposed in the next section by substituting the aforementioned unknown quantities with pilot estimates. This is the same procedure that is commonly adopted when considering the plug-in bandwidth selector $\hat{h}_{\text{PI},1}$ for density gradient estimation (see Jones, 1992; Chacón and Duong, 2013). With reference to Remark 2.4, note that due to sample variability, resorting to the considered plug-in strategy makes highly

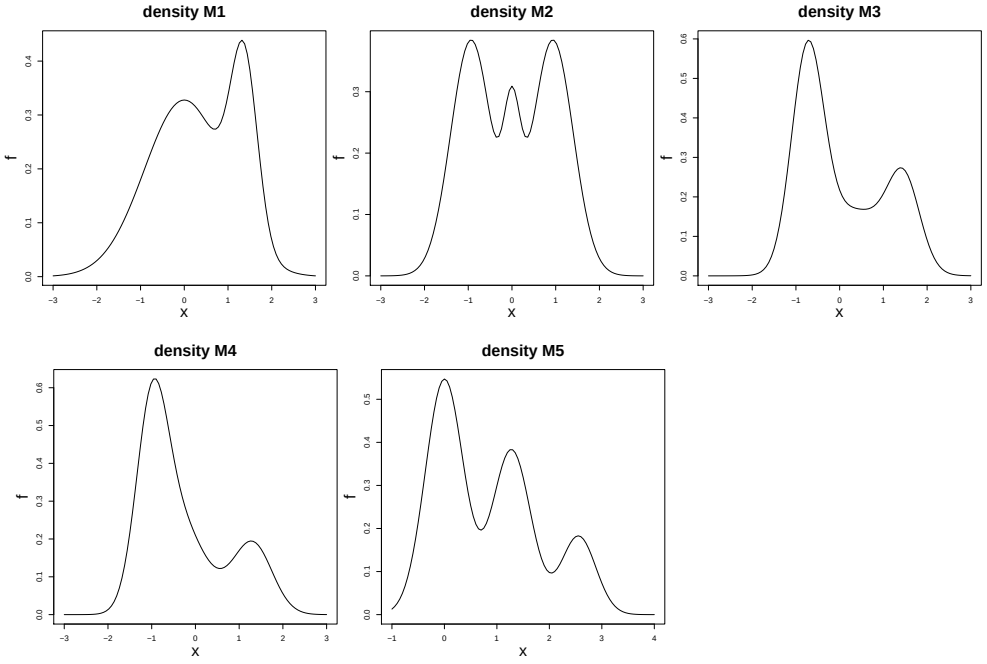


FIGURE 2.4: Univariate density functions selected for simulations.

unlikely to encounter a situation of perfect symmetry around a minimum in practice.

Remark 2.6 Theorem 2.1 assumes that f is a Morse function with compact support. Since the support of a probability distribution is always a closed set, any other assumption (smoothness, critical points, etc) is intended to be made with respect to the interior of this support. In practice any sample takes values in a bounded set, so we may extend the applicability of Theorem 2.1 to densities with unbounded support, provided that we consider their *significant support* (Baillo *et al.*, 2001), i.e. a subset of the support where most of the probability mass lies. More formally, the significant support of a density f is defined as the density level set $L(c) = \{x \in \mathbb{R} : f(x) > c\}$, where $c = c_\alpha$ is the largest constant such that $\mathbb{P}(L(c_\alpha)) \geq 1 - \alpha$, for some small $\alpha > 0$. Note that, by construction, the *significant support* is always bounded hence respecting the theorem's assumptions.

2.3 Numerical results

The idea of estimating the density for clustering purposes, via the minimization of the expected distance in measure – or its asymptotic counterpart – is explored in this section via simulations. All the analyses have been performed in the R environment (R Core Team, 2019) with the aid of the `ks` (Duong, 2019), `meanShiftR` (Lisic, 2018), `clue` (Hornik, 2018), and `multimode` (Ameijeiras-Alonso *et al.*, 2018) packages, as well as a number of routines specifically designed for the scope. All the functions to implement

the proposed selectors will be made available publicly.

A total of $B = 1000$ samples for each of the sizes $n \in \{100, 1000, 10000\}$ are generated from the univariate densities depicted in Figure 2.4 and whose parameters are reported in the Appendix. The selected densities are designed to illustrate different modal structures to encompass different possible behaviors from a clustering perspective. In order to respect the assumptions of Theorem 2.1, in the following analysis we restrain results to the *significant support* of the considered densities, as discussed in Remark 2.6, with $\alpha = 0.01$.

The first goal of the study was to evaluate the quality of the asymptotic approximation of the EDM and the behavior of the two bounds derived in Corollary 2.3. Since an explicit expression for the EDM was not available, we obtained a Monte Carlo approximation based on the $B = 1000$ synthetic samples.

The plots displayed in Tables 2.1 to 2.5 show the behavior of the asymptotic approximations, with respect to the EDM, as a function of the bandwidth h. As expected, the approximations improve as the sample size increases. The two bounds show a quite different behavior, with characteristics that reflect the theoretical properties pointed out in Section 2.2.2. The first bound is closer to the AEDM in uniform terms, but despite having a diverging behavior for large h the second bound is usually closer to the AEDM around the location of the minimizer h_{AEDM}.

With regard to the EDM, it presents a nearly flat pattern around its minimizer, thus suggesting a range of plausible bandwidths with very similar performance as the optimal one. This is especially true for densities with a simpler modal structure, captured by the kernel estimate for a wide range of bandwidth values.

To appreciate how much is lost by changing the target from the optimal h_{EDM} to the oracle surrogates h_{AEDM} and $h_{\text{MISE},1}$, the first three lines in each table also present the values for the corresponding EDM, all computed under a full knowledge of the density and its involved features. By construction, $\text{EDM}(h_{\text{EDM}})$ is the lowest of these values and, being derived as an asymptotic approximation, the oracle h_{AEDM} stands close to this optimal value, especially for larger sample sizes. However, it is remarkable that $h_{\text{MISE},1}$, despite being based on a different optimality criterion, also leads to comparable or even improved results over h_{AEDM} in terms of the EDM.

As a second goal, we propose new data-based bandwidth selectors specifically designed for modal clustering purposes. The first step consists in estimating the number of local minima, and their location. This is achieved by numerically finding the roots of a pilot estimate of $f^{(1)}$, constructed as the derivative of the kernel density estimator using the plug-in gradient bandwidth $\hat{h}_{\text{PI},1}$. Then, similarly, we obtain pilot estimates

TABLE 2.1: Top panel: the EDM (solid line), the AEDM (dashed grey line), and the bounds AB1 (dotted line) and AB2 (dot-dashed line) versus h, for $n = 100, 1000, 10000$. All the expressions are evaluated by assuming f and all the involved quantities known. The minimum EDM is reported below the plots, together with the EDM for the oracle bandwidths h_{AEDM} and $h_{\text{MISE},1}$. Middle panel: average distances in measure (and their standard error) for the proposed bandwidth selectors and the plug-in bandwidth for density gradient estimation. Bottom panel: percentages of times when the estimated number of cluster $\hat{r}$ matches the true one r. Results refer to density M1.

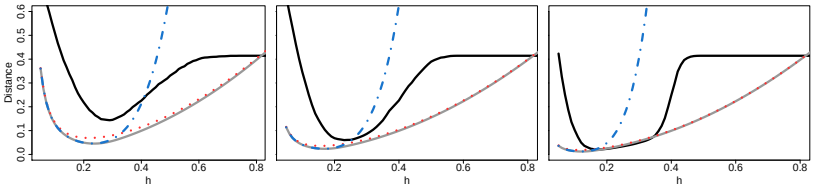
	n =100	n=1000	n=10000
			
h_{EDM}	0.144	0.060	0.020
h_{AEDM}	0.164	0.103	0.050
$h_{\text{MISE},1}$	0.146	0.081	0.044
$\hat{h}_{\text{AEDM}}$	0.267 (0.173)	0.103 (0.130)	0.045 (0.075)
$\hat{h}_{\text{AB1}}$	0.256 (0.174)	0.105 (0.127)	0.056 (0.084)
$\hat{h}_{\text{AB2}}$	0.265 (0.173)	0.102 (0.129)	0.048 (0.079)
$\hat{h}_{\text{PI},1}$	0.221 (0.176)	0.063 (0.084)	0.029 (0.052)
$\%\,\hat{r} = r$	54.5	91.7	92.6

TABLE 2.2: Cf. Table 2.1. Results refer to density M2.

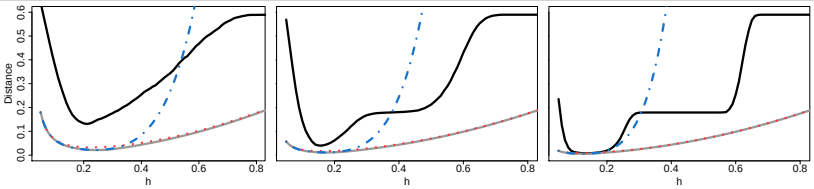
	n =100	n=1000	n=10000
			
h_{EDM}	0.131	0.040	0.008
h_{AEDM}	0.143	0.047	0.008
$h_{\text{MISE},1}$	0.165	0.041	0.011
$\hat{h}_{\text{AEDM}}$	0.324 (0.200)	0.061 (0.070)	0.010 (0.016)
$\hat{h}_{\text{AB1}}$	0.301 (0.195)	0.053 (0.066)	0.011 (0.017)
$\hat{h}_{\text{AB2}}$	0.318 (0.199)	0.058 (0.069)	0.010 (0.016)
$\hat{h}_{\text{PI},1}$	0.256 (0.159)	0.092 (0.076)	0.008 (0.005)
$\%\,\hat{r} = r$	2.8	58.0	100.0

TABLE 2.3: Cf. Table 2.1. Results refer to density M3.

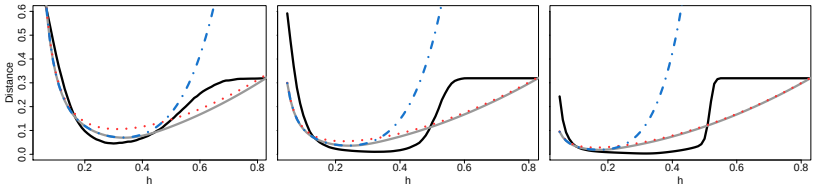
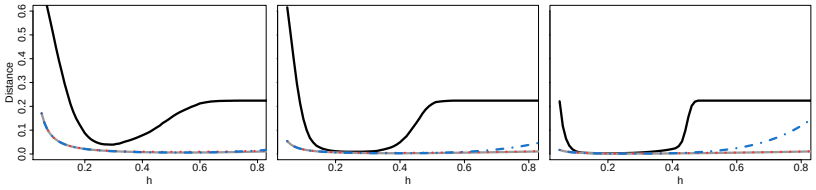
	n =100	n=1000	n=10000
			
h_{EDM}	0.045	0.010	0.003
h_{AEDM}	0.051	0.016	0.011
$h_{\text{MISE},1}$	0.054	0.034	0.022
$\hat{h}_{\text{AEDM}}$	0.090 (0.110)	0.039 (0.057)	0.026 (0.036)
$\hat{h}_{\text{AB1}}$	0.087 (0.104)	0.042 (0.058)	0.028 (0.035)
$\hat{h}_{\text{AB2}}$	0.091 (0.109)	0.040 (0.058)	0.026 (0.036)
$\hat{h}_{\text{PI},1}$	0.050 (0.072)	0.024 (0.025)	0.019 (0.017)
$\%\,\hat{r} = r$	91.0	91.6	88.1

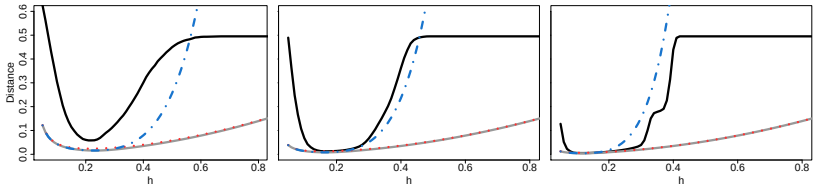
TABLE 2.4: Cf. Table 2.1. Results refer to density M4.

	n =100	n=1000	n=10000
			
h_{EDM}	0.039	0.009	0.003
h_{AEDM}	0.187	0.040	0.005
$h_{\text{MISE},1}$	0.040	0.015	0.005
$\hat{h}_{\text{AEDM}}$	0.077 (0.088)	0.030 (0.057)	0.007 (0.016)
$\hat{h}_{\text{AB1}}$	0.074 (0.086)	0.029 (0.053)	0.009 (0.021)
$\hat{h}_{\text{AB2}}$	0.076 (0.089)	0.030 (0.057)	0.007 (0.017)
$\hat{h}_{\text{PI},1}$	0.051 (0.069)	0.011 (0.014)	0.005 (0.005)
$\%\,\hat{r} = r$	85.4	97.2	99.8

of f, $f^{(2)}$ and $f^{(3)}$ at the estimated local minima using kernel estimates with the same bandwidth $\hat{h}_{\text{PI},1}$. These quantities are subsequently plugged-in in the formulas of the AEDM, AB1 and AB2, and the minimizers of the resulting estimated criteria are found; in the case of the estimated AEDM by numerical minimization, and according to expressions (2.7) and (2.8) for AB1 and AB2 respectively. The data-based bandwidths thus obtained are denoted $\hat{h}_{\text{AEDM}}$, $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$, respectively.

Occasionally (although rarely) the first step in the procedure above yielded a single

TABLE 2.5: Cf. Table 2.1. Results refer to density M5.

	n =100	n=1000	n=10000
			
h_{EDM}	0.058	0.012	0.005
h_{AEDM}	0.059	0.012	0.005
$h_{\text{MISE},1}$	0.058	0.013	0.005
$\hat{h}_{\text{AEDM}}$	0.160 (0.175)	0.017 (0.034)	0.006 (0.007)
$\hat{h}_{\text{AB1}}$	0.144 (0.169)	0.017 (0.030)	0.006 (0.007)
$\hat{h}_{\text{AB2}}$	0.157 (0.174)	0.017 (0.034)	0.006 (0.007)
$\hat{h}_{\text{PI},1}$	0.179 (0.158)	0.013 (0.009)	0.005 (0.003)
$\%\,\hat{r} = r$	42.7	99.7	100.0

mode, and then the AEDM was undefined. In those cases, and according to the rationale exposed in Remark 2.1, a sensible choice for h is the *critical bandwdith* proposed by Silverman (1981),

$$\hat{h}_{\text{crit}} = \inf\{h > 0 : \hat{f}_h(\cdot) \text{ has exactly one mode}\},$$

so in that case we set $\hat{h}_{\text{AEDM}} = \hat{h}_{\text{AB1}} = \hat{h}_{\text{AB2}} = \hat{h}_{\text{crit}}$.

Tables 2.1 to 2.5 also contain the Monte Carlo averages and standard deviations of the distances in measure obtained when performing modal clustering using the bandwidth selectors $\hat{h}_{\text{AEDM}}$, $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$. For completeness, their performance is also compared to that of $\hat{h}_{\text{PI},1}$, which so far probably represents their most sensible competitor in the clustering framework (see Chacón and Monfort, 2006).

In general, $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$ led to more accurate clusterings than $\hat{h}_{\text{AEDM}}$, with a slight preference for $\hat{h}_{\text{AB1}}$. The gradient-based bandwidth $\hat{h}_{\text{PI},1}$, in turn, not only produces competitive results, but its Monte Carlo average distance in measure appears lower than the one produced by the asymptotic EDM minimizers. In fact, a deeper insight into the standard errors of the obtained distances shows that $\hat{h}_{\text{AEDM}}$, as well as $\hat{h}_{\text{AB1}}$ and $\hat{h}_{\text{AB2}}$, produce more variable results. The higher variability seems to be due to the sensitivity of the minimizers to the plugged in pilot estimates, which strongly depend on local features of the density. Some further investigations, not fully reported here, suggest that the main responsible for this behaviour is not the pilot estimate of the local minima but the pilot density derivatives estimates at the minimum points. On the other hand,

while relying as well on some plug-in estimates, the gradient-based bandwidth $\hat{h}_{\mathrm{PI},1}$ produces more robust clusterings, as the quantities to be estimated refer conversely to global features of the density. As expected, this diverging behavior tends to vanish with increasing sample size since the asymptotic approximations improve. As a confirmation, with $n = 10000$, all the considered bandwidths perform comparably.

2.4 Multidimensional generalization

The concepts discussed so far refer to the one-dimensional setting where a mathematically rigorous treatment is feasible. The multidimensional generalization poses some difficulties since obtaining an asymptotic approximation of the EDM appears far from trivial. Hence, in order to gain some insight into the problem of selecting the amount of smoothing for nonparametric clustering in more than one dimension, some numerical comparisons are performed assuming the true density as known.

Denote by $f : \mathbb{R}^d \to \mathbb{R}$ the true density function and by

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} |\mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i)\right) \ , \tag{2.10}$$

its kernel estimate based on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and indexed by a symmetric positive definite $d \times d$ bandwidth matrix $\mathbf{H}$. The problem of bandwidth selection is considered by studying the EDM between the clustering induced by the kernel estimate $\hat{\mathcal{C}}_{\mathbf{H}}$ and the ideal population clustering $\mathcal{C}_0$. These clusterings are not so easily identifiable as in the unidimensional setting, due to the arbitrary forms that the cluster boundaries may adopt, however an approximation of the distance in measure $d(\hat{\mathcal{C}}_{\mathbf{H}}, \mathcal{C}_0)$ can be computed by resorting to a discretization scheme as follows (see Chacón and Monfort (2006) for further details):

1. Take a grid over the sample space and rule the grid by considering hyper-rectangles centered at each grid point.

2. Assign a cluster membership to each grid point by running a population version of the mean-shift algorithm i.e. using the true density. This produces a discretized version of $\mathcal{C}_0$.

3. Similarly, obtain the data-based partition $\hat{\mathcal{C}}_{\mathbf{H}}$ induced by $\hat{f}_{\mathbf{H}}$.

4. Compute the probability mass of each single hyper-rectangle in $\mathcal{C}_0$.

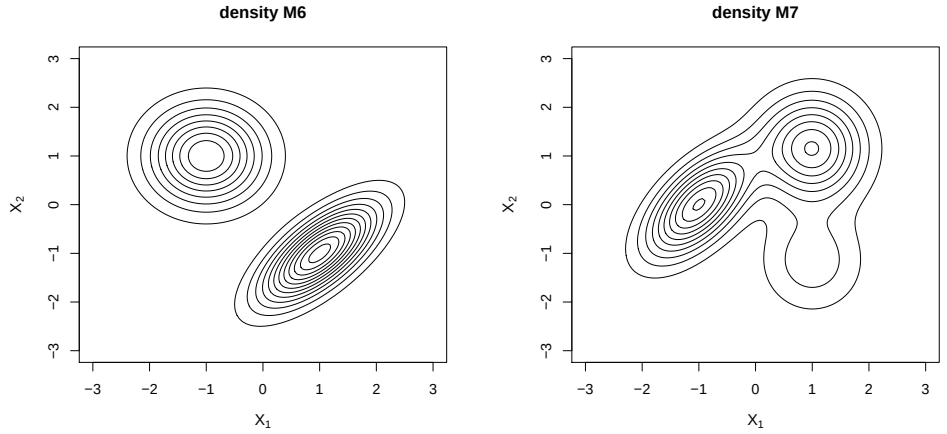


FIGURE 2.5: Bivariate density functions selected for simulations.

5. Compute the distance in measure as in (2.1) where the involved probabilities are evaluated based on the previous step.

For the multidimensional simulation study, a total of $B = 1000$ samples for each of the sizes $n \in \{100, 1000\}$ were generated from the bivariate densities whose contour plots are shown in Figure 2.5 and described in Appendix. The densities have been chosen to generalize the settings M1 and M5 included in the univariate study.

Three different parametrizations for the bandwidth matrix were considered: a scalar bandwidth $\mathbf{H} = h^2\mathbf{I}$, with $\mathbf{I}$ the identity matrix, a diagonal bandwidth $\mathbf{H} = \text{diag}(h_1^2, h_2^2)$, and a full, unconstrained bandwidth matrix $\mathbf{H}$. For density and density derivative estimation, Wand and Jones (1993) and Chacón *et al.* (2011) showed that the use of the simplest scalar bandwidth can be quite detrimental in practice, a diagonal bandwidth may suffice in some scenarios, but in general it is advantageous to employ unconstrained bandwidth matrices (see also Chacón and Duong, 2018). However, such results have never been obtained in a modal clustering framework; thus one of the goals of this simulation study is to examine how the bandwidth matrix parametrization affects the performances of the procedures.

Using the synthetic samples from each density in the study, it was possible to obtain a Monte Carlo estimate of the (discretized version of the) EDM, which was then minimized over the class of scalar, diagonal and unconstrained bandwidth matrices. The EDM was computed also for the MISE-optimal bandwidth for density gradient estimation over the same matrix classes. In both cases, the true density as well as all the involved quantities were assumed to be known. The EDM minimizers were determined numerically, by running the procedure over a grid of sensible values of the entries, while the optimal matrices for gradient estimation were determined as in Chacón *et al.* (2011).

TABLE 2.6: Minimum EDM associated with a density estimate with bandwidth matrix $\mathbf{H}$ selected to minimize the EDM ($\mathbf{H}_{\text{EDM}}$) and the MISE for gradient estimation ($\mathbf{H}_{\text{MISE},1}$). Different parametrizations for $\mathbf{H}$ are considered. In both cases, the true density as well as all the involved quantities are assumed to be known. Results refer to density M6.

	$\mathbf{H}_{\text{EDM}}$		$\mathbf{H}_{\text{MISE},1}$	
	n =100	n=1000	n =100	n=1000
$\begin{pmatrix} h^2 & 0 \\ 0 & h^2 \end{pmatrix}$	0.006	0.004	0.064	0.040
$\begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$	0.006	0.004	0.064	0.040
$\begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}$	0.005	0.003	0.042	0.024

The results are reported in Tables 2.6 and 2.7. Clustering based on the optimal bandwidth according to the EDM is very accurate in both of the considered examples, and improves considerably for increasing sample size. The use of more complex bandwidth parametrizations does not seem worth for modal clustering since results obtained with a full, unconstrained bandwidth matrix are comparable with those obtained with a scalar bandwidth, while the latter requires a substantially smaller computational effort.

In the multidimensional setting, the gradient bandwidth is quite competitive in terms of EDM, as in the univariate case. Again the comparable performance of unconstrained bandwidth matrices does not seem to justify the use of more complex parametrizations.

TABLE 2.7: Cf. Table 2.6. Results refer to density M7.

	$\mathbf{H}_{\text{EDM}}$		$\mathbf{H}_{\text{MISE},1}$	
	n =100	n=1000	n =100	n=1000
$\begin{pmatrix} h^2 & 0 \\ 0 & h^2 \end{pmatrix}$	0.114	0.044	0.116	0.054
$\begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$	0.114	0.042	0.115	0.055
$\begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}$	0.110	0.040	0.121	0.054

2.5 Conclusions

The modal clustering methodology provides a framework to perform cluster analysis with a clear and explicit population goal. It allows clusters of arbitrary shape and size, which can be captured by means of a nonparametric density estimator. In this context, the distance in measure represents a natural and easily interpretable error criterion. Therefore, in this chapter we have presented an asymptotic study of this criterion for the case where density estimates of kernel type are employed to obtain a whole-space clustering via the mean shift algorithm.

Our asymptotic approximations are useful to gain insight into the fundamental problem of bandwidth selection for modal clustering and, at the same time, serve as the basis to propose practical data-based bandwidth choices specifically designed for clustering purposes.

The finite-sample performance of the new proposals was investigated in a thorough simulation study, and compared to the oracle bandwidths i.e. the optimal choices when the true population is fully known. The gradient bandwidth, designed for the closely related problem of density gradient estimation, was also included as a natural competitor in the study.

The results of this simulation study have suggested that all the methods perform quite satisfactorily, and exhibit a very similar behavior for large sample sizes. For smaller samples, the performance of the gradient bandwidth was rather remarkable, since it obtained comparable or even better results than the new proposals, even without being specifically conceived for modal clustering.

This phenomenon resembles the conclusions obtained in Saavedra-Nieves *et al.* (2014) regarding the related problem of level set estimation. There, it was shown that the traditional bandwidth selectors for density estimation often outperformed more sophisticated methods designed for level set estimation purposes. The common pattern in both situations is that the optimal choices for the specific problems (level set estimation and modal clustering, respectively) depend on very subtle local features of the unknown density function, which are difficult to estimate, so that choices based on a more global, yet somehow related, perspective represent a sensible alternative.

Chapter 3

Ensemble density-based clustering

3.1 Introduction

In virtually any scientific domain we are witnessing an explosion in the availability of the data, coupled with a tremendous growth in their complexity. As a straightforward consequence, the number of choices that has to be made is increasing as well as the number of sophisticated modelling strategies proposed to deal with such newly introduced challenges. These choices are practically involved in any phase of the modelling process, spanning a wide landscape of possible options: from choosing a class of models or an appropriate approach to analyze a set of data, to more specific decisions as the selection of subsets of relevant variables or suitable parametrizations. Therefore, nowadays model selection steps, helping to formally extricate ourselves from the labyrinth of all these possible alternatives, are ubiquitous in any data analysis routine. Some commonly considered ways forward hence consist among the others in estimating a set of different models and then selecting the best one according to some information criterion (Claeskens and Hjort, 2008) or resorting to penalization schemes aimed at balancing fit and complexity (see Tibshirani *et al.* (2015) for an introduction).

Nevertheless, basing predictions and inference on a single model could turn out to be suboptimal. In the latter case, model averaging approaches have been proposed as a viable alternative, intended to estimate quantities by computing weighted averages of different estimates. Such approaches may lead to improvements in the estimation process by accounting for model uncertainty. In turn, from a predictive point of view, ensemble techniques have shown remarkable performances in a lot of different applications by building predictions as combinations of the ones given by a set of different models. Well established methods as *bagging*, *stacking*, *boosting* or the *random forests* (see Friedman *et al.*, 2001, for a review) have become the state of the art in the supervised learning

framework. Even if model averaging and ensemble approaches focus on different phases of the modelling process, respectively estimation and prediction, they share the same founding rationale as they aim to improve performances of the base models by combining their strenghts, while simultaneously circumventing their limits. For this reason the two expressions will be used interchangeably in the rest of the dissertation.

While extensively studied in the classification context, ensemble techniques have been scarcely pursued in the clustering one. A possible explanation can be found in the unsupervised nature of the problem itself; the absence of a response variable introduces relevant issues in evaluating the quality of a model and of the corresponding partition. As a consequence, weighting models in order to combine them turns out to be an awkward problem. Nonetheless mixing different partitions in a final one could in principle allows to combine clustering techniques based on different focuses to give a multiresolution view of the data and possibly improve the stability and the robustness of the solutions. Fern and Brodley (2003) exploit the concept of *similarity matrix* in order to aggregate partitions obtained on multiple random projections, while a similar approach is followed by Kuncheva and Hadjitodorov (2004) to study the concept of diversity among partitions. Monti *et al.* (2003) consider again a similarity matrix in order to evaluate the robustness of a discovered cluster under random resampling. In turn, the work by Strehl and Ghosh (2002) introduces three different solutions to the ensemble problem in the unsupervised setting by exploiting hypergraph representations of the partitions.

In this chapter we focus mainly on the parametric, or model-based, approach to cluster analysis where, as discussed in Chapter 1, a one-to-one correspondence among clusters and components of an appropriate mixture model is drawn. In this framework, the usual working routine is based on the *single best model paradigm*, i.e. a set of models is fitted and only the best one is chosen and considered to obtain a partition. The goal of the chapter is to go beyond this paradigm by proposing a model averaging methodology to give partitions resulting from an ensemble of models, thus possibly achieving a greater accuracy and robustness. Averaging is pursued directly on the estimated mixture densities in order to build a new and more accurate estimate which will be used to obtain a grouping of the data.

The rest of the chapter is organized as follows. In Section 3.2 the proposed methodology is outlined with specific attention to the estimation procedure. In Section 3.3 we discuss some specific aspects of our proposal and highlight connections with other models. Lastly in Section 3.4 we show the performances of our method on both simulated and real datasets, comparing them with some competitors. Section 3.5 presents some

concluding remarks.

3.2 Model averaging in model-based clustering

3.2.1 Framework and model specification

In the model-based clustering framework introduced in Section 1.2.1, the observed data $\mathbb{X} = (x_1, \ldots, x_n)', x_i \in \mathbb{R}^d$ are assumed to be generated from a density $f : \mathbb{R}^d \to \mathbb{R}$ adequately described by a mixture model. A partition of $\mathbb{X}$ is then obtained by associating clusters to the components of the mixture, in practice each observation is assigned to the most likely component. Since the estimation step is performed conditionally to the specification of the number of clusters, the choice of the model for each component and its parametrization, different models are usually fitted and the best one, according to an information criterion, is selected and used to obtain a clustering of the data.

We argue that this so called *single best model paradigm* could be sub-optimal especially when differences among values of the information criterion across competing models are close. In this setting mixing competitive models together may lead to a gain in robustness, stability and in the quality of the partition, as often witnessed in the supervised framework.

As an illustrative example we consider the widely known Iris dataset. In Figure 3.1 the left panel shows the partition obtained by the best model according to the BIC, a two-components VEV model (see Table 1.1 for details on the parametrization). On the right, the clustering induced by the second best model, a three-components VEV model, is shown. Even if no formal criteria are available in order to check if their difference is significant, the values assumed by the BIC for the two models appears quite close. Therefore it seems natural to ask if, discarding completely the second best model, useful information on the data is thrown away. In fact, the true labels indicate the presence of three groups, here adequately captured by the second best model.

In a model-based clustering framework the idea of mixing different models has been developed in order to obtain partitions based on an average of different models rather than on a single one. Both the works of Russell *et al.* (2015) and Wei and McNicholas (2015) propose a Bayesian model averaging approach to postprocess the results of model-based clustering. A key issue pointed out in both the proposals consists in the need of selecting an invariant quantity, i.e. a quantity having the same meaning across all the models in the ensemble, to average on. In parametric clustering this represents a cumbersome problem since the models to mix together could possibly have different number of groups; as a consequence, parameters spaces have different dimensions, thus

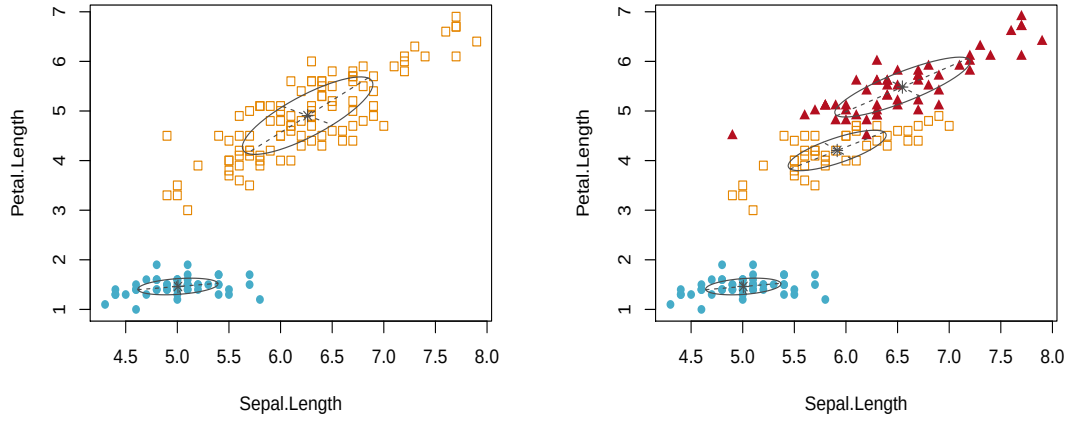


FIGURE 3.1: Example on Iris data: on the left the partition induced by the best model according to the Bayesian information criterion (BIC = −561.72). On the right the partition induced by the second best model (BIC = −562.55).

preventing the chance to average directly parameters estimates. Wei and McNicholas (2015) overcome this issue by introducing a component merging step in the procedure. Alternatively, Russell *et al.* (2015), consider the similarity matrix as the invariant quantity. They obtain an ensemble similarity matrix by averaging the candidate models ones. Afterwards the resulting matrix, where the $(i, j)th$ entry represents the averaged probability of x_i and x_j to belong to the same cluster, is considered to obtain partitions adopting a hierarchical clustering approach.

In this work we take a different path with respect to the ones mentioned above. The issue is tackled directly at its roots, by exploiting the essential role assumed by the density in the considered framework. Therefore, recasting the problem as a density estimation one, the density itself is chosen as the invariant quantity to be averaged. Let $\{f_m(\cdot|\hat{\Theta}_m)\}_{m=1,\dots,M}$ be a set of estimated candidate mixture models. In the rest of the chapter, we focus specifically on mixtures of normal densities, but the choice is not binding for the subsequent developments. Additionally, the number M of models to average is here considered as given, and we refer the reader to Section 3.3 for a discussion about this aspect. A new estimator, being a convex linear combination of the estimated densities $f_m(\cdot|\hat{\Theta}_m)$, is introduced:

$$\tilde{f}(x) = \sum_{m=1}^{M} \alpha_m f_m(x|\hat{\Theta}_m) \ , \tag{3.1}$$

with $\alpha_m > 0$, $\sum_m \alpha_m = 1$, representing the weight to assign to the mth model $\forall m =$

$1, \ldots, M$. A key aspect, as it will be discussed in Section 3.2.2, consists in properly estimate the model weights in order to guarantee that models describing more adequately the underlying density will count more in the resulting estimator.

The rationale behind our proposal heuristically exploits some results obtained by Rigollet and Tsybakov (2007). Here the authors show that, under some fairly general regularity assumptions, linearly aggregating density estimators leads asymptotically to an improvement in the resulting one under L_2-loss perspective. Hence, by possibly improving the quality of the density estimates, we aim at obtaining better characterizations of the relevant patterns in the data, leading to more refined partitions.

Even if the estimator in (3.1) is still a mixture model we cannot obtain a partition as usually carried out in parametric clustering, thus resorting to the one-to-one correspondence among groups and components. As an illustrative example let consider an ensemble formed by two mixture models, with two and three components. In this situation $\tilde{f}(\cdot)$ will result in a five component mixture model hence giving contradictory indications about the number of groups with respect to the models that have been mixed together. The problem is naturally circumvented by shifting the concept of cluster, and recasting it to the modal formulation hence searching for the modes of the estimated density and associating the groups to their domains of attraction.

The proposed solution, staying in the realm of density-based clustering, inherits and enjoys its relevant strenghts as the chance to frame the problem in a standard inferential setting where proper statistical tools can be employed for evaluation, and to obtain whole sample space partitions whose features are inferentially explorable. Moreover it has already been shown (see Scrucca, 2016; Chacón, 2019) that blending together parametric and nonparametric approaches to clustering coud lead to some relevant improvements in some, otherwise troublesome, situations.

3.2.2 Model estimation

The procedure outlined in Section 3.2.1 requires a practical way to estimate the density as in (3.1). Note that, since $\hat{\Theta}_m$ has been previously estimated, the only unknown parameters involved are the α_ms. These parameters represent the weights to be assigned at each single model in the ensemble, hence their estimation is crucial in governing the resulting shape of the density, its modal structure and consequently the final partition. A reasonable estimation procedure would result in giving nearly zero weights to those models in the ensemble which do not suitably capture the features of the underlying density, while weighting more the adequate ones.

In order to obtain an estimate for the weight vector $\alpha = (\alpha_1, \ldots, \alpha_M)$, we can aim at maximizing the log-likelihood of the model (3.1), defined as

$$\ell(\alpha) = \sum_{i=1}^{N} \log \sum_{m=1}^{M} \alpha_m f_m(x_i|\hat{\Theta}_m). \tag{3.2}$$

However, if the quantity in (3.2) is considered as the objective function to maximize, the procedure will incur in the overfitting problem since the most complex models in the ensemble, which provide a better fit by construction, will weight more. This behaviour will commonly result in wiggler estimates not appropriately seizing the relevant features of the density hence some regularization has to be considered in the estimation.

A tentative solution has been proposed by Smyth and Wolpert (1999) where a *stacking* procedure is adapted to the density estimation framework. The authors avoid to fall into the overfitting trap by exploiting a cross-validation scheme when combining the candidate models to obtain ensemble density estimates.

We take a different path by replacing the log-likelihood in (3.2) with a penalized version, generally defined as

$$\ell_P(\alpha) = \ell(\alpha) - \lambda g(\alpha, \nu) \ . \tag{3.3}$$

Here $g(\cdot)$ is a penalty function to be specified, $\nu = (\nu_1, \ldots, \nu_M)$ is a vector measuring the complexity of the models in the ensemble, while λ is a parameter controlling for the strength of the penalization. Within this general framework, we set ν_m to be the cardinality of $\hat{\Theta}_m$, as it appears a sensible proxy of the complexity of the *mth* model. Additionally, we consider $g(\alpha, \nu) = \sum_m \alpha_m \nu_m$ as a simple choice which guarantees a stronger penalization to the most complex models.

Due to the mixture structure easily recognizable in (3.1), and since the only unknown parameters are the mixture weights $\alpha_1, \ldots, \alpha_M$, we can resort to a slightly simplified version of the EM-algorithm in order to maximize the penalized log-likelihood (3.3). In the *E-step*, conditionally to an estimate $\hat{\alpha}^{(t)}$ for the vector α at iteration t, we compute

$$\tau_{mi}^{(t)} = \frac{\hat{\alpha}_m^{(t)} f_m(x_i|\hat{\Theta}_m)}{\sum_{m'=1}^{M} \hat{\alpha}_{m'}^{(t)} f_{m'}(x_i|\hat{\Theta}_{m'})} \ . \tag{3.4}$$

Then the *M-step* will consist in maximizing, with respect to α, the expected value of the complete-data penalized log-likelihood, in our setting expressed as

$$Q_p(\alpha; \hat{\alpha}^{(t)}) = \sum_{m=1}^{M} \sum_{i=1}^{n} \tau_{mi}^{(t)} [\log \alpha_m + \log f_m(x_i|\hat{\Theta}_m)] - \lambda \sum_{m=1}^{M} \alpha_m \nu_m \ , \tag{3.5}$$

under the constraint $\sum_m \alpha_m = 1$. Since closed form solutions are not available, $\hat{\alpha}^{(t+1)}$ is obtained by maximizing (3.5) numerically. As usual, the two steps will be iterated until a convergence criterion is met.

Regarding the choice of λ, some more caution is needed, since an accurate selection turns out to be essential in order to obtain a meaningful estimate properly reflecting the modal structure of the underlying density. In this work a few different options have been taken in consideration as, for example, the ones inspired to some information criteria as the *AIC-type* or the *BIC-type* penalizations. In our framework AIC and BIC may be expressed

$$\begin{aligned} \text{AIC} &= 2\ell(\alpha) - 2M \\ \text{BIC} &= 2\ell(\alpha) - \log(n)M \ , \end{aligned}$$

where M represents the cardinality of α being the number of parameters to estimate in (3.1). Stemming directly from these expressions, and according to the formulation in (3.3), we obtain two penalized log-likelihoods defined as

$$\ell_{P,AIC}(\alpha) = \ell(\alpha) - \sum_{m=1}^{M} \alpha_m \nu_m \tag{3.6}$$

$$\ell_{P,BIC}(\alpha) = \ell(\alpha) - \frac{\log(n)}{2} \sum_{m=1}^{M} \alpha_m \nu_m \ , \tag{3.7}$$

implying, as a consequence, $\lambda_{AIC} = 1$ and $\lambda_{BIC} = \log(n)/2$.

Another possible strategy consists in keeping λ unconstrained and estimating it by means of the observed data. A sensible approach resorts to a cross-validation strategy defined as follows:

- Partition the sample $\mathbb{X}$ randomly into S subsamples, where one subsample is retained as test set $\mathbb{X}_{\text{test}}$ while the remaining ones are used as a training set $\mathbb{X}_{\text{train}}$;
- Build a reasonable grid for the regularization parameter then, for each λ in the grid, obtain the estimated density based on $\mathbb{X}_{\text{train}}$ and then use it to compute $\tilde{f}(x_{\text{test}}|\mathbb{X}_{\text{train}}, \lambda)$ defined as the predicted density of $\mathbb{X}_{\text{test}}$;
- Repeat the previous steps for $s = 1, \ldots, S$ obtaining an out-of-sample predicted density estimate for the whole dataset $\mathbb{X}$;

- Select

$$\lambda_{\text{CV}} = \arg\max \ell_{\text{test}}(\lambda)$$

with $\ell_{\text{test}}(\lambda) = \sum_{x \in \mathbb{X}_{\text{test}}} \tilde{f}(x|\mathbb{X}_{\text{train}}, \lambda)$ a test log-likelihood. The selected λ_{CV} is finally used to estimate the vector of weights α based on the whole sample.

Although requiring an higher computational effort, this approach introduces some relevant advantages in the regularization process. By resorting to a data-driven selection of λ, we end up with a parameter being more adaptive, with respect to λ_{BIC} and λ_{AIC}, both to the sample size and to the features of the observed data.

Once the density (3.1) is estimated, a partition is operationally obtained by identifying its modal regions. To this aim, the most natural choice in the considered parametric framework, is the Modal EM algorithm (Li *et al.*, 2007) briefly introduced in Section 1.3.2.

3.3 Discussion

In this section we discuss further the procedure introduced in Section 3.2 pointing out some practical considerations and highlighting its properties and some links with other existing methods.

Remark 3.1. In Section 3.2 the dimension of the ensemble M has been considered as fixed. Nonetheless its selection is needed in order to practically resort to the estimator (3.1) and it could have some impact on the resulting partitions. Finding substantial arguments that motivate some general recommendations for choosing M is challenging and cannot leave aside the specificities of the data and of the problem at hand.
A natural strategy would consists in considering all the estimated models being a set of reasonable candidates selected by some prior knowledge, as a wide batch of alternatives recording a general uncertainty. Another alternative, being the one we followed in the empirical section of the chapter, may consist in choosing M subjectively by picking those models, among the estimated ones, resulting in a good fitting of the data. In this case M should vary also reflecting the case specific uncertainty witnessed in the modelling process. Lastly a viable approach we explore consists in considering an *Occam's window* to choose a set of model as proposed by Madigan and Raftery (1994). The main idea is to discard those models providing estimates being qualitatively too far from the ones provided by the best model. Practically the *ith* model can be discarded if $|\text{BIC}_{best} - \text{BIC}_i| > 10$, where BIC_{best} and BIC_i represent respectively the values of the BIC for the best model and for the *ith* one.

Remark 3.2. The estimation procedure outlined in Section 3.2.2 is fully frequentist in nature. Alternatively, a Bayesian approach could be an interesting development claiming some advantages. The work by Malsiner-Walli *et al.* (2017) faces, from a Bayesian perspective, the estimation of mixtures of mixture models. Even if the underlying motivation is different some ideas could be fruitfully borrowed and exploited in order to average different mixture models. As an example, the consideration of a shrinkage prior on the weights of the models in the ensemble could practically overcome the previously discussed issue of selecting M.

Remark 3.3. When considering the number of components as an unknown parameter, mixture models can be seen as a semi-parametric compromise between classical parametric model and non-parametric methods as, for example, kernel density estimators where the number of components equals the sample size. The model we introduced has an increased number of components inherited by the averaging procedure hence it takes another step forwards the non-parametric approach to density estimation. This partially motivates the way we identify the ensemble partitions by searching for the domains of attraction of the density modes. We believe indeed that, being model-based and modal clustering two sides of the density-based coin, our proposal finds a relevant strength in the coherency to not resort to distance-based approaches to practically identify a grouping of the data. Moreover, staying in the density-based clustering realm, it enjoys some of the relevant properties as for example the mathematically sound formalization and the concept of ideal population clustering.

Remark 3.4. Model selection often precedes inference that is usually conducted considering the chosen model as fixed. However, since the selection is itself data-dependent, it possesses its own variability. Drawing inference without accounting for the selection of the model corresponds to neglect completely a source of uncertainty usually resulting in anti-conservative statements (Leeb and Pötscher, 2005). Even in the full awareness of the fact that, in parametric clustering, the main focus usually lies on obtaining partitions rather than on inference or uncertainty quantification, we believe that a model averaging approach can entail better estimation properties and more informative confidence intervals for the parameters when needed.

Remark 3.5. In the supervised framework ensemble approaches have been found tremendously effective in improving predictions of a plethora of different models. For those techniques it has been frequently noticed (see, e.g. Dietterich, 2000) how the concept of *diversity* is a key factor in increasing classification performances of the *base learners* that are combined. As a consequence, often weak learners are considered in the supervised context. These classifiers are highly unstable, consequently different one from

the others, as they possibly focus on distinct features of the observed data. Even in a clustering framework the impact of the diversity among the combined partitions has been empirically studied and proved to be impactful by Fern and Brodley (2003) and Kuncheva and Hadjitodorov (2004).
We are aware that, when the proposed method is used to go beyond the *single best model* paradigm, the models in the ensemble cannot be considered as weak and consequently diversity among them is not achieved. Nonetheless, even if introduced with a specific aim, the proposal can in principle be exploited in all those cases where averaging multiple density-induced clusterings could be fruitful. As a consequence the diversity can be somehow determined for example averaging densities computed on bootstrap samples or on general subsamples of the observed data. Another possibly appealing application consists in combining models estimated using different starting values hence probably more heterogeneous because of the well known initialization issues encountered in the model-based clustering framework.

Remark 3.6. The model introduced so far, despite being based on a different rationale, shares some connections with the general framework of *Deep Gaussian Mixture Models* investigated by Viroli and McLachlan (2019). Deep Gaussian Mixture Models are networks of multiple layers of latent variables distributed as a mixture of Gaussian densities. Since the outlined representation encompasses the specification of a mixture of mixtures (Li, 2005), model (3.1) can be seen as a two layers Deep Gaussian Mixture Model where the parameters involved in the inner layer are fixed.

3.4 Results

3.4.1 Syntethic data

The idea of averaging together different densities to obtain a more informative summary for clustering purposes is explored in this section via simulations. All the reported analyses have been conducted in the R environment (R Core Team, 2019) with the aid of the `mclust` (Scrucca *et al.*, 2016), `ks` (Duong, 2019) and `EMMIXskew` (Wang *et al.*, 2018) packages. The code implementing the proposed procedure is meant to be made publicly available.

A total of $B = 200$ samples have been drawn, with sizes $n \in \{500, 5000\}$, for each of the bivariate densities depicted in Figure 3.2 and whose parameters are reported in the Appendix. These densities have been considered in order to encompass different situations posing different challenges from a model-based clustering perspective. The

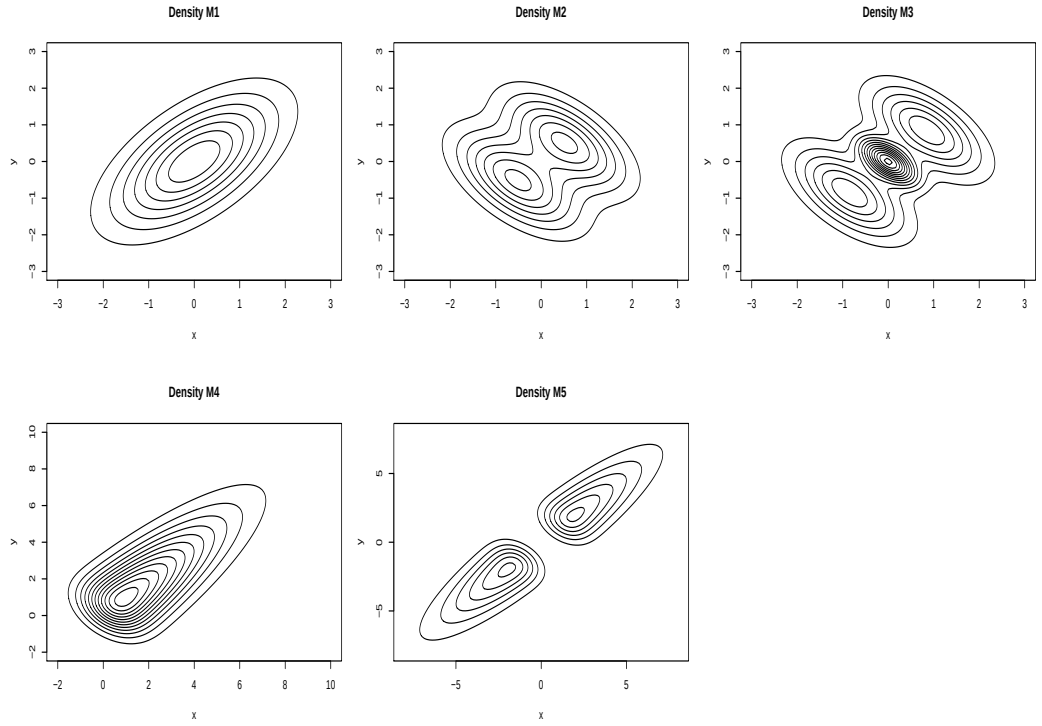


FIGURE 3.2: Bivariate density functions selected for simulations.

densities on the top panels of Figure 3.2 represent indeed settings where the single best model is expected to display satisfactory results, being the data generated from Gaussian mixtures. On the other hand the densities on the bottom panels, showing strong asymmetric behaviors, constitute more challenging settings where Gaussian mixture models generally produce inadequate partitions.

Throughout the simulations we have considered $M = 30$ best models ranked according to their BIC values, coherently with Remark 3.1 in Section 3.3; this choice moves towards the direction of retaining a large number of models, letting the estimation procedure to select the most relevant ones, while keeping the computations feasible. We also explored the option of selecting M by the *Occam's window* to build the ensemble as discussed in Remark 3.1; nonetheless results, not reported here, indicate that this strategy often leads to the selection of a small set of models implying again a strong reliance on the BIC. The three options λ_{AIC}, λ_{BIC} and λ_{CV} discussed in Section 3.2.2 are evaluated, the last one resorting to a *k-fold cross validation* scheme with $k = 5$.

The simulation study has multiple goals. On one side we want to evaluate the performances of our proposal in terms of the quality of the produced density estimates. These performances are studied with respect to the true and known density function considering the MISE as evaluating criterion. On the other hand the clustering performances of the proposed method are investigated. As an assessment criterion we employ the

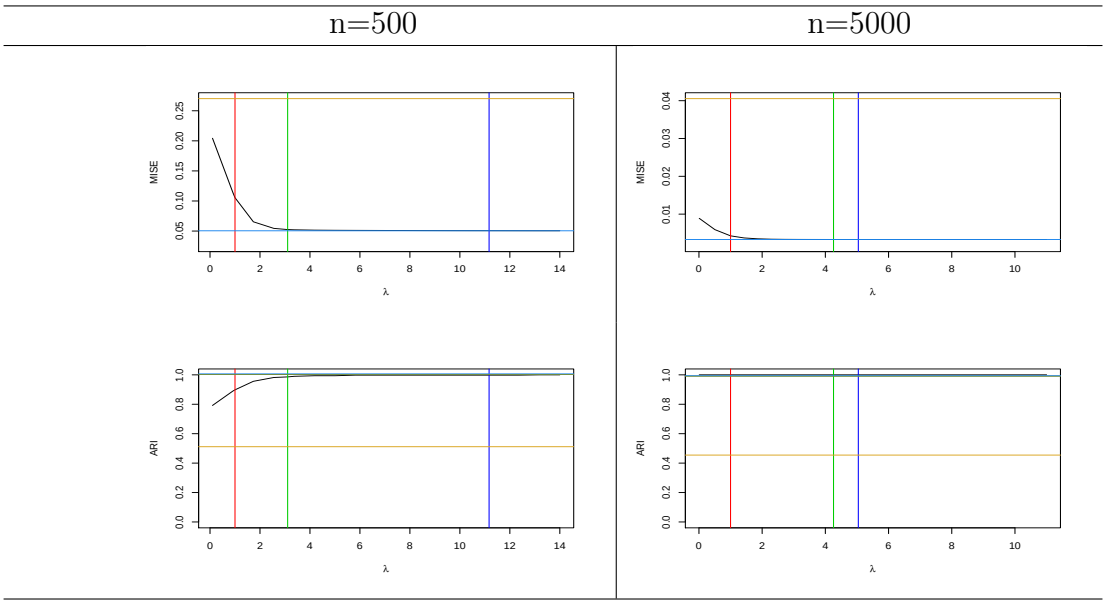
Adjusted Rand Index (ARI, Hubert and Arabie, 1985) between the obtained partitions and the true component memberships of the observations. An additional aim consists in evaluating how the sample size impacts on these comparisons.

As a side goal of the numerical explorations we want to study which penalization strategy introduced in Section 3.2.2 produces more satisfactory results. In particular we evaluate if the increased computational costs implied by the cross-validation worth the effort or if less intensive strategies as the *BIC-type* and *AIC-type* penalizations produce comparable results. Lastly we want to compare our proposals with some reasonable competitors. We consider a fully parametric approach, using the single best model chosen among the alternatives in Table 1.1. Moreover we consider a completely non-parametric counterpart relying on the kernel density estimator and on the *mean-shift algorithm* to obtain the partition as discussed in Section 1.3.1 where we use, as a bandwidth matrix, the unconstrained gradient one as it constitutes a standard choice (see Chacón and Duong, 2018, for a detailed tractation). Furthermore we examine also an hybrid approach consisting in finding the modes, via Modal EM algorithm, of the density estimated by the single best model. The possible improvements introduced by our proposal may be due to two different motivations: the first related to a better estimation of the underlying density while the second is concerned with the modal-inspired allocation procedure. Considering an hybrid approach as a competitor can help to disentangle properly these distinct sources.

Results are reported in Tables 3.1 to 3.5. As a first, expected, behavior the performances of the methods considered tend to improve, both from a clustering and from a density estimation point of view, as the sample size increases.

Generally speaking our proposal, regardless of the penalization used, produces satisfactory density estimates and partitions of the datasets. The first three scenarios have been considered to see how the ensemble approach behaves in situations where the *single best model* has a head start; in these cases the true generative model is indeed among the ones estimated in the model-based clustering routine. Even in these somewhat unfavourable settings, where in some sense an ensemble approach is not strictly needed, the proposed method behaves well producing overall comparable results with respect to the parametric ones.

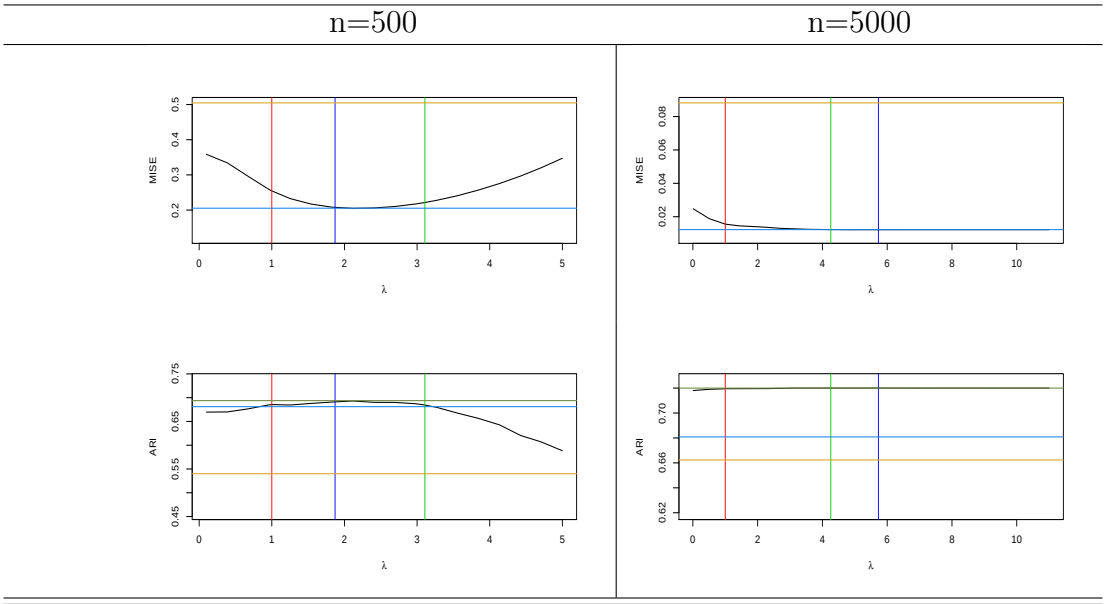
In the skewed scenarios M4 and M5, where Gaussian mixture models are known to be less effective as a clustering tool, the ensemble approach induces remarkable improvements in the performances, both in terms of MISE and ARI. Note that, regarding the relation between performances and sample size, we are witnessing some results constituting an exception with respect to what we pointed out before. Indeed, especially for



	n=500		n=5000	
	MISE	ARI	MISE	ARI
SB	0.050 (0.037)	1.000 (0.000)	0.003 (0.002)	1.000 (0.000)
NP	0.270 (0.091)	0.528 (0.500)	0.041 (0.012)	0.455 (0.499)
SB-NP	-	1.000 (0.000)	-	1.000 (0.000)
λ_{AIC}	0.105 (0.088)	0.898 (0.303)	0.004 (0.003)	1.000 (0.000)
λ_{BIC}	0.052 (0.038)	0.986 (0.118)	0.003 (0.002)	1.000 (0.000)
λ_{CV}	0.051 (0.037)	0.998 (0.045)	0.003 (0.002)	1.000 (0.000)

TABLE 3.1: Top panel: the MISE (up to a density-dependent multiplicative constant) and the ARI (black lines) as functions of λ for $n = 500, 5000$. Light blue, gold and dark green horizontal lines represent the same quantities respectively for the single best model (SB), the nonparametric approach (NP) and the hybrid approach (SB-NP). The vertical lines represent the values of λ_{AIC} (in red), λ_{BIC} (in light green) and the mean over the B samples of λ_{CV} (in blue). Bottom panel: numerical values of the MISE (up to a density-dependent multiplicative constant) and ARI (and their standard errors) for the competing considered methods. Results refer to density M1.

the setting M5, the increased availability of data points forces Gaussian mixture models to resort to an higher number of components, even if in the presence of two groups, to properly model the asymmetry thus deteriorating the clustering results. In commenting these results some words of caution are needed since obtaining the allocation according to the modal concept of groups can have a strong impact in these two settings. Nonetheless comparisons with the hybrid approach help shedding light on this and to study further the improvements intrinsically introduced by averaging together distinct densities. The method proposed, despite showing comparable results when $n = 500$, attains notable enhancements when $n = 5000$ along with decreased standard errors. This could constitute an indication about the increased quality, from a clustering standpoint, of the density estimates produced considering model 3.1 with respect to the ones



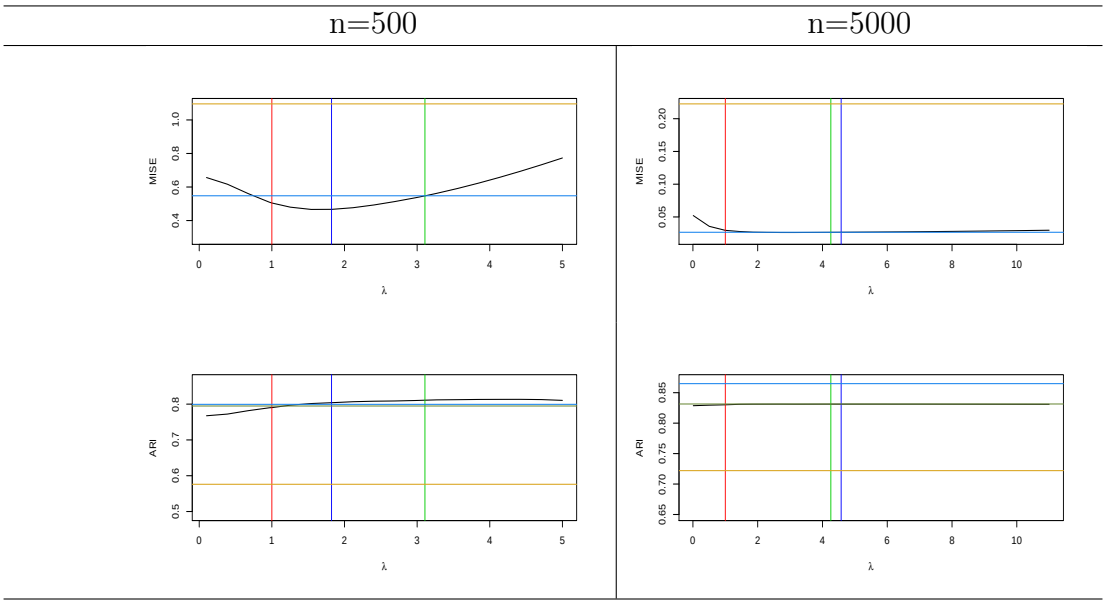
	MISE	ARI	MISE	ARI
SB	0.205 (0.223)	0.681 (0.119)	0.012 (0.008)	0.680 (0.067)
NP	0.505 (0.144)	0.540 (0.177)	0.089 (0.021)	0.661 (0.083)
SB-NP	-	0.694 (0.113)	-	0.720 (0.012)
λ_{AIC}	0.255 (0.146)	0.686 (0.076)	0.015 (0.010)	0.719 (0.013)
λ_{BIC}	0.221 (0.171)	0.685 (0.125)	0.012 (0.008)	0.720 (0.012)
λ_{CV}	0.208 (0.130)	0.691 (0.092)	0.012 (0.008)	0.720 (0.012)

TABLE 3.2: Cf. Table 3.1. Results refer to density M2

produced by a single mixture model; better ARI values could indeed indicate smoother estimates, being easier to be explored when searching for the modes.

The aforementioned decrease in the variability of the results of the proposal with respect to the competitors is witnessed across all the scenarios. This represents a substantial and somewhat expected advantage of the ensemble approach, since a gain in robustness and stability moves towards the desired direction when mixing models together.

With regard to the choice of the penalization scheme some different considerations arise. As expected, building on a data-based rationale, λ_{CV} seems to be more reliable when the aim is to obtain an accurate estimate of the density. Choosing the amount of the penalization via cross-validation appears to be particularly suitable especially when $n = 500$ while, with increasing sample size, the performances of the three considered schemes tend to be more similar. However, when clustering is the final aim of the analysis λ_{BIC} turns out to be a serious candidate as it often produces better results with respect to λ_{CV} and λ_{AIC}; this constitutes a notable result since the *BIC-type* penalization, unlike the cross-validation based one, requires a null computational cost when dealing



	n=500		n=5000	
	MISE	ARI	MISE	ARI
SB	0.548 (0.431)	0.798 (0.107)	0.027 (0.013)	0.865 (0.008)
NP	1.096 (0.226)	0.588 (0.152)	0.224 (0.041)	0.713 (0.097)
SB-NP	-	0.797 (0.099)	-	0.831 (0.012)
λ_{AIC}	0.505 (0.247)	0.790 (0.070)	0.029 (0.014)	0.829 (0.017)
λ_{BIC}	0.547 (0.330)	0.811 (0.065)	0.027 (0.013)	0.830 (0.012)
λ_{CV}	0.467 (0.254)	0.804 (0.063)	0.027 (0.013)	0.830 (0.012)

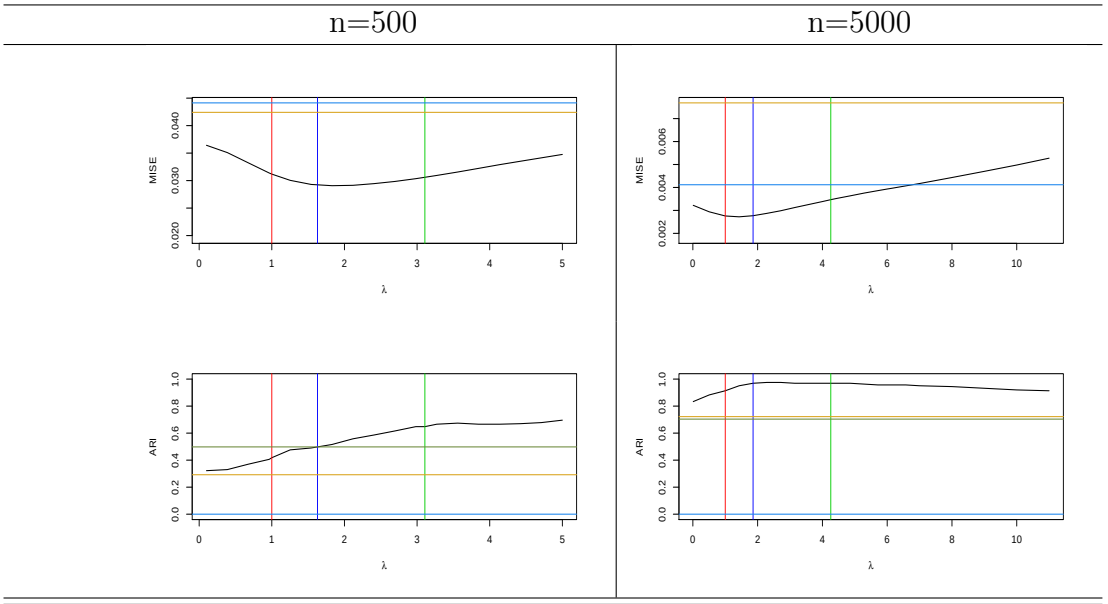
TABLE 3.3: Cf. Table 3.1. Results refer to density M3

with the selection of λ. On the other hand, as expected not even depending on the sample size, λ_{AIC} tends to produce the most unsatisfactory results among the three.

Lastly note that the performances of the fully nonparametric approach appear not to be competitive with the other approaches considered. Nonetheless we believe that some tuning in choosing the smoothing parameters used could lead to an improvement in the results. Anyway in our numerical explorations this chance is not explored since appropriate bandwidth selection is not the aim of the study hence it appears reasonable to resort to a standard selector as we did.

3.4.2 Real data

In this section we consider three illustrative examples on real datasets. As in the previous section, we fit our proposed model considering the three different penalization schemes introduced in Section 3.2.2 and we use as competitors the parametric, the nonparametric and the hybrid approaches. The number of models in the ensemble is set to $M = 30$ following the same rationale as the one discussed in the simulated examples. The focus



	MISE	ARI	MISE	ARI
SB	0.044 (0.018)	0.000 (0.000)	0.004 (0.002)	0.000 (0.000)
NP	0.042 (0.013)	0.292 (0.455)	0.008 (0.002)	0.736 (0.442)
SB-NP	-	0.498 (0.500)	-	0.727 (0.447)
λ_{AIC}	0.031 (0.015)	0.418 (0.494)	0.003 (0.001)	0.917 (0.277)
λ_{BIC}	0.031 (0.013)	0.648 (0.478)	0.003 (0.001)	0.968 (0.177)
λ_{CV}	0.029 (0.013)	0.498 (0.500)	0.003 (0.001)	0.967 (0.178)

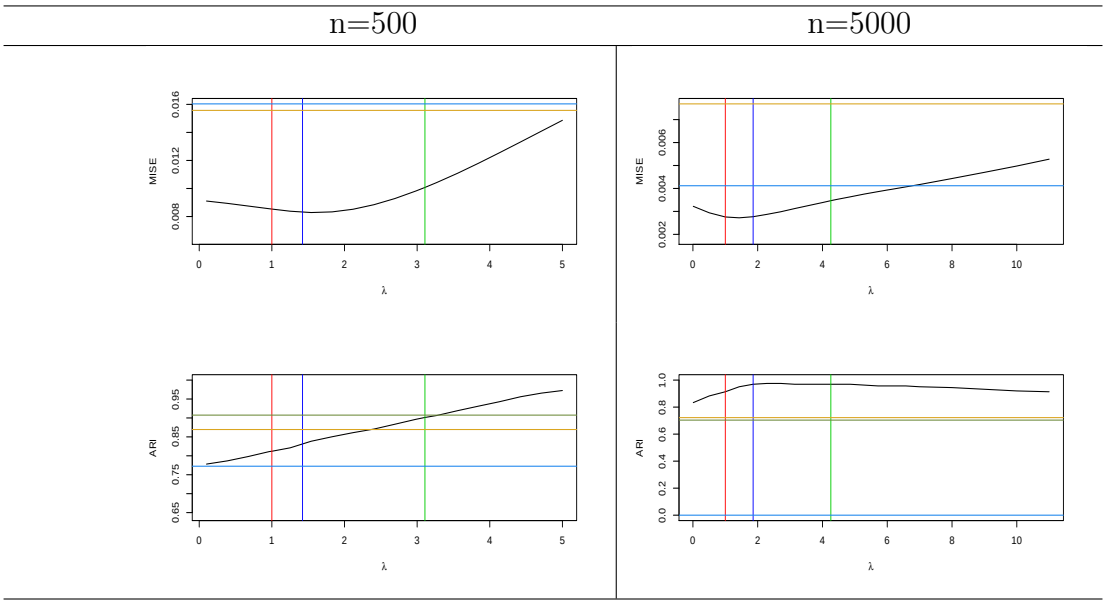
TABLE 3.4: Cf. Table 3.1. Results refer to density M4

of the analyses, not having a true density to refer at, is set on the quality of the obtained partitions, evaluated via *Adjusted Rand Index*.

3.4.2.1 Iris data

The *Iris* dataset (available at `https://archive.ics.uci.edu/ml/datasets/Iris`), already mentioned in Section 3.2.1 to motivate our proposal, have been thoroughly studied since the seminal paper by Fisher (1936) and it consists in $d = 4$ variables (sepal length and width, petal length and width) measured on $n = 150$ iris plants with $K_{true} = 3$ classes equally sized. A visual illustration of the data is given in Figure 3.3; note that one class is linearly separable from the other two, in turn hardly to detect as separate groups.

Results are shown in Table 3.6. The method proposed here clearly outperforms all the considered competitors. As seen in Section 3.2.1 the BIC select a two-component model hence giving wrong indications about the number of groups. As a consequence, both the parametric and the hybrid approaches, relying on the single best model, tend to produce unsatisfactory results. On the other hand the detection of 7 groups, via modal clustering



	MISE	ARI	MISE	ARI
SB	0.016 (0.006)	0.773 (0.190)	0.001(0.001)	0.367 (0.033)
NP	0.016 (0.003)	0.869 (0.144)	0.003 (< 0.001)	0.868 (0.159)
SB-NP	-	0.907 (0.160)	-	0.839 (0.175)
λ_{AIC}	0.009 (0.003)	0.812 (0.167)	0.001 (< 0.001)	0.960 (0.091)
λ_{BIC}	0.010 (0.003)	0.901 (0.152)	0.001 (< 0.001)	0.986 (0.052)
λ_{CV}	0.008 (0.003)	0.831 (0.164)	0.001 (< 0.001)	0.966 (0.086)

TABLE 3.5: Cf. Table 3.1. Results refer to density M5

	SB	NP	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.568	0.556	0.568	0.845	0.941	0.869
$\hat{K}$	2	7	2	4	3	4

TABLE 3.6: Results obtained on the Iris dataset. The true number of cluster is $K_{true} = 3$.

based on kernel density estimation, is a symptom of an undersmoothed density estimate. Note that the high degree of rounding in the dataset could affect nonparametric performances since the estimator is built to work with continuous data, hence without duplicated values. Our method, regardless of the penalization scheme, produces strong improvements in the clustering results. The *AIC-type* and the cross-validation-based penalizations wrongly find 4 clusters with one spurious, yet small, group detected. A deeper examination of the results reveals that conversely, λ_{BIC} assumes a grossly doubled value with respect to λ_{AIC} and λ_{CV} and allows for the correct identification of 3 groups.

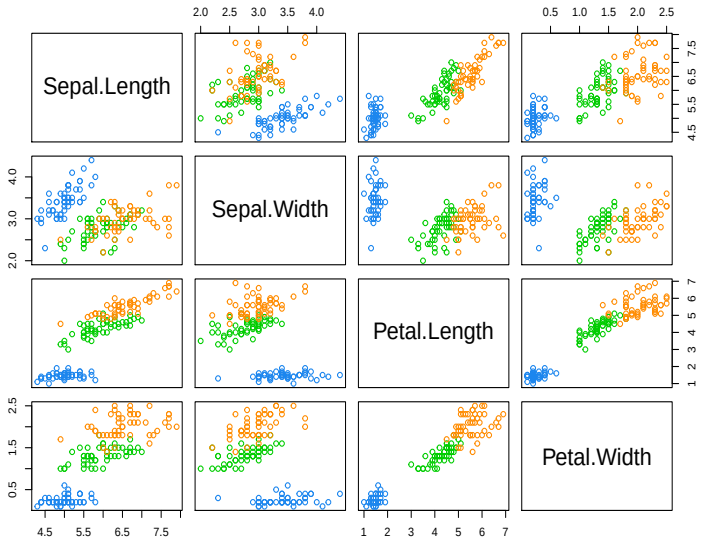


FIGURE 3.3: Bivariate scatter plots of the Iris data with colors representing the true clustering labels.

3.4.2.2 DLBCL data

The *Diffuse Large B-cell Lymphoma* (DLBCL) dataset is provided by the British Columbia Cancer Agency (Spidlen *et al.*, 2012; Aghaeepour *et al.*, 2013). The sample consists in fluorescent intensities of $d = 3$ markers, namely CD3, CD5 and CD19, measured on $n = 8183$ lymph nodes cells from subjects with a DLBCL diagnosis. A scatter plot of the data is shown in Figure 3.4. In flow cytometry analysis these measurements are used to study normal and abnormal cell structures and to monitor human diseases and response to therapies. An essential step in this framework consists in obtaining a grouping of the cells according to their fluorescences. This task is usually accomplished via the so called *gating* process: the experts obtain a partition manually by visually inspecting the data. This approach is usually time-consuming and infeasible in high-dimensional situations, therefore clustering tools could come in aid to automate the gating process. The 3-dimensional structure of the data, illustrated in Figure 3.4, allows us to visually inspect the true cluster configuration, displaying elongated and skewed group shapes. Results in Table 3.7 show how the model-based approach, as noted in the simulated scenarios, tends to perform badly when dealing with such situations, since it detects an higher number of groups with respect to the true one. In this setting, building mixtures on more flexible, possibly skew component densities could help in improving the fit by means of a single model. Conversely, the nonparametric and the hybrid approaches, which search for the modes of the density, do not suffer of the same drawbacks and outperform the parametric strategy. Nonetheless while the former appears to undersmooth

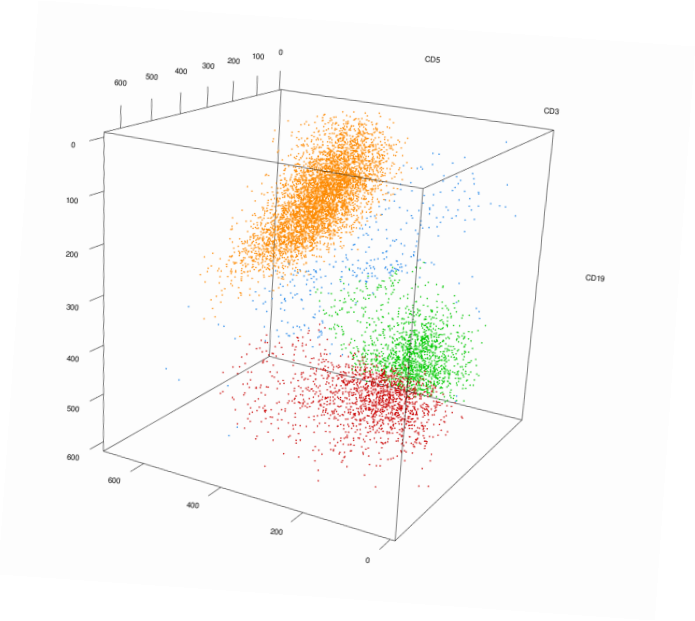


FIGURE 3.4: 3D scatter plot of the DLBCL data with colors representing the true clustering labels.

	SB	NP	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.401	0.857	0.867	0.909	0.910	0.912
$\hat{K}$	7	5	4	4	4	4

TABLE 3.7: Results obtained on the DLBCL dataset. The true number of cluster is $K_{true} = 4$.

again the density, the latter detects the true number of clusters, yet with improvable performance in the allocation of units.

Our proposal, regardless of the penalization scheme adopted, enjoys the very same advantage of nonparametric tools when dealing with asymmetric shapes. Nonetheless the results obtained improve with respect to the hybrid approach thus indicating that our model produces a density estimate better tailored for the clustering scope. In this case different penalization schemes lead to irrelevant changes in the ARI values, and indicate a weaker dependency on the strength of the penalization itself.

3.4.2.3 Olive oil data

As a last example we consider the *Olive oil* dataset, originally introduced in Forina *et al.* (1986). The data consist of $d = 8$ chemical measurements on $n = 572$ olive oils produced in 9 regions of Italy (North and South Apulia, Calabria, Sicily, Sardinia coast and inland, Umbria, East and West Liguria) that can be further aggregated in three macro-areas (Centre-North, South and Sardinia island). Clustering tools may come in

	SB	NP	SB-NP	λ_{AIC}	λ_{BIC}	λ_{CV}
ARI	0.782	0.604	0.792	0.902	0.892	0.902
$\hat{K}$	6	20	6	8	8	8

TABLE 3.8: Results obtained on the Olive oil dataset. The unaggregated regions have been considered as true labels hence $K_{true} = 9$.

		1	2	3	4	5	6	7	8
South	Apulia north	24	1	0	0	0	0	0	0
	Apulia south	0	6	200	0	0	0	0	0
	Calabria	0	56	0	0	0	0	0	0
	Sicily	6	30	0	0	0	0	0	0
Sardinia	Sardinia inland	0	0	0	65	0	0	0	0
	Sardinia coast	0	0	0	0	33	0	0	0
Centre-North	Liguria east	0	0	0	0	0	1	42	7
	Liguria west	0	0	0	0	0	0	0	50
	Umbria	0	0	0	0	0	48	3	0

TABLE 3.9: Olive oil results, partition obtained with penalization parameter λ_{AIC}

aid in reconstructing the geographical origin of the oils on the basis of their chemical compositions.

This example allows us to explore the performances of the proposal in a moderately higher dimensional setting with respect to the two considered above. Results in Table 3.8 show how our proposal outperforms the competitors, regardless of the penalization adopted, allowing to obtain a more faithful partition of the data into the 9 considered regions. The parametric and the hybrid approaches detect 6 groups, aggregrating Sardinia coast and inland oils and highlighting some issues concerning the correct classification of oils produced in South macro-area. On the other hand, probably suffering of the higher dimensionality of the data, the fully nonparametric approach clearly produces a partition based on an severely undersmoothed density with 20 modes.

As it happened in Section 3.4.2.2 the clustering performances of our proposal appear to be quite insensitive to the specific penalization adopted. In Table 3.9 we report the partition induced considering λ_{AIC} as penalizing parameter. Again it appears harder to discriminate the oils produced in the southern macro-area, with calabrian and sicilian ones assigned mainly to the same cluster, while oils in the other two macro-areas are substantially correctly identified.

3.5 Conclusions

In this chapter we have addressed the issue of overcoming the strong reliance of model-based clustering on a single best model, selected according to some information criterion. Making reference to a single model may be suboptimal both for clustering and for density estimation, since alternative well-fitted models may provide useful information by uncovering different and complementary features which are otherwise discarded. It has been pointed out that possible solutions may be found in the ensemble learning literature. In this setting, we have proposed a clustering method building on a density function which averages different estimated models, and whose modal regions are then associated to the groups. The introduced density estimator is defined as a convex linear combination of the estimates of the models in the ensemble, with weights estimated via penalized maximum likelihood. This choice allows assigning relevance to the only models which better fit the data while avoiding the risk of overfitting.

The introduced approach can be comprehensively viewed as an attempt to bind together the parametric and the nonparametric formulations of density-based clustering, thus inherit their intrinsic strengths. From one side, the modal concept of clusters is considered, which allows to identify groups of arbitrary shape which naturally comply with the geometric intuition. From the other side, by resorting to parametric tools and to model average, density estimation is strengthened, allowing to obtain more accurate results of both nonparametric tools and single parametric models. The performances of the proposal have been investigated both on simulated and on real data, selected to encompass different situations and to pose distinct challenges. The method produces satisfactory results both from a density estimation and from a clustering perspective, and it compares favorably with the considered competitors. A deeper examination of the results leads to disentangle the reasons of the improvements into two different sources: on one side partitioning the data according to the modal formulation produces promising results in some specific scenarios, on the other hand several clues have been obtained which highlight enhancements in the density estimation process. Concerning the introduced penalization schemes, the results seem to suggest the use of the *BIC-type* penalization, being more suitable for clustering, or of the cross-validation-based one, being able to adapt more to the features of the considered dataset.

Chapter 4

Co-clustering of time-dependent data

4.1 Introduction

Time dependent data, arising when measurements are taken on a set of units in different time occasions, are pervasive in a plethora of different fields. Non exhaustive examples are data describing the time evolution of asset prices and volatility in finance, the growth of countries as measured by economic indices, heart or brain activities as monitored by medical instruments, disease evolution evaluated by suitable bio-markers in epidemiology, data streams on websites or electronic devices. The analysis of such data shares a common aim of proper modelling typical time courses by accounting for the individual correlation over time. In fact, while nomenclature and taxonomy in this setting are not always consistent, some relevant differences in time-dependent data structures and the subsequent different challenges in the modelling process can be highlighted. On opposite poles we may thus distinguish functional from longitudinal data analysis. In the former case the quantity of interest is supposed to vary over a continuum and usually a huge number of regularly sampled observations is available, allowing to treat each sample element as a function. On the other hand, in longitudinal studies, time series are often shorter with sparse and irregular measurements. See Rice (2004) for further details.

Very often standard methods take into account properly these features while, at the same time, assume homogeneity among individuals as if the observed curves were generated by the same mechanism. Nonetheless this is often not the case, and tools being able to identify and describe the heterogeneity across curves are necessary. In the outlined landscape, clustering methods may be particularly useful to capture heterogeneous behaviors by assuming that some groups, each of them characterized by its

own generative model, are present. To this aim, several tools, addressing the above-mentioned criticalities and aiming at finding groups in a set of observed curves in time, have been proposed. For a thorough review of these works the reader may refer to Liao (2005) and Frühwirth-Schnatter (2011). Some contributions that specifically worth a mention, developed in a model-based framework coherently with this thesis, are the ones of De la Cruz-Mesía *et al.* (2008), McNicholas and Murphy (2010), Bouveyron and Jacques (2011) and Bouveyron *et al.* (2015). While the first two works specifically deal with longitudinal data, the latter two address, with a clustering aim in mind, the issues arising in a functional analysis setting.

All these methods deal with situations where a single feature is measured over time for a number of subjects. Data of such type have a two-way structure and may be represented by a $n \times T$ matrix, being n and T respectively the number of subjects and of observed time instants. In fact, nowadays it is increasingly common to encounter multivariate time-dependent data, where several variables are measured over time for different individuals. These data may be represented according to three-way structured matrices of dimension $n \times d \times T$ where d here is the number of features. The introduction of an additional layer entails some new challenges and criticalities that have to be faced and taken into account by clustering, and more generally, modelling tools. Research in this framework has been conducted in a considerably more scattered way. Indeed, as highlighted by Anderlucci and Viroli (2015), models have to "account simultaneously for three goals of the analysis, which arise from the three layers of the data structure: heterogeneous units, correlated occasions and dependent variables".

To extract useful information and unveil patterns from such complex structured and high-dimensional data, standard clustering strategies would require the specification and the estimation of severely parametrized models. To induce parsimony, such situation has often lead to neglect the correlation structure among different variables. A possible clever workaround, specifically proposed in a parametric setting, is represented by the contributions of Viroli (2011a,b) where, in order to handle three-way data, mixtures of Gaussian matrix-variate distributions are exploited.

In this work a different direction has been taken, and a co-clustering strategy is pursued to address the mentioned issues. Aiming at simultaneously cluster rows and columns of the observed data matrix, co-clustering models turn out to be particularly well suited in the presence of heterogeneous high-dimensional data where also relations among the variables are of interest. In this setting, we propose a parametric model conceived for time-dependent data and we introduce a new estimation strategy being able to handle the peculiar characteristics of the model.

The rest of the chapter is organized as follows. In Section 4.2 we introduce the main ingredients considered for the specification of the proposed method. This is in turn described, along with the estimation procedure, in Section 4.3. In Section 4.4 the performances of the methodology are illustrated both on simulated and real examples. Lastly some concluding remarks are outlined in Section 4.5.

4.2 Building blocks

4.2.1 Modelling time-dependent data

In the quest for a flexible approach to handle the heterogeneous landscape of time dependent data outlined in the previous section, a variety of modelling approaches are sensible to be pursued. The one we follow in this thesis borrows the rationale from *curve registration* (Ramsay and Li, 1998), according to which observed curves often exhibit common patterns but with some variations. Methods for curve registration, also known as *curve alignment* or *time warping*, are based on the ideas of aligning prominent features in a set of curves via either an *amplitude variation*, a *phase variation* or a combination of the two via scale transformation. The first one is concerned with vertical variations while the latter regards horizontal, hence time related, ones. As an example it is possible to think about modelling the evolution of a specific disease. Here the observable heterogeneity of the raw curves can be often disentangled in two different sources: on one hand it should indeed depend on differences in the intensities of the disease among subjects, on the other hand there could be different ages of onset, i.e. the age at which an individual experiences the first symptoms. Therefore, after having properly taken into account of these causes of variation, often the curves result to be more homogeneously behaving, with a so called *warping function*, which synchronizes the observed curves and allows for visualization and estimation of a common mean shape curve.

Coherently with the aforementioned rationale, in this work time dependency is accounted for via a *self-modelling regression* approach (Lawton *et al.*, 1972) and, more specifically, via the so called *Shape Invariant Model* (SIM, Lindstrom, 1995), based on the idea that an individual curve is a simple transformation of a common shape function. Let be $\mathbb{X} = \{x_i(\mathbf{t_i})\}_{1 \leq i \leq n}$ the set of curves, observed on n individuals, with $x_i(t)$ representing the level of the *i-th* curve at time t and $t \in \mathbf{t_i} = (t_1, \ldots, T_{n_i})$, hence with the number of observed measurements allowed to be subject-specific. Stemming from the

Shape Invariant Model we define

$$x_i(t) = \alpha_{i,1} + \mathrm{e}^{\alpha_{i,2}} m(t - \alpha_{i,3}) + \epsilon_i(t) \tag{4.1}$$

where

- $m(\cdot)$ denotes a general common shape function whose specification is arbitrary. In the following we consider B-spline basis functions (De Boor, 1978), i.e. giving $m(t) = m(t; \beta) = \mathcal{B}(t)\beta$, where $\mathcal{B}(t)$ and β are respectively a vector of B-spline basis evaluated at time t and a vector of basis coefficients whose dimension allows for different degrees of flexibility;

- $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}) \sim \mathcal{N}_3(\mu^\alpha, \Sigma^\alpha)$ $i = 1, \ldots, n$ is a vector of subject specific normally distributed random effects. These random effects are responsible for the individual specific transformations of the mean shape curve $m(\cdot)$ assumed to give birth to the observed ones. In particular $\alpha_{i,1}$ and $\alpha_{i,3}$ govern respectively amplitude and phase variations while $\alpha_{i,2}$ accounts for scale transformations. Moreover they allow taking into account the correlation among observations on the same subject measured at different time points. Note that, following Lindstrom (1995), the parameter $\alpha_{i,2}$ is optimized in the log-scale to avoid identifiability issues;

- $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2_\epsilon)$ is a Gaussian distributed error term.

Due to its flexibility, and even if with different purposes, the SIM has already been considered as a stepping stone to model different types of time-dependent data as functional and longitudinal one (Telesca and Inoue, 2008; Telesca *et al.*, 2012). Indeed, if on one hand the smoothing involved in the specification of $m(\cdot; \beta)$ allows to handle function-like data, on the other hand random effects, borrowing information across curves, make this approach fruitful even with short, irregular and sparsely sampled time series. Hence we find such model particularly appealing and suitable for our scopes, being potentially able to handle temporal dependent data in a quite comprehensive way.

4.2.2 Latent Block Model

Even in a co-clustering framework a taxonomy of the approaches proposed in literature, coherent with the *distance-based* versus *density-based* dualism in the clustering framework, is possible. With regard to the first class of methods, referred to as *metric approaches* in Govaert and Nadif (2013), unsurprisingly the task boils down to the selection of an appropriate distance measure to be minimized among the original matrix

and a block-structured one. Conversely, the density-based approach aims at embedding co-clustering in a probabilistic framework. It reflects the idea of a density being partitionable in several blocks, and builds a common framework to handle different type of data. In this work, coherently with the rest of the thesis, we pursue the latter approach.

In the model-based co-clustering framework the *Latent Block Model* (LBM, Govaert and Nadif, 2013) represents unarguably the most popular approach. Consider a data set represented in a matrix form $\mathbb{X} = \{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq d}$, where by now we should intend x_{ij} as a random variable of generic nature. To aid the definition of the model, and in accordance with the parametric approach to clustering, two latent random vectors $\mathbf{z} = \{z_i\}_{1\leq i\leq n}$, with $z_i = (z_{i1}, \ldots, z_{iK})$, and $\mathbf{w} = \{w_j\}_{1\leq j\leq d}$, with $w_j = (w_{j1}, \ldots, w_{jL})$, are introduced, indicating respectively the row and column cluster memberships, with K and L the number of row and column clusters. Here, the standard binary partition holds for the latent variables; hence $z_{ik} = 1$ if the *i-th* observation belongs to the *k-th* row cluster and 0 otherwise and, coherently, $w_{jl} = 1$ if the *j-th* variable belongs to the *l-th* column cluster and 0 otherwise. The model formulation relies on a local independence assumption, i.e. the $n \times d$ random variables $\{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq d}$ are assumed to be independent conditionally on $\mathbf{z}$ and $\mathbf{w}$. Moreover $\mathbf{z}$ and $\mathbf{w}$ are in turn considered as independent. The LBM can be thus written as

$$p(\mathbb{X}; \Theta) = \sum_{z\in Z} \sum_{w\in W} p(\mathbf{z}; \Theta) p(\mathbf{w}; \Theta) p(\mathbb{X}|\mathbf{z}, \mathbf{w}; \Theta) \ , \tag{4.2}$$

where:

- Z and W are respectively the set of all the possible partitions of rows in K groups and columns in L groups;
- a multinomial distribution is assumed for both the latent vectors $\mathbf{z}$ and $\mathbf{w}$. Therefore $p(\mathbf{z}; \Theta) = \prod_{ik} \pi_k^{z_{ik}}$ and $p(\mathbf{w}; \Theta) = \prod_{jl} \rho_l^{w_{jl}}$ where π_k and ρ_l are the row and column mixture proportions, hence lying in $[0, 1]$ with $\sum_k \pi_k = \sum_l \rho_l = 1$;
- as a consequence of the local independence assumption, we may decompose $p(\mathbb{X}|\mathbf{z}, \mathbf{w}; \Theta)$ such as $p(\mathbb{X}|\mathbf{z}, \mathbf{w}; \Theta) = \prod_{ijkl} p(x_{ij}; \theta_{kl})^{z_{ik} w_{jl}}$ where θ_{kl} is the vector of parameters specific to block (k, l);
- $\Theta = (\pi_k, \rho_l, \theta_{kl})_{1\leq k\leq K, 1\leq l\leq L}$ is the full parameter vector of the model.

It is straightforward to note, from the formulation outlined in (4.2), how the introduction of an additional latent variable basically adds a supplementary mixture layer to the model (1.1) used as a cornerstone in parametric clustering. For a more detailed tractation of the link among the LBM and mixture models the reader may refer to Govaert

and Nadif (2013). The authors indeed highlight how, conditionally on the partition $\mathbf{w}$, the density function of $\mathbb{X}$ is a mixture model, and the same holds conditioning on $\mathbf{z}$.

The LBM turns out to be particularly flexible in modelling different data types as they are handled by a proper specification of the marginal density $p(x_{ij}; \theta_{kl})$. As a consequence several versions of the LBM may be found in literature dealing with binary (Govaert and Nadif, 2003), count (Govaert and Nadif, 2010), continuous (Lomet, 2012), categorical (Keribin *et al.*, 2015), and ordinal data (Jacques and Biernacki, 2018; Corneli *et al.*, 2019). At the best of our knowledge the only works proposing a parametric co-clustering approach to model time-dependent data is represented by the ones of Ben Slimen *et al.* (2018) and Bouveyron *et al.* (2018) where the LBM is extended to a functional setting, where the block partitions result from clustering the basis expansion coefficients and not directly the observed curves.

Several estimation approaches have been proposed for LBM as for example likelihood-based (Govaert and Nadif, 2008), Bayesian (Wyse and Friel, 2012) and greedy search methods (Wyse *et al.*, 2017). Coherently with the framework introduced in Section 1.2.2, we focus on methods based on likelihood maximization, and exploit the double missing data structure of the problem, induced by the unknown row and column labels described by the latent variables $\mathbf{z}$ and $\mathbf{w}$.

As in the clustering setting, the aim consists in maximizing the *complete data log-likelihood*, defined as

$$\ell_c(\Theta, \mathbf{z}, \mathbf{w}) = \sum_{ik} z_{ik} \log \pi_k + \sum_{jl} w_{jl} \log \rho_l + \sum_{ijkl} z_{ik} w_{jl} \log p(x_{ij}; \theta_{kl}) \tag{4.3}$$

where the first two terms account for the proportions of row and column clusters while the third one depends on the probability density function of each block.
A sensible approach to maximize (4.3) would resort to the EM-algorithm. Unfortunately in the co-clustering case this approach is unfeasible as the E-step would require the computation of the joint conditional distribution of the missing labels which involves terms that cannot be factorized as conversely happens in a standard mixture model framework. As a consequence, several modifications have been explored, searching for a workaround when dealing with the E-step; examples are the *Stochastic EM-Gibbs* (SEM) algorithm, the *Classification EM* (CEM) and the *Variational EM* (VEM) among the others. In the subsequent developments, we shall focus on the SEM algorithm, where a SE-step takes the place of the E-step by replacing the intractable computation of the expected value of $(\mathbf{z}, \mathbf{w})$ by simulating according to their conditional distribution via Gibbs sampling.

4.3 Time-dependent Latent Block Model

4.3.1 Model specification

Once the LBM structure has been properly defined, extending its rationale to handle time-dependent data in a co-clustering framework boils down to a suitable specification of $p(x_{ij};\theta_{kl})$. Note that this reveals one of the main advantage of such an highly-structured model, consisting in the chance to search for patterns in multivariate and complex data by specifying only the model for the variable x_{ij}.

As introduced in Section 4.1, multidimensional time-dependent data may be represented according to a three-way structure where the third *mode* accounts for the time evolution. The observed data assume an array configuration as $\mathbb{X} = \{x_{ij}(\mathbf{t_i})\}_{1\leq i\leq n, 1\leq j\leq d}$ with $\mathbf{t_i} = (t_1, \ldots, T_{n_i})$ as outlined in Section 4.2.1; different observational lengths are taken can be handled by a suitable use of missing entries. Consistently with the (4.1), we consider as a generative model for the curve in the (i,j)-*th* entry, belonging to the generic block (k,l)), the following

$$(x_{ij}(t)|z_{ik}=1, w_{jl}=1) = \alpha^{kl}_{ij,1} + \mathrm{e}^{\alpha^{kl}_{ij,2}} m(t-\alpha^{kl}_{ij,3};\beta_{kl}) + \epsilon_{ij}(t)\ . \tag{4.4}$$

Two main differences may be highlighted with respect to the original SIM model. Here we do not have individual-specific random effects but cell-specific ones since we are indeed modelling directly the (i,j)-*th* cell. Moreover, reasoning conditionally to the block membership of the cell, the parameters involved are block-specific, coherently with the co-clustering setting. As a consequence:

- $m(t;\beta_{kl}) = \mathcal{B}(t)\beta_{kl}$ where the quantities are defined as in Section 4.2.1, with the only difference that β_{kl} is a vector of block-specific basis coefficients, hence allowing different mean shape curves across different blocks;
- $\alpha^{kl}_{ij} = (\alpha^{kl}_{ij,1}, \alpha^{kl}_{ij,2}, \alpha^{kl}_{ij,3}) \sim \mathcal{N}_3(\mu^{\alpha}_{kl}, \Sigma^{\alpha}_{kl})$ is a vector of cell-specific random effects distributed according to a block-specific Gaussian distribution;
- $\epsilon_{ij}(t) \sim \mathcal{N}(0, \sigma^2_{\epsilon,kl})$ being the error term distributed as a block-specific Gaussian;
- $\theta_{kl} = (\mu^{\alpha}_{kl}, \Sigma^{\alpha}_{kl}, \sigma^2_{\epsilon,kl}, \beta_{kl})$.

Note that the ideas borrowed from the *curve registration* framework are here accounted for according to a rather different perspective. While *curve alignment* aims at synchronizing the curves to estimate a common mean one, in our setting the Shape Invariant model works as a suitable tool to model the heterogeneity inside a block and

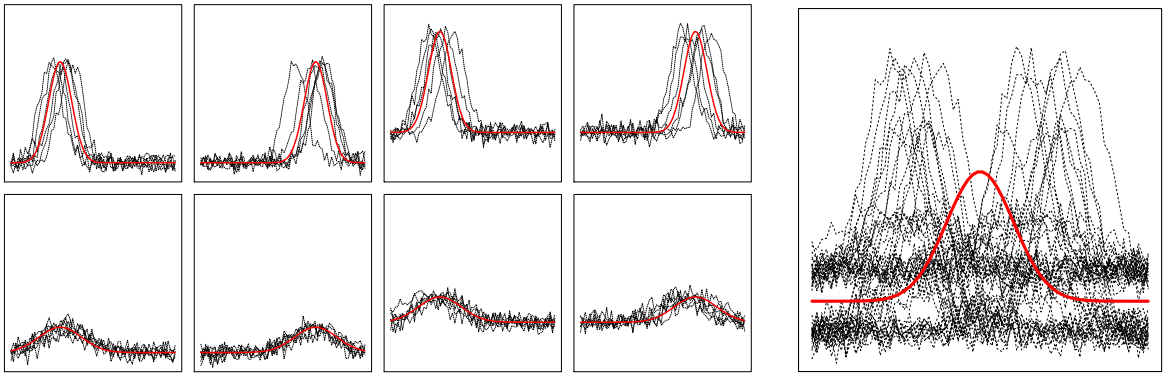


FIGURE 4.1: In the left panels curves in dotted line arise as random fluctuations of the superimposed red curves, but they are all time, amplitude or scale transformations of the same mean-shape function on the right panel.

to introduce a flexible notion of cluster. The rationale behind considering the SIM in a co-clustering framework consists in looking for blocks characterized by a different mean shape function $m(\cdot;\beta_{kl})$. Then, the curves belonging to the same block arise as random shifts and scale transformations of $m(\cdot;\beta_{kl})$, governed by the block-specifically distributed random effects. Consider, for the sake of illustration, the small panels on the left side of Figure 4.1, displaying a number of curves which arise as transformations induced by non-zero values of $\alpha_{ij,1}$, $\alpha_{ij,2}$, or $\alpha_{ij,3}$. In other words, beyond the sample variability, the curves differ for a (phase) random shift on the $x-$ axes, an amplitude shift on the $y-$ axes, and a scale factor. According to model (4.4), all those curves belong to the same cluster, since they share the same mean shape function (see the right panel of Figure 4.1).

In fact, further flexibility can be naturally introduced within the model by "switching off" one or more random effects, depending on subject-matter considerations and on the concept of cluster one has in mind. If there are reasons to support that similar time evolutions associated to different intensities are, in fact, expression of different clusters, it makes sense to switch off the random intercept $\alpha_{ij,1}$. In the example illustrated in Figure 4.1 this ideally leads to a two-clusters structure (4.2, left panels). Similarly, switching off the random effect $\alpha_{ij,3}$ would lead to blocks characterized by a shifted time evolution (Figure 4.2, right panels), while removing the random effect $\alpha_{ij,2}$ determines different blocks varying for a scale factor (Figure 4.2, middle panels).

4.3.2 Model estimation

In order to estimate model (4.4), we propose a modification of the SEM algorithm being able to properly take into account the presence of the random effects. To ease readability,

and with a slight abuse of notation, in the following we suppress the dependency on the time t i.e. x_{ij} has to be intended as $x_{ij}(\mathbf{t_i})$.
Computation of the *complete data log-likelihood* (4.3) associated to model (4.4) is not straightforward, since the marginal density $p(x_{ij};\theta_{kl})$ is defined as

$$p(x_{ij};\theta_{kl}) = \int p(x_{ij}|\alpha_{ij}^{kl};\theta_{kl})p(\alpha_{ij}^{kl};\theta_{kl})\,d\alpha_{ij}^{kl}\ , \tag{4.5}$$

not lending itself to a closed-form expression. To overcome this problem, we propose a *Marginalized SEM-Gibbs* (M-SEM) algorithm as an iterative procedure, in the guise of the SEM. The novelty we introduce consists in considering an additional step, the *Marginalization step*, in order to properly take into account of the random effect and handle (4.5).

Given an initial value for the parameters $\Theta^{(0)}$ and an initial column partition $\mathbf{w}^{(0)}$, the $(q+1)$-*th* iteration of the M-SEM algorithm alternates the following steps:

- **Marginalization step**: obtain the marginal density of each matrix cell by means of Monte Carlo integration as follow

$$p(x_{ij};\theta_{kl}^{(q)}) \simeq \frac{1}{M}\sum_{m=1}^{M} p(x_{ij};\alpha_{ij}^{kl,(m)},\theta_{kl}^{(q)}) \tag{4.6}$$

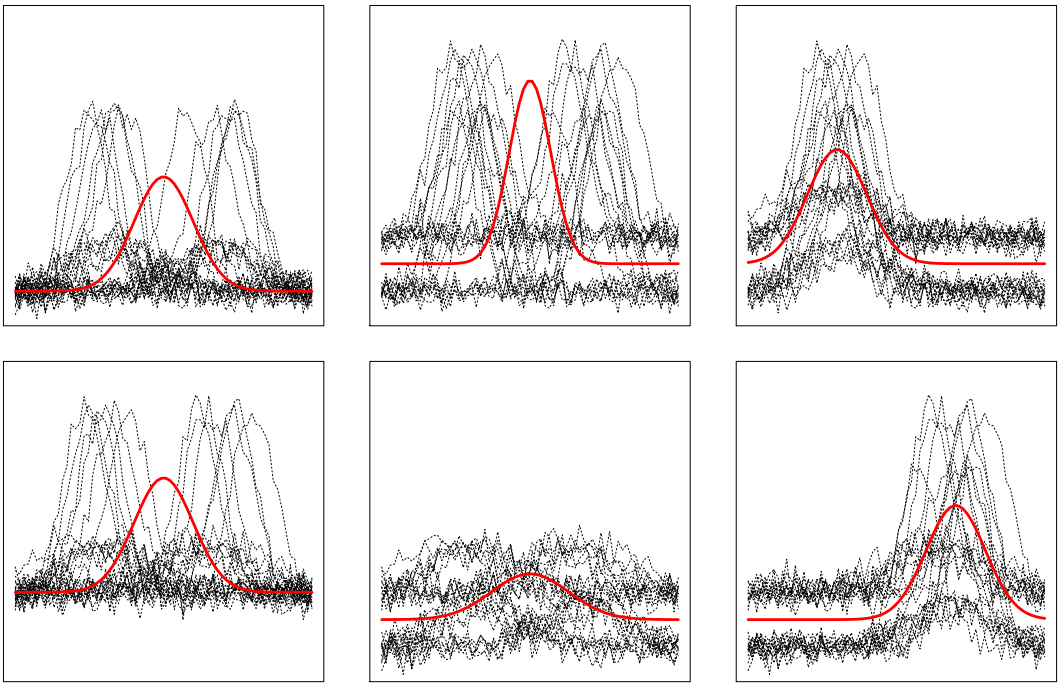


FIGURE 4.2: Pairs of plots in each column represent the two-cluster configurations arising from switching off respectively $\alpha_{ij,1}$ (left), $\alpha_{ij,2}$ (middle), $\alpha_{ij,3}$ (right).

for $i = 1, \ldots, n$, $j = 1, \ldots, d$, $k = 1, \ldots, K$ and $l = 1, \ldots, L$ and being M the number of Monte Carlo samples. The values of the vectors $\alpha_{ij}^{kl,(1)}, \ldots, \alpha_{ij}^{kl,(M)}$ are drawn from a Gaussian distribution $\mathcal{N}_3(\mu_{kl}^{\alpha,(q)}, \Sigma_{kl}^{\alpha,(q)})$;

- **SE step**: repeat, for a number of iterations, the following Gibbs sampling steps:

  1. generate the row partition $z_i^{(q+1)} = (z_{i1}^{(q+1)}, \ldots, z_{iK}^{(q+1)})$, $i = 1, \ldots, n$ according to a multinomial distribution $z_i^{(q+1)} \sim \mathcal{M}(1, \tilde{z}_{i1}, \ldots, \tilde{z}_{iK})$, with

$$\begin{aligned} \tilde{z}_{ik} &= p(z_{ik} = 1 | \mathbb{X}, \mathbf{w}^{(q)}; \Theta^{(q)}) \\ &= \frac{\pi_k^{(q)} p_k(\mathbf{x}_i | w^{(q)}; \Theta^{(q)})}{\sum_{k'} \pi_{k'}^{(q)} p_{k'}(\mathbf{x}_i | w^{(q)}; \Theta^{(q)})} , \end{aligned}$$

     for $k = 1, \ldots, K$, $\mathbf{x}_i = \{x_{ij}\}_{1 \leq j \leq d}$ and $p_k(\mathbf{x}_i | \mathbf{w}^{(q)}; \Theta^{(q)}) = \prod_{jl} p(x_{ij}; \theta_{kl}^{(q)})^{w_{jl}^{(q)}}$.

  2. generate the column partition $w_j^{(q+1)} = (w_{j1}^{(q+1)}, \ldots, w_{jL}^{(q+1)})$, $j = 1, \ldots, d$ according to a multinomial distribution $w_j^{(q+1)} \sim \mathcal{M}(1, \tilde{w}_{j1}, \ldots, \tilde{w}_{jL})$, with

$$\begin{aligned} \tilde{w}_{jl} &= p(w_{jl} = 1 | \mathbb{X}, \mathbf{z}^{(q+1)}; \Theta^{(q)}) \\ &= \frac{\rho_l^{(q)} p_l(\mathbf{x}_j | \mathbf{z}^{(q+1)}; \Theta^{(q)})}{\sum_{l'} \rho_{l'}^{(q)} p_{l'}(\mathbf{x}_j | \mathbf{z}^{(q+1)}; \Theta^{(q)})} , \end{aligned}$$

     for $l = 1, \ldots, L$, $\mathbf{x}_j = \{x_{ij}\}_{1 \leq i \leq n}$ and $p_l(\mathbf{x}_j | \mathbf{z}^{(q+1)}; \Theta^{(q)}) = \prod_{ik} p(x_{ij}; \Theta_{kl}^{(q)})^{z_{ik}^{(q+1)}}$.

- **M step**: Estimate $\Theta^{(q+1)}$ conditionally on $\mathbf{z}^{(q+1)}$ and $\mathbf{w}^{(q+1)}$. Mixture proportions are updated as $\pi_k^{(q+1)} = \frac{1}{n} \sum_i z_{ik}^{(q+1)}$ and by $\rho_l^{(q+1)} = \frac{1}{d} \sum_j w_{jl}^{(q+1)}$.
  An estimate of $\theta_{kl} = (\mu_{kl}^{\alpha}, \Sigma_{kl}^{\alpha}, \sigma_{\epsilon,kl}^2, \beta_{kl})$ can be obtained by noting that the specification in (4.4) results in a *non-linear mixed effect model*. Estimation in this framework is not straighforward and closed-form solutions are not available. In this work we use the approximate maximum likelihood approach as proposed in Lindstrom and Bates (1990); the variance and the mean components are estimated by approximating and then maximizing the marginal density of the latter near the mode of the posterior distribution of the random effects. Conditional or shrinkage estimates are afterwards used for the estimation of the random effects.

The M-SEM algorithm is run for a certain number of iterations until a convergence criterion on the *complete data log-likelihood* is met. Since a burn-in period is considered, the final estimate for Θ, denoted as $\hat{\Theta}$, is given by the mean of the sample distribution. A sample of $(\mathbf{z}, \mathbf{w})$ is then generated according to the SE step as illustrated above with

$\Theta = \hat{\Theta}$. The final block-partition $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$ is then obtained as the mode of their sample distribution.

Coherently with the parametric clustering formulation, the choice of the number of groups is recasted here to a model selection problem. In accordance with a standard model-based framework (Section 1.2.2), the *single best model paradigm* is operationally considered. Several models, corresponding to different combinations of K and L and, in our case, to different configurations for turning on and off the random effects, are estimated and the best one is selected according to a specific information criterion. Keribin *et al.* (2015) have noted that in the co-clustering framework the BIC is not a viable solution, as the penalization term no longer remains valid due to the dependency structure of the observed data $\mathbb{X}$. As an alternative, we consider an approximated version of the ICL (Biernacki *et al.*, 2000) that, relying on the *complete data likelihood* does not suffer of the same issues as the BIC. The considered criterion is then defined as

$$\text{ICL} = \log p(\mathbb{X}, \hat{z}, \hat{w}; \hat{\Theta}) - \frac{K-1}{2} \log n - \frac{L-1}{2} \log d - \frac{KL\nu}{2} \log nd \ , \tag{4.7}$$

where ν denotes number of specific parameters for each block and

$$\log p(\mathbb{X}, \hat{z}, \hat{w}; \hat{\Theta}) = \prod_{ik} \hat{z}_{ik} \log \hat{\alpha}_k + \prod_{jl} \hat{w}_{jl} \log \hat{\beta}_l \sum_{ijkl} \hat{z}_{ik} \hat{w}_{jl} \log p(x_{ij}, \hat{\theta}_{kl}) \ . \tag{4.8}$$

The selected model will be the one corresponding to the pair (K, L) attaining the highest value for the ICL.

4.3.3 Computational remarks

The model introduced so far inherits the advantages of both the building ingredients it embeds. Thanks to the local independence assumption of the LBM, it allows handling multivariate, possibly high dimensional complex data structures in a relatively parsimonious way. Differences among the subjects are captured by the random effects, while curve summaries can be expressed as a functional of the mean shape curve. Additionally, the recourse to a smoother when modeling the mean shape function allows for a flexible handling of functional data while the presence of random effects lends the model to be applied to a longitudinal setting. Finally, clustering is pursued directly on the observed curves, without resorting to intermediate transformation steps, as in Bouveyron *et al.* (2018).

These reasons of attractiveness should not distract from the caution required by some aspects, especially of computational nature, of the proposed method, discussed in the following.

- *Initialization* The M-SEM algorithm encloses different numerical steps which require the suitable specification of starting values.
  The EM algorithm and its modifications are known to be very sensible to the initialization. Furthermore since the convergence towards a global maximum of the likelihood is not guaranteed, to avoid local solutions a proper initialization strategy is crucial. Assuming K and L to be known, the M-SEM algorithm requires starting values for $\mathbf{z}$ and $\mathbf{w}$ in order to implement the first M-step. In this work we consider two different initialization strategies:

  – (Multiple) random initialization: the row and column partitions are sampled independently from multinomial distributions with uniform weights. If multiple initializations are considered, the one eventually leading to the highest value of the *complete data log-likelihood* is retained;

  – K-means initialization: two k-means algorithms are independently run for the rows and the columns of $\mathbb{X}$ and the M-SEM algorithm is initialized with the obtained partitions $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$.

  Anyhow it has been pointed out (see e.g. Govaert and Nadif, 2013) that the SEM, being a stochastic algorithm, can attenuate in practice the impact of the initialization on the resulting estimates.
  Lastly note that a further initialization is required, to estimate the nonlinear mean shape function within the M-step.

- *Convergence and other numerical problems.* Although the benefits of including random effects in the considered framework are undeniable, parameter estimation is known not to be straightforward in mixed effect models, especially in the nonlinear setting (see, e.g. Harring and Liu, 2016). The nonlinear dependence of the conditional mean of the response on the random effects requires multidimensional integration over the random effects distribution to derive the marginal distribution of the data. In fact, this integral is almost always intractable. While several methods such as, for example, likelihood approximation have been proposed to overcome this issue, serious convergence problems are often encountered. In such situations, some devices can come in aid to ease convergence of the estimation algorithm. Examples are to try different sets of starting values, to scale the data

prior to the modeling step, or to simplify the nonlinear structure of the model (e.g. in the case of B-splines, by reducing the number of knots). Even when convergence is eventually achieved, in fact, addressing these issues often results in considerable computational times. Depending on the specific data at hand, it is also possible to consider alternative mean shape formulations, such as polynomial functions, which result in easier estimation procedures.

- *Curse of flexibility.* Including random effects for both phase and amplitude shifts and scale transformations allows for a virtually excellent fitting of any arbitrarily shaped curve. This flexibility, albeit desirable, may achieve excessive extents, turning out to estimation troubles. This is especially true in a clustering framework, when data are expected to exhibit a remarkable heterogeneity. From a practical point of view our experience suggests that the estimation of the parameters $\alpha_{ij,2}$ turns out to be the most troublesome, sometimes leading to convergence issues and instability in the resulting estimates. In agreement with the final considerations of Section 4.3.1, and when a preliminary exploration of data suggests highly heterogeneous curve shapes, a sensible workaround consists in switching off $\alpha_{ij,2}$, resulting in the search for clusters which are homogeneous in the curve scale. Operationally, switching off a random effect boils down to resort to a constrained estimation scheme in the M step of the M-SEM algorithm. Let us consider a diagonal specification for Σ^{α}_{kl} for all the blocks. A random effect is then practically turned off by constraining its mean and variance estimates to be exactly equal to zero.

- *Label switching.* As every stochastic algorithm due to the resort to a Gibbs sampling, SEM is in principle subject to label switching (see Frühwirth-Schnatter, 2006, for a detailed tractation of the topic). Nonetheless it has been pointed out by Keribin *et al.* (2015) that most of the time this phenomenon is not encountered in practical situations. This indication has been confirmed by the results we obtained in Section 4.4.

4.4 Numerical examples

4.4.1 Simulation study

This section aims at exploring the main features of the proposed approach on some synthetic data. We focus on the capability of the method to properly partition the data into blocks, analyzing thoroughly the subsequent groups obtained from a clustering

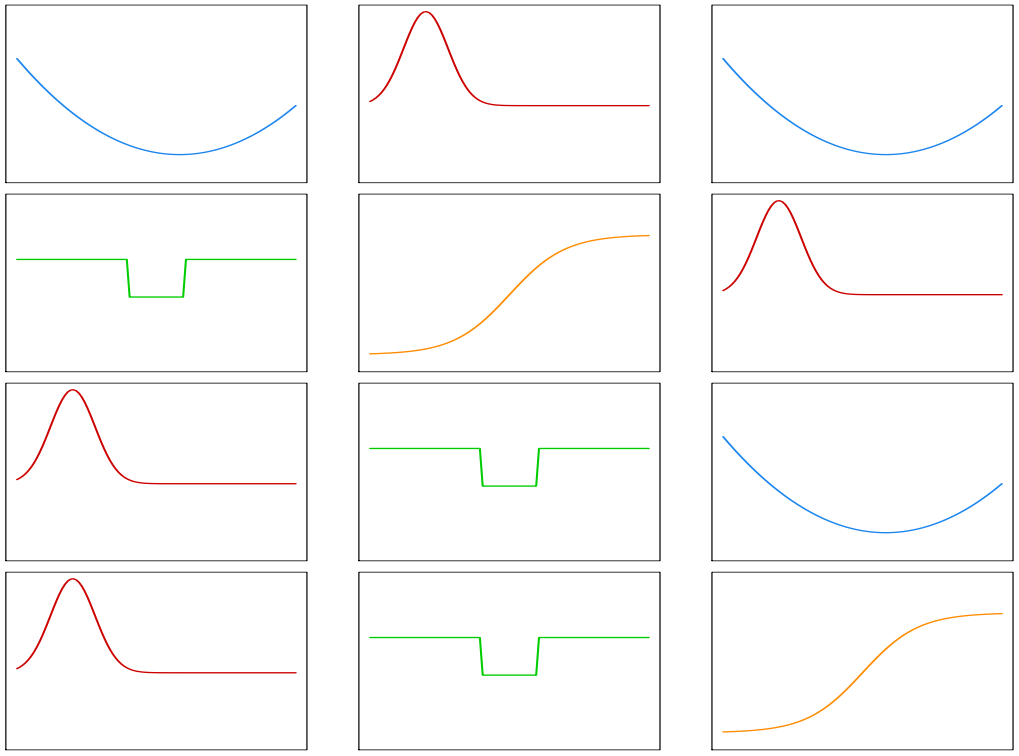


FIGURE 4.3: Block specific mean shape curves employed in the simulation study.

quality perspective. Moreover we explore the validity of the ICL criterion when used to choose the number of blocks and when considered to select an appropriate random effect configuration.

All the analyes have been conducted in the R environment (R Core Team, 2019) with the aid of `nlme` (Pinheiro *et al.*, 2019), `funLBM` (Bouveyron and Jacques, 2018) and `splines` packages. The code implementing the proposed procedure is available upon request.

The examined simulation setup is defined as follows. We generated $B = 100$ Monte Carlo samples of curves according to the general specification (4.4), with block-specific mean shape function $m_{kl}(\cdot)$ and both the parameters involved in the error term and the ones describing the random effects distribution considered constant across the blocks. In fact, in the light of the final considerations made in Section 4.3.3, the random scale parameter is switched off in the data generative mechanism, i.e. $\alpha_{ij,2}$ is constrained to be degenerate in zero. The number of row and column clusters has been fixed to $K_{\text{true}} = 4$ and $L_{\text{true}} = 3$ and the mean shape functions $m_{kl}(\cdot)$ are chosen among four different curves namely $m_{11} = m_{13} = m_{33} = m_1$, $m_{12} = m_{32} = m_{31} = m_{41} = m_2$, $m_{21} = m_{32} = m_{42} = m_3$ and $m_{22} = m_{43} = m_4$, as illustrated in Figure 4.3.

We set $n = 100$ rows, $d = 20$ columns and $T = 15$ equi-spaced time points ranging in

TABLE 4.1: Mean over the Monte Carlo samples of the Adjusted Rand Index for both the row and column partitions obtained, as a function of the detected number of groups. The true number of row clusters is 4, while the true number of column clusters is 3.

	2	3	4	5
ARI (row)	-	0.81	0.98	0.89
ARI (column)	-	1.00	0.85	0.69

$[0, 1]$. Further details about the simulation setting parameters, as well as the specification of the curves considered, are provided in the Appendix.

Model estimation is performed by setting the first 10 iterations of the M-SEM algorithm as a burn-in period while, for the B-spline basis functions, we considered 4 knots.

With the aim of evaluating the ability of the proposed methodology in recovering the true clustering structure, we have computed the Adjusted Rand Index for both the row and the column partitions, disregarding the number of detected blocks. Results are illustrated in Table 4.1. The obtained performances are extremely satisfactory. Indeed, when the selected number of groups corresponds to the true one, the ARI values indicates nearly perfect classifications of rows and columns. Moreover, even when the ICL detects a wrong number of blocks, a satisying clustering quality is still witnessed; this gives an indication about the identification of clusters of curves that retain the homogeneity of the curves inside the groups.

As a second goal of the numerical analysis we have explored the capabilities of the ICL to detect the right number of blocks in the data, disregarding the random effect configuration. To this aim, the best model has been selected among different choices for K and L, with the random effect configuration assumed to be known and fixed. Table 4.2 reports the percentage of samples for which each number of considered row and column cluster has been selected across the Monte Carlo samples. Again we obtained satisfactory results with a slight tendency to overestimate the number of blocks. While it would be interesting to further investigate this behavior, it is nonetheless conceptually preferable with respect to underestimation since, potentially, it does not hinder the homogeneity within a block, being the final aim of cluster analysis.

As a third goal of the study, we assessed if the ICL criterion represents a valid strategy also when used to select among different random effects configuration. To this aim, the number of blocks has been set to the true values and models corresponding to different configuration for α_{ij}^{kl} are estimated. As highlighted in Section 4.3.3, the estimation of model configurations in the presence of the scale random effect is sometimes

TABLE 4.2: Percentage of selection for each model (K, L) on the 100 simulated datasets. Bold cell represents the true number of blocks.

K/L	2	3	4	5
2	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0
4	0.0	**57.6**	10.1	8.1
5	0.0	18.2	4.0	1.0

TABLE 4.3: Percentage of selection for each random effects configuration over 100 simulated datasets. `T` means that the corresponding random effect is switched on while `F` means that is switched off. As an example `FTT` represent a model where $\alpha_{ij,1}$ is constrained to be a random variable with degenerate distribution in zero. Bold cell represents the true data generative model.

	FFF	TFF	FTF	FFT	TTF	TFT	FTT	TTT
% of selection	0.0	0.0	0.0	0.0	0.0	**96.9**	0.0	3.1

troublesome, and has not converged for a few cases, so that results, reported in Table 4.3, are somewhat biased toward the selection of the true generative mechanism, as well as towards the other model configurations with $\alpha_{ij,2}$ switched off. In fact, none among the latter configurations has been selected across any of the Monte Carlo samples, and the true generative configuration of random effects is selected in the great majority of the cases.

Due to the mentioned computational complexity of the proposed methodology the numerical exploration has been limited to the mentioned goals. In fact, an improved understanding of the strengths and weaknesses of the proposed methodology is left for future work, and meant to be reached by extending the focus of the simulation to alternative choices of n, d, and T, and to the comparison with some competitors. Also, note that further aspects such as the selection of the number of knots in the splines has not been thoroughly studied in this work but it could constitute an interesting further development; by removing some knots we may indeed, losing something in terms of flexibility, obtain faster and simpler procedures from a computational point of view.

4.4.2 Real data illustration

The data we consider in this section (publicly available at `https://www.pollens.fr/en/reports/database`) are provided by the *Réseau National de Surveillance Aérobiologique* (RNSA), the French institute which analyzes the biological particles content of the air and studies their impact on the human health. RNSA collects data on concentration of pollens and moulds in the air, along with some clinical data, in more than 70 municipalities in France.

The analyzed dataset contains daily observations of the concentration of 21 pollens for 71 cities in France in 2016. Concentration is measured as the number of pollens detected over a cubic meter of air and carried on by means of some pollen traps located in central urban positions over the roof of buildings, in order to be representative of the trend air quality.

The aim of the analysis is to identify homogeneous trends in the pollen concentration over the year and across different geographical areas. For this reason, we focus on finding groups of pollens differentiating one from the others for either the period of maximum exhibition or the timespan they are present. Consistently with this choice, only models with the y-axis shift parameter $\alpha_{ij,1}$ are estimated (i.e. $\alpha_{ij,2}$ and $\alpha_{ij,3}$ are switched off), for varying number of row and column clusters, and the best one selected via ICL. We consider monthly data by averaging the observed daily concentrations over months. The resulting dataset may be represented as a matrix with $n = 71$ rows (cities), $p = 21$ columns (pollens) where each entry is a sequence of $T = 12$ time-indexed measurements.

In order to practically apply our proposed procedure on the data a preprocessing step has been carried out. We work on a logarithmic scale and, in order to improve the stability of the estimation procedure, the data have been standardized.

Results are graphically displayed in Figure 4.4. The ICL selects a model with $K = 3$ row clusters and $L = 5$ column ones. A first visual inspection of the pollen time evolutions reveals that the procedure is able to discriminate the pollens according to their seasonality. Pollens in the first two column groups are mainly present during the summer, with a difference in the intensity of the concentration. In the remaining three groups pollens are more active, grossly speaking, during winter and spring months but with a different time persistence and evolution.

Digging deeper substantially in the cluster configuration obtained is beyond the scope of this work and may benefit from some help and insights from experts of botanical and geographical disciplines. Anyway it stands out that column clusters are roughly grouping together trees pollens, distinguishing them from weed and grass ones (left panel of Figure 4.5). Results are also coherent with the usually considered typical seasons, with groups of pollens from trees mostly present in winter and spring while the ones from grass spreading in the air mainly during the summer months. With respect to the row partition, displayed in the row panel of Figure 4.5, three clusters have been detected, with the one in light blue on the map roughly corresponding to the Mediterranean region. The situation, for what it concerns the other two clusters, appears to be more heterogeneous. One of these groups (in red on the map) tends to gather together cities in the northern region and on the Atlantic coast while the other

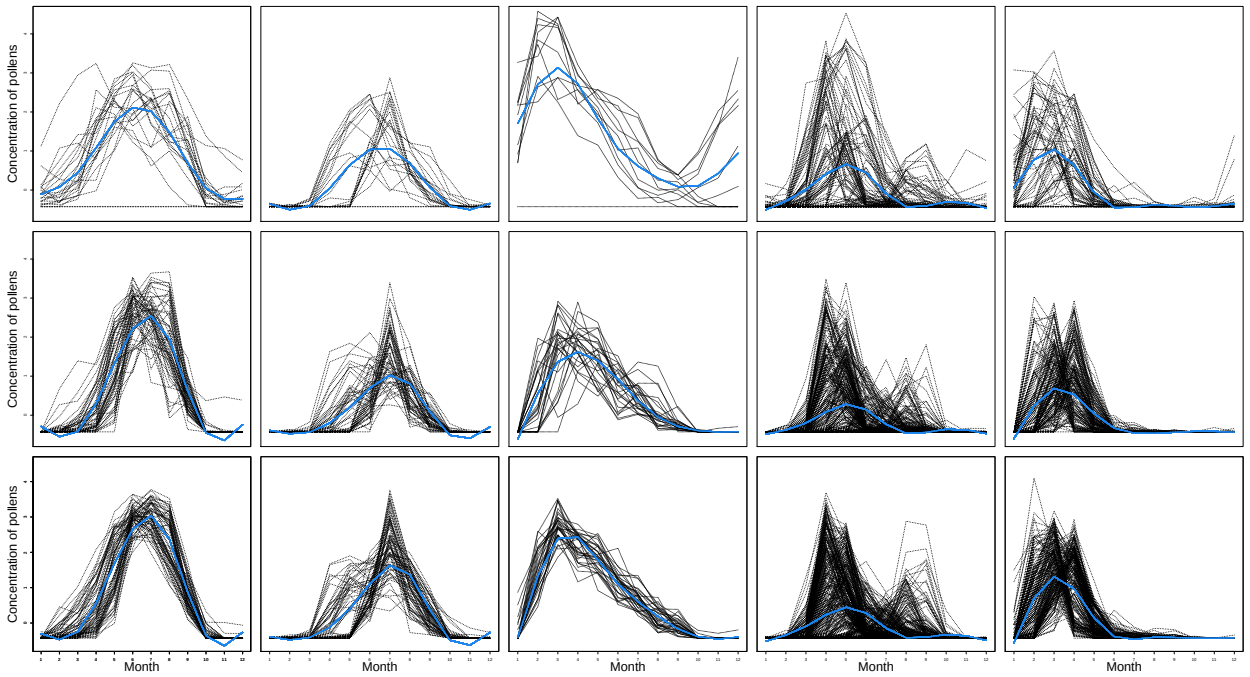


FIGURE 4.4: Curves belonging to each single block with superimposed the corresponding block specific mean curve (in light blue).

(in green) cover mainly the central and continental part of the country. Again the results appear promising but it may be beneficial a cross analysis with some climate scientists in order to get a more informative and substantiated point of view.

4.5 Conclusions

Multivariate time-dependent data can be suitably arranged in three-way structures where each layer introduces its own peculiar characteristics. When exploring appropriate modelling strategies it is required to account for heterogeneous subjects, relations among variables and correlation across different time instants.

This chapter has aimed at answering these challenges by proposing a new parametric co-clustering methodology, recasting to the widely known Latent Block Model in a time-dependent fashion. The co-clustering model, by simultaneously searching for row and column clusters, partitions three-way matrices in some blocks formed by homogeneous curves. Such an approach seems particularly reasonable in the considered framework since it takes into account the mentioned features of the data while building parsimonious and meaningful summaries. As a data generative mechanism for a single curve we have considered the *Shape Invariant Model* that has turned out to be particularly flexible when embedded in a co-clustering context. The model allows describing

	Pollens group
1	Gramineae, Urticaceae
2	Chestnut, Plantain
3	Cypress
4	Ragweed, Mugwort, Birch, Beech Morus, Olive, Platanus, Oak, Sorrel Linden
5	Alder, Hornbeam, Hazel, Ash Poplar, Willow

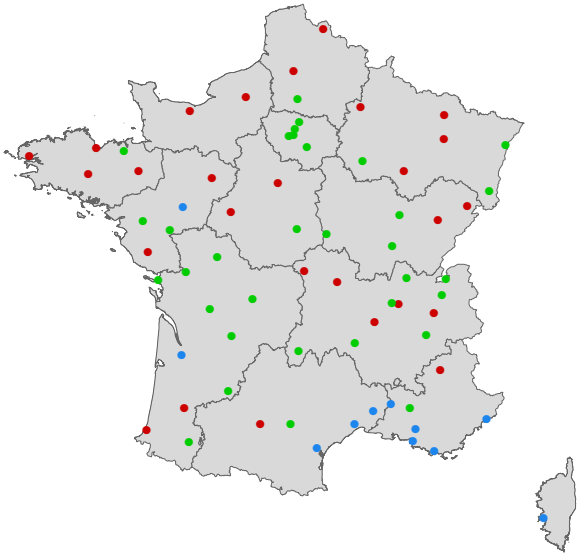


FIGURE 4.5: On the left: Pollens organized according to the column cluster memberships. On the right: French map with overimposed the points indicating the cities colored according to their row cluster memberships.

arbitrary time evolution patterns while adequately capturing dependencies among different temporal instants. The further chance of "switching off" some of the random effects, although in principle simplifying the model structure, increases its flexibility, as it allow encompassing different concepts of cluster possibly depending on the specific applications and on subject-matter considerations.

While further analyses and comparisons with alternative models are required to increase our understanding about the general performance of the proposed model, its first application to both simulated and real data has proved overall satisfactory results and highlighted some aspects which will worth further investigation. Among them, we shall introduce the idea of the *curse of flexibility*, as in some specific situations, the considered specification may induce a degree of flexibility possibly entailing issues from a computational point of view and in the obtained results. Some alternative choices, for example to model the block mean curves, could be considered and compared with the ones adopted here. A further direction for future work would consist in exploring the chance to resort to a fully Bayesian estimation approach, possibly handling more easily the random parameters in the model.

Appendix

Parameter settings - Chapter 2

In the following the parameter settings of the densities selected for the simulations in Chapter 2 are presented. Since all the densities are mixtures of Gaussian models, we adopt the usual notation where, for a given k component, π_k represents the k-th mixture weight, μ_k and σ_k^2 (Σ_k for the bivariate models) the mean and variance (covariance matrix).

Unidimensional parameter settings

Density M1

Components	π_k	μ_k	σ_k^2
1	0.75	0.00	0.83
2	0.25	1.37	0.09

Density M2

Components	π_k	μ_k	σ_k^2
1	0.45	-0.93	0.22
2	0.45	0.93	0.22
3	0.1	0.00	0.04

Density M3

Components	π_k	μ_k	σ_k^2
1	0.5	-0.74	0.14
2	0.3	0.37	0.55
3	0.2	1.47	0.14

Density M4

Components	π_k	μ_k	σ_k^2
1	0.15	0.00	0.44
2	0.15	-0.33	0.19
3	0.5	-0.99	0.14
4	0.2	1.32	0.19

Density M5

Components	π_k	μ_k	σ_k^2
1	0.5	0.00	0.14
2	0.35	1.28	0.14
3	0.15	2.56	0.11

Bidimensional settings

Asymmetric bimodal

Components	π_k	μ_k	Σ_k
1	0.5	$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 0.44 & 0.31 \\ 0.31 & 0.44 \end{pmatrix}$
2	0.5	$\begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.44 & 0 \\ 0 & 0.44 \end{pmatrix}$

Trimodal

Components	π_k	μ_k	Σ_k
1	0.43	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.36 & 0.25 \\ 0.25 & 0.49 \end{pmatrix}$
2	0.43	$\begin{pmatrix} 1 \\ 1.15 \end{pmatrix}$	$\begin{pmatrix} 0.36 & 0 \\ 0 & 0.49 \end{pmatrix}$
3	0.14	$\begin{pmatrix} 1 \\ -1.15 \end{pmatrix}$	$\begin{pmatrix} 0.36 & 0 \\ 0 & 0.49 \end{pmatrix}$

Parameter settings - Chapter 3

In the following the parameter settings of the densities selected for the simulations in Chapter 3 are presented. The notation remains unchanged for Density M1, M2 and M3 being Gaussian mixture models. On the other hand for Density M4 and M5 we consider multivariate skew normal distributions (or mixture of) hence the additional parameter δ_k regulates the skeweness of the k-th component.

Density M1

Components	π_k	μ_k	Σ_k
1	1	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1.25 & 0.75 \\ 0.75 & 1.25 \end{pmatrix}$

Density M2

Components	π_k	μ_k	Σ_k
1	0.5	$\begin{pmatrix} -0.53 \\ -0.53 \end{pmatrix}$	$\begin{pmatrix} 0.68 & -0.41 \\ -0.41 & 0.68 \end{pmatrix}$
2	0.5	$\begin{pmatrix} 0.53 \\ 0.53 \end{pmatrix}$	$\begin{pmatrix} 0.68 & -0.41 \\ -0.41 & 0.68 \end{pmatrix}$

Density M3

Components	π_k	μ_k	Σ_k
1	0.4	$\begin{pmatrix} -0.85 \\ -0.85 \end{pmatrix}$	$\begin{pmatrix} 0.58 & -0.35 \\ -0.35 & 0.58 \end{pmatrix}$
2	0.4	$\begin{pmatrix} 0.85 \\ 0.85 \end{pmatrix}$	$\begin{pmatrix} 0.58 & -0.35 \\ -0.35 & 0.58 \end{pmatrix}$
3	0.2	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.16 & -0.09 \\ -0.09 & 0.16 \end{pmatrix}$

Density M4

Components	π_k	μ_k	Σ_k	δ_k
1	1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0.8 & -0.4\\-0.4 & 0.8\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$

Density M5

Components	π_k	μ_k	Σ_k	δ_k
1	0.5	$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}0.8 & -0.4\\-0.4 & 0.8\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$
2	0.5	$\begin{pmatrix}-1\\-1\end{pmatrix}$	$\begin{pmatrix}0.8 & -0.4\\-0.4 & 0.8\end{pmatrix}$	$\begin{pmatrix}-3\\-3\end{pmatrix}$

Parameter settings - Chapter 4

In the following the parameter considered for the simulations in Chapter 4, as well as the specification of the functions used as block specific mean curves, are presented.
The curves considered are specified as follows

$$\begin{aligned} m_1(t) &\propto 6t^2 - 7t + 1 \\ m_2(t) &\propto \phi(t; 0.2, 0.008) \\ m_3(t) &\propto 0.75 - 0.8\mathbb{1}_{\{t\in(0.4,0.6)\}} \\ m_4(t) &\propto \frac{1}{(1+\exp(-10t+5))} \end{aligned}$$

Concerning the parameters involved in the error terms and in the random effects distribution $\sigma_{\epsilon,kl} = 0.3$, $\mu^{\alpha}_{kl} = (0,0,0)$ and $\Sigma^{\alpha}_{kl} = \mathrm{diag}(1,0,0.1)$ $\forall k = 1,\dots,K, l = 1,\dots,L$.

Bibliography

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., Scheuermann, R. H., FlowCAP Consortium and DREAM Consortium (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nature methods* **10**(3), 228.

Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society: Series A (General)* **144**(4), 419–448.

Ameijeiras-Alonso, J., Crujeiras, R. M. and Rodríguez-Casal, A. (2018) multimode: an r package for mode assessment. *arXiv preprint arXiv:1803.00472* .

Anderlucci, L. and Viroli, C. (2015) Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics* **9**(2), 777–800.

Azzalini, A. and Torelli, N. (2007) Clustering via nonparametric density estimation. *Statistics and Computing* **17**(1), 71–80.

Baillo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001) Convergence rates in nonparametric estimation of level sets. *Statistics & probability letters* **53**(1), 27–35.

Banfield, J. D. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics* pp. 803–821.

Ben-David, S., Von Luxburg, U. and Pál, D. (2006) A sober look at clustering stability. In *In Proceedings of the 19th Annual Conference on Learning Theory (G. Lugosi and H.-U. Simon, eds.)*, pp. 5–19.

Ben Slimen, Y. S., Allio, S. and Jacques, J. (2018) Model-based co-clustering for functional data. *Neurocomputing* **291**, 97–108.

Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* **22**(7), 719–725.

Bouveyron, C., Bozzi, L., Jacques, J. and Jollois, F. X. (2018) The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(4), 897–915.

Bouveyron, C., Celeux, G., Murphy, T. B. and Raftery, A. E. (2019) *Model-Based Clustering and Classification for Data Science: With Applications in R*. Volume 50. Cambridge University Press.

Bouveyron, C., Côme, E. and Jacques, J. (2015) The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics* **9**(4), 1726–1760.

Bouveyron, C. and Jacques, J. (2011) Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification* **5**(4), 281–300.

Bouveyron, C. and Jacques, J. (2018) *funLBM: Model-Based Co-Clustering of Functional Data*. R package version 1.0.

Carmichael, J., George, J. and Julius, R. (1968) Finding natural clusters. *Systematic Zoology* **17**(2), 144–150.

Carreira-Perpinán, M. A. (2008) Generalised blurring mean-shift algorithms for non-parametric clustering. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.

Celeux, G. and Diebolt, J. (1985) The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly* **2**, 73–82.

Celeux, G. and Govaert, G. (1992) A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis* **14**(3), 315–332.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern recognition* **28**(5), 781–793.

Chacón, J. E. (2015) A population background for nonparametric density-based clustering. *Statistical Science* **30**(4), 518–532.

Chacón, J. E. (2019) Mixture model modal clustering. *Advances in Data Analysis and Classification* **13**(2), 379–404.

Chacón, J. E. and Duong, T. (2013) Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics* **7**, 499–532.

Chacón, J. E. and Duong, T. (2018) *Multivariate kernel smoothing and its applications*. Chapman and Hall/CRC.

Chacón, J. E., Duong, T. and Wand, M. (2011) Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica* pp. 807–840.

Chacón, J. E. and Monfort, P. (2006) A comparison of bandwidth selectors for mean shift clustering. In *Theoretical and Applied Issues in Statistics and Demography (C. H. Skiadas, ed.)*, pp. 47–59.

Chen, Y.-C., Genovese, C. R., Wasserman, L. *et al.* (2016) A comprehensive approach to mode clustering. *Electronic Journal of Statistics* **10**(1), 210–241.

Chen, Y.-C., Genovese, C. R., Wasserman, L. *et al.* (2017) Statistical inference using the morse-smale complex. *Electronic Journal of Statistics* **11**(1), 1390–1433.

Cheng, Y. (1995) Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence* **17**(8), 790–799.

Chernoff, H. (1964) Estimation of the mode. *Annals of the Institute of Statistical Mathematics* **16**(1), 31–41.

Claeskens, G. and Hjort, N. (2008) *Model selection and model averaging*. Cambridge University Press.

Comaniciu, D. and Meer, P. (2002) Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (5), 603–619.

Corneli, M., Bouveyron, C. and Latouche, P. (2019) Co-clustering of ordinal data via latent continuous random variables and a classification em algorithm. *HAL preprint hal-01978174* .

De la Cruz-Mesía, R., Quintana, F. A. and Marshall, G. (2008) Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis* **52**(3), 1441–1457.

Cuevas, A., Febrero, M. and Fraiman, R. (2000) Estimating the number of clusters. *Canadian Journal of Statistics* **28**(2), 367–382.

Cuevas, A., Febrero, M. and Fraiman, R. (2001) Cluster analysis: a further approach based on density estimation. *Computational Statistics & Data Analysis* **36**(4), 441–459.

De Boor, C. (1978) *A practical guide to splines*. Springer-Verlag, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: the L_1 View*. Wiley, New York.

Dietterich, T. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning* **40**(2), 139–157.

Doss, C. R. and Weng, G. (2018) Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. *Electronic Journal of Statistics* **12**(2), 4313–4376.

Duong, T. (2019) *ks: Kernel Smoothing*. R package version 1.11.4.

Einbeck, J. (2011) Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of pattern recognition research.* **6**(2), 175–192.

Fern, X. Z. and Brodley, C. E. (2003) Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 186–193.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of eugenics* **7**(2), 179–188.

Forina, M., Armanino, C., Castino, M. and Ubigli, M. (1986) Multivariate data analysis as a discriminating method of the origin of wines. *Vitis* **25**(3), 189–201.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97**(458), 611–631.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The elements of statistical learning*. Springer-Verlag, New York.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer series in Statistics.

Frühwirth-Schnatter, S. (2011) Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification* **5**(4), 251–280.

Fukunaga, K. and Hostetler, L. (1975) The estimation of the gradient of a density fujnction, with applications in pattern recognition. *IEEE Transactions on information theory* **21**(1), 32–40.

Govaert, G. and Nadif, M. (2003) Clustering with block mixture models. *Pattern Recognition* **36**(2), 463–473.

Govaert, G. and Nadif, M. (2008) Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis* **52**(6), 3233–3245.

Govaert, G. and Nadif, M. (2010) Latent block model for contingency table. *Communications in Statistics - Theory and Methods* **39**(3), 416–425.

Govaert, G. and Nadif, M. (2013) *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Grund, B. and Hall, P. (1995) On the minimisation of the L^p error in mode estimation. *The Annals of Statistics* **23**(6), 2264–2284.

Hall, P. and Marron, J. (1991) Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields* **90**, 149–173.

Hall, P. and Wand, M. P. (1988) On the minimization of absolute distance in kernel density estimation. *Statistics & probability letters* **6**(5), 311–314.

Harring, J. R. and Liu, J. (2016) A comparison of estimation methods for nonlinear mixed-effects models under model misspecification and data sparseness: A simulation study. *Journal of Modern Applied Statistical Methods* **15**(1), 27.

Hartigan, J. (1975) *Clustering Algorithms*. J. Wiley & Sons, New York.

Hennig, C., Meila, M., Murtagh, F. and Rocci, R. (2016) *Handbook of Cluster Analysis*. Chapman and Hall.

Hornik, K. (2018) *Clue: Cluster ensembles*. R package version 0.3-55.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of classification* **2**(1), 193–218.

Ingrassia, S. and Rocci, R. (2007) Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis* **51**(11), 5339–5351.

Ingrassia, S. and Rocci, R. (2011) Degeneracy of the em algorithm for the mle of multivariate gaussian mixtures and dynamic constraints. *Computational statistics & data analysis* **55**(4), 1715–1725.

Jacques, J. and Biernacki, C. (2018) Model-based co-clustering for ordinal data. *Computational Statistics & Data Analysis* **123**, 101–115.

Jones, M. (1992) Potential for automatic bandwidth choice in variations on kernel density estimation. *Statistics & probability letters* **13**(5), 351–356.

Keribin, C., Brault, V., Celeux, G. and Govaert, G. (2015) Estimation and selection for the latent block model on categorical data. *Statistics and Computing* **25**(6), 1201–1216.

Kuncheva, L. and Hadjitodorov, S. (2004) Using diversity in cluster ensembles. In *2004 IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pp. 1214–1219.

Lawton, W., Sylvestre, E. and Maggio, M. (1972) Self modeling nonlinear regression. *Technometrics* **14**(3), 513–532.

Leeb, H. and Pötscher, B. (2005) Model selection and inference: Facts and fiction. *Econometric Theory* **21**(1), 21–59.

Leone, F., Nelson, L. and Nottingham, R. (1961) The folded normal distribution. *Technometrics* **3**(4), 543–550.

Li, J. (2005) Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* **14**(3), 547–568.

Li, J., Ray, S. and Lindsay, B. G. (2007) A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research* **8**, 1687–1723.

Liao, T. W. (2005) Clustering of time series data - a survey. *Pattern recognition* **38**(11), 1857–1874.

Lin, T. I. (2009) Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* **100**(2), 257–265.

Lin, T. I. (2010) Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing* **20**(3), 343–356.

Lindstrom, M. J. (1995) Self-modelling with random shift and scale parameters and a free-knot spline shape function. *Statistics in Medicine* **14**(18), 2009–2021.

Lindstrom, M. J. and Bates, D. M. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**(3), 673–687.

Lisic, J. (2018) *MeanShiftR: A Computationally Efficient Mean Shift Implementation*. R package version 0.52.

Lomet, A. (2012) *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Compiègne.

Madigan, D. and Raftery, A. (1994) Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association* **89**(428), 1535–1546.

Malsiner-Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2017) Identifying mixtures of mixtures using bayesian estimation. *Journal of Computational and Graphical Statistics* **26**(2), 285–295.

Matsumoto, Y. (2002) *An introduction to Morse Theory*. American Mathematical Society.

McLachlan, G. and Krishnan, T. (2007) *The EM algorithm and extensions*. Volume 382. John Wiley & Sons.

McLachlan, G. J. and Peel, D. (1998) Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, volume 1451, pp. 658–666.

McNicholas, P. D. (2016) *Mixture model-based classification*. Chapman and Hall/CRC.

McNicholas, P. D. and Murphy, T. B. (2010) Model-based clustering of longitudinal data. *Canadian Journal of Statistics* **38**(1), 153–168.

Meilă, M. (2007) Comparing clusterings - an information based distance. *Journal of multivariate analysis* **98**(5), 873–895.

Meila, M. (2016) Criteria for comparing clusterings. In *Handbook of Cluster Analysis*, pp. 619–635. CRC Press.

Menardi, G. (2016) A review on modal clustering. *International Statistical Review* **84**(3), 413–433.

Menardi, G. and Azzalini, A. (2014) An advancement in clustering via nonparametric density estimation. *Statistics and Computing* **24**(5), 753–767.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**(1-2), 91–118.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2019) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-139.

Qiao, W. (2018) Asymptotics and optimal bandwidth selection for nonparametric estimation of density level sets. *arXiv preprint arXiv:1707.09697* .

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Li, X. (1998) Curve registration. *Journal of the Royal Statistical Society: Series B (Methodological)* **60**(2), 351–363.

Rice, J. A. (2004) Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* pp. 631–647.

Rigollet, P. and Tsybakov, A. (2007) Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics* **16**(3), 260–280.

Romano, J. P. (1988) On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics* **16**(2), 629–647.

Russell, N., Murphy, T. B. and Raftery, A. E. (2015) Bayesian model averaging in model-based clustering and density estimation. *arXiv preprint arXiv:1506.09035* .

Saavedra-Nieves, P., González-Manteiga, W. and Rodr'ıguez-Casal, A. (2014) Level set estimation. In *Topics in Nonparametric Statistics (M.G.Akritas, S.N. Lahiri and D.N.Politis)*, pp. 299–307.

Samworth, R. J. and Wand, M. P. (2010) Asymptotics and optimal bandwidth selection for highest density region estimation. *The Annals of Statistics* **38**(3), 1767–1792.

Schwarz, G. (1978) Estimating the dimension of a model. *The annals of statistics* **6**(2), 461–464.

Scrucca, L. (2016) Identifying connected components in gaussian finite mixture models for clustering. *Computational Statistics & Data Analysis* **93**, 5–17.

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**(1), 205–233.

Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)* **43**(1), 97–99.

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

Singh, R. S. (1987) Mise of kernel estimates of a density and its derivatives. *Statistics & probability letters* **5**(2), 153–159.

Smyth, P. and Wolpert, D. (1999) Linearly combining density estimators via stacking. *Machine Learning* **36**(1-2), 59–83.

Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N. and Brinkman, R. R. (2012) Flowrepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry Part A* **81**(9), 727–731.

Strehl, A. and Ghosh, J. (2002) Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* **3**, 583–617.

Stuetzle, W. (2003) Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of classification* **20**(1), 025–047.

Stuetzle, W. and Nugent, R. (2010) A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics* **19**(2), 397–418.

Telesca, D., Erosheva, E., Kreager, D. A. and Matsueda, R. L. (2012) Modeling criminal careers as departures from a unimodal population age–crime curve: the case of marijuana use. *Journal of the American Statistical Association* **107**(500), 1427–1440.

Telesca, D. and Inoue, L. Y. T. (2008) Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103**(481), 328–339.

Thom, R. (1949) Sur une partition en cellules associée à une fonction sur une variété. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences* **228**, 973–975.

Tibshirani, R., Wainwright, M. and Hastie, T. (2015) *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall.

Viroli, C. (2011a) Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing* **21**(4), 511–522.

Viroli, C. (2011b) Model based clustering for three-way data structures. *Bayesian Analysis* **6**(4), 573–602.

Viroli, C. and McLachlan, G. (2019) Deep gaussian mixture models. *Statistics and Computing* **29**(1), 43–51.

Von Luxburg, U. (2010) Clustering stability: an overview. *Foundations and Trends in Machine Learning* **2**(3), 235–274.

Wand, M. P. and Jones, M. C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association* **88**(422), 520–528.

Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing*. Chapman and Hall.

Wang, K., Ng, A. and McLachlan., G. (2018) *EMMIXskew: The EM Algorithm and Skew Mixture Distribution*. R package version 1.0.3.

Wei, Y. and McNicholas, P. D. (2015) Mixture model averaging for clustering. *Advances in Data Analysis and Classification* **9**(2), 197–217.

Wishart, D. (1969) Mode analysis: a generalization of nearest neighbour which reduces chaining effects (with discussion). *Numerical taxonomy* pp. 282–311.

Wolfe, J. H. (1963) *Object cluster analysis of social areas*. Ph.D. thesis, University of California.

Wyse, J. and Friel, N. (2012) Block clustering with collapsed latent block models. *Statistics and Computing* **22**(2), 415–428.

Wyse, J., Friel, N. and Latouche, P. (2017) Inferring structure in bipartite networks using the latent blockmodel and exact icl. *Network Science* **5**(1), 45–69.

Alessandro Casa

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4111
e-mail: casa@stat.unipd.it

Current Position

Since October 2016; (expected completion: September 2019)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Climbing modes and exploring mixtures: a journey in density-based clustering
Supervisor: Prof. Giovanna Menardi
Co-supervisor: Prof. José E. Chacón.

Research interests

- Density-based clustering
- Density estimation
- Time-dependent data
- Nonparametric statistics

Education

October 2014 – September 2016
Master degree (*laurea magistrale***) in Statistical Sciences** .
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Semi-supervised detection of collective anomalies using nonparametric methods with an application to particle physics"
Supervisor: Prof. Giovanna Menardi
Final mark: 110/110 cum laude

October 2011 – July 2014
Bachelor degree (*laurea triennale***) in Statistics, Economics and Finance**.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Dimension reduction for model-based clustering: a simulation study and an application to genetic data"
Supervisor: Prof. Giovanna Menardi
Final mark: 110/110 cum laude.

Visiting periods

October 2018 – December 2018
University of Cote d'Azur,
Nice, France.
Supervisor: Prof. Charles Bouveyron

April 2018
University of Perugia,
Perugia, Italy .
Supervisor: Prof. Luca Scrucca

October 2017 – December 2017
University of Cambridge,
Cambridge, United Kingdom.
Supervisor: Prof. Richard Samworth

Computer skills

- R (advanced), Stata (basic), SAS (basic)
- LaTeX, Microsoft Office tools

Language skills

Italian: native; English: fluent

Publications

Articles in journals
Casa, A. (2018). On the choice of the weight function for the integrated likelihood. *SM Journal of Biometrics & Biostatistics* **3**(3), 1033.

Chapters in books
Cabassi, A., Casa, A., Farcomeni, A., Fontana, M., Russo, M. (2018). Three testing perspectives on connectome data. In *Studies in Neural Data Science*, eds. Springer Volume "Proceedings in Mathematics & Statistics", ISBN-9783030000394.

Working papers
Casa, A., Chacón, J.E., Menardi, G. (2019). Modal clustering asymptotics with applications to bandwidth selection. *arXiv preprint* arXiv:1901.07300.

Casa, A., Menardi, G. (2019). Nonparametric semi-supervised classification with application to signal detection in high energy physics. *arXiv preprint* arXiv:1809.02977.

Conference proceedings
Casa, A., Chacón, J.E., Menardi, G. (2019). Asymptotics for bandwidth selection in nonparametric clustering. *Book of short Papers CLADAG-2019*.

Casa, A., Menardi, G. (2019). Nonparametric semisupervised classification and variable selection

for new physics searches. *EMS 2019 - Program and Book of Abstracts.*

Pascali, G., Casa, A., Menardi G. (2019). Co-clustering TripAdvisor data for personalized reccomendations.*Book of short Papers SIS 2019*. ISBN-978889191510.

Casa, A., Scrucca, L., Menardi, G. (2018). Averaging via stacking in model-based clustering. *Book of abstracts of the 4th International Workshop on Model-Based Clustering and Classification (MBC2).*

Casa, A., Scrucca, L., Menardi, G. (2018). On the selection uncertainty in parametric clustering. *Book of abstracts of the European Conference on Data Analysis (ECDA).*

Casa, A., Chacón, J.E., Menardi, G. (2018). On the choice of an appropriate bandwidth for modal clustering. *Book of short Papers SIS 2018*. ISBN-9788891910233.

Casa, A., Chacón, J.E., Menardi, G. (2018). Clustering-oriented selction of the amount of smoothing in kernel density estimation. *Book of abstracts - ISNPS 2018.* ISBN: 978-88-61970- 00-7.

Casa, A., Menardi, G. (2017). Signal detection in high energy physics via a semisupervised nonparametric approach. *Proceedings of the Conference of the Italian Statistical Society "Statistics and Data Sciences: new challenges, new generations".* ISBN: 978-88-6453-521-0.

Casa, A., Menardi, G. (2017). Nonparametric semi-supervised classification with application to signal detection in high energy physics. *Books of abstracts of the International Federation of Classification Societies (IFCS).*

Conference presentations

Casa, A., Chacón, J.E., Menardi, G. (2019). Asymptotics for bandwidth selection in nonparametric clustering. (Invited talk) *12th Scientific Meeting - Classification and Data Analysis Group*, Cassino, Italy, September 2019.

Casa, A., Menardi, G. (2019). Nonparametric semisupervised classification and variable selection for new physics searches. (Invited talk) *European Meeting of Statisticians*, Palermo, Italy, July 2019.

Casa, A., Bouveyron, C., Erosheva, E., Menardi, G. (2019). Co-clustering of time-dependent data. (Poster) *Working group on Model-based Clustering*, Wien, Austria, July 2019.

Casa, A., Scrucca, L., Menardi, G. (2018). Averaging via stacking in model-based clustering. (Poster) *Workshop on Advanced Statistics for Physics Discovery*, Padova, Italy, September 2018.

Casa, A., Scrucca, L., Menardi, G. (2018). Averaging via stacking in model-based clustering. (Poster) *Workshop on Model-Based Clustering and Classification*, Catania, Italy, September 2018.

Casa, A., Scrucca, L., Menardi, G. (2018). On the selection uncertainty in parametric clustering. (Contributed talk) *European Conference on Data Analysis*, Paderborn, Germany, July 2018.

Casa, A., Chacón, J.E., Menardi, G. (2018). On the choice of an appropriate bandwidth for modal clustering. (Contributed talk) *SIS - 49th Scientific meeting of the Italian Statistical Society*, Palermo, Italy, June 2018.

Casa, A., Chacón, J.E., Menardi, G. (2018). Clustering-oriented selction of the amount of smoothing in kernel density estimation. (Contributed talk) *ISNPS - 4th Conference of the International Society for Nonparametric Statistics*, Salerno, Italy, June 2018.

Casa, A., Menardi, G. (2017). Signal detection in high energy physics via a semisupervised nonparametric approach. (Contributed talk) *SIS - Intermediate conference of the Italian Statistical Society*, Florence, Italy, June 2017.

Teaching experience

Bike sharing in Paris: a case study. Specialist lecture during the class Statistica Iterazione (Master degree, Academic year 2018/2019)

References

Prof. Giovanna Menardi
Department of Statistics
University of Padova
via Cesare Battisti, 241-243
35121 Padova, Italy
Phone: +39 049 827 4119
e-mail: menardi@stat.unipd.it

Prof. Luca Scrucca
Department of Economics,
University of Perugia
via Alessandro Pascoli, 20
06123 Perugia, Italy
Phone: +39 075 5855229
e-mail: luca.scrucca@unipg.it

Prof. José E. Chacón
Department of Mathematics
University of Extremadura
Avenida de Elvas
06006 Badajoz
Phone: +34 927289300
e-mail: jechacon@unex.es