Electronic Journal of Statistics
Vol. 9 (2015) 266–297
ISSN: 1935-7524
DOI: 10.1214/15-EJS995

On clustering procedures and nonparametric mixture estimation

Stéphane Auray

CREST-Ensai and EQUIPPE (EA4018), ULCO
e-mail: stephane.auray@ensai.fr

Nicolas Klutchnikoff

CREST-Ensai, IRMA (UMR CNRS 7501) and Université de Strasbourg
e-mail: nicolas.klutchnikoff@ensai.fr

and

Laurent Rouvière

CREST-Ensai, IRMAR (UMR CNRS 6625) and UEB
e-mail: laurent.rouviere@ensai.fr

Abstract: This paper deals with nonparametric estimation of conditional densities in mixture models in the case when additional covariates are available. The proposed approach consists of performing a preliminary clustering algorithm on the additional covariates to guess the mixture component of each observation. Conditional densities of the mixture model are then estimated using kernel density estimates applied separately to each cluster. We investigate the expected L_1-error of the resulting estimates and derive optimal rates of convergence over classical nonparametric density classes provided the clustering method is accurate. Performances of clustering algorithms are measured by the *maximal misclassification error*. We obtain upper bounds of this quantity for a single linkage hierarchical clustering algorithm. Lastly, applications of the proposed method to mixture models involving electricity distribution data and simulated data are presented.

MSC 2010 subject classifications: Primary 62G07; secondary 62H30.
Keywords and phrases: Nonparametric estimation, mixture models, clustering.

Received June 2014.

Contents

1. Introduction

Finite mixture models are widely used to account for population heterogeneities. In many fields such as biology, econometrics and social sciences, experiments are based on the analysis of a variable characterized by a different behavior depending on the group of individuals. A natural way to model heterogeneity for a real random variable Y is to use a mixture model. In this case, the density f of Y can be written as

$$f(t) = \sum_{i=1}^{M} \alpha_i f_i(t), \quad t \in \mathbb{R}. \tag{1}$$

Here M is the number of subpopulations, α_i and f_i are respectively the mixture proportion and the probability density function of the i^{th} subpopulation. We refer the reader to Everit and Hand (1981), McLachlan and Basford (1988) and McLachlan and Peel (2000) for a broader picture of mixture density models as well as for practical applications.

When dealing with mixture density models such as (1), some issues arise. In some cases, the number of components M is unknown and needs to be estimated. To this end, some algorithms have been developed to provide consistent estimates of this parameter. For instance, when M corresponds to the number of modes of f, Cuevas, Febrero and Fraiman (2000) and Biau, Cadre and Pelletier (2007) propose an estimator based on the level sets of f. Model identifiability is an additional issue that has received some attention in the literature. Actually, model (1) is identifiable only by imposing restrictions on the vector $(\alpha_1, \ldots, \alpha_M, f_1, \ldots, f_M)$. In order to provide the minimal assumptions such that (1) becomes identifiable, Celeux and Govaert (1995) and Bordes, Mottelet

and Vandekerkhove (2006) (see also the references therein) assume that the density functions f_i's belong to some parametric or semi-parametric density families. However, in a nonparametric setting, it turns out that identifiability conditions are more difficult to provide. Hall and Zhou (2003) define mild regularity conditions to achieve identifiability in a multivariate nonparametric setting while Kitamura (2004) considers the case where appropriate covariates are available.

When the model (1) is identifiable, the statistical problem consists of estimating mixture proportions α_i and density functions f_i. In the parametric case, some algorithms have been proposed such as maximum likelihood techniques (Lindsay (1983a,b) and Redner and Walker (1984)) as well as Bayesian approaches (Diebolt and Robert (1994) and Biernacki, Celeux and Govaert (2000)). When the f_i's belong to nonparametric families, it is often assumed that training data are observed, *i.e.*, the component of the mixture from which Y is distributed is available. In that case, the model is identifiable and some algorithms allow to estimate both the α_i's and the f_i's (see Titterington (1983), Hall and Titterington (1984, 1985) and Cerrito (1992)). However, as pointed out by Hall and Zhou (2003), inference in mixture nonparametric density models becomes more difficult without training data. These authors introduce consistent nonparametric estimators of the conditional distributions in a multivariate setting. We also refer to Bordes, Mottelet and Vandekerkhove (2006) who provide efficient estimators under the assumption that the unknown mixed distribution is symmetric. These estimates are extended by Benaglia, Chauveau and Hunter (2009, 2011) for multivariate mixture models.

The framework we consider takes place between the two above situations. More precisely, training data are not observed but we assume to have at hand some covariates that may provide information on the components of the mixture from which Y is distributed. Our approach consists of performing a preliminary clustering algorithm on these covariates to guess the mixture component of each observation. Density functions f_i are then estimated using a nonparametric density estimate based on the predictions of the clustering method.

Many authors have already proposed to carry out a preliminary clustering step to improve density estimates in mixture models. Ruzgas, Rudzkis and Kavaliauskas (2006) conduct a comprehensive simulation study to conclude that a preliminary clustering using the EM algorithm allows to some extent to improve performances of some density estimates (see also Jeon and Landgrebe (1994)). However, to our knowledge, no work has been devoted so far to measure the effects of the clustering algorithm on the resulting estimates of the distribution functions f_i. This paper proposes to fill this gap, studying the L_1-error of these estimates. To do so, we measure the performance of clustering methods by the maximal misclassification error (4). This criterion allows us to derive optimal rates of convergence over classical nonparametric density classes, provided the clustering method used in the first step performs well with respect to this notion.

The paper is organized as follows. In Section 2, we present the two-step estimator and give the main results. Examples of clustering algorithms are worked

out in Section 3. In particular, the maximal misclassification error of a hierarchical clustering algorithm is studied under mild assumptions on the model. Applications on simulated and real data are presented in Sections 4 and 5. A short conclusion including a discussion of the implications of the work is given in Section 6 and proofs are gathered in Section 7.

2. A two-step nonparametric estimator

2.1. The statistical problem

Our focus is on the estimation of conditional densities in a univariate mixture density model. Formally we let (Y, I) be a random vector taking values in $\mathbb{R} \times [\![1, M]\!]$ where $M \geq 2$ is a known integer. We assume that the distribution of Y is characterized by a density f defined, for all $t \in \mathbb{R}$, by

$$f(t) = \sum_{i=1}^{M} \alpha_i f_i(t),$$

where, for all $i \in [\![1, M]\!]$, $\alpha_i = \mathbb{P}(I = i)$ are the prior probabilities (or the weights of the mixture) and f_i are the densities of the conditional distributions $\mathcal{L}(Y|I = i)$ (or the components of the mixture).

If we have at hand n observations $(Y_1, I_1), \ldots, (Y_n, I_n)$ drawn from the distribution of (Y, I), one can easily find efficient estimates for both the α_i's and the f_i's. For example, if we denote $N_i = \#\{k \in [\![1, n]\!] : I_k = i\}$, then we can estimate α_i using the empirical proportion $\bar{\alpha}_i = N_i/n$ and f_i by the kernel density estimate $\bar{f}_i$ defined for all $t \in \mathbb{R}$ by

$$\bar{f}_i(t) = \frac{1}{N_i} \sum_{k=1}^{n} K_h(t, Y_k) \mathbb{I}_i(I_k) \tag{2}$$

if $N_i > 0$. For the definiteness of $\bar{f}_i$ we conventionally set $\bar{f}_i(t) = 0$ if $N_i = 0$. Here K is a kernel which belongs to $L_1(\mathbb{R}, \mathbb{R})$ and such that $\int K = 1$, $h > 0$ is a bandwidth and

$$K_h(t, y) = \frac{1}{h} K\left(\frac{t - y}{h}\right) \tag{3}$$

is the classical convolution kernel located at point t (see Rosenblatt (1956) and Parzen (1962) for instance). Estimate (2) is just the usual kernel density estimate defined from observations in the i^{th} subpopulation. It follows that, under classical assumptions regarding the smoothing parameter h and the kernel K, $\bar{f}_i$ has similar properties as those of the well-known kernel density estimate. In particular, the expected L_1-error

$$\mathbb{E}\|\bar{f}_i - f_i\|_1 = \mathbb{E} \int_{\mathbb{R}} |\bar{f}_i(t) - f_i(t)| dt$$

achieves optimal rates when f_i belongs to regular density classes such as Hölder or Lipschitz classes (see Devroye and Györfi (1985)).

The problem is more complicated when the random variable I is not observed. In this situation, $\bar{\alpha}_i$ and $\bar{f}_i$ are not computable and one has to find another way to define efficient estimates for both α_i and f_i. In this work, we assume that one can obtain information on I through another covariate X which takes values in $\mathbb{R}^d$ where $d \geq 1$. This random variable is observed and its conditional distribution $\mathcal{L}(X|I = i)$ is characterized by a density $g_i = g_{i,n} : \mathbb{R}^d \to \mathbb{R}$ which could depend on n. In this framework, the statistical problem is to estimate both the components and the weights of the mixture model (1) using the n-sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ extracted from $(Y_1, X_1, I_1), \ldots, (Y_n, X_n, I_n)$ randomly drawn from the distribution of (Y, X, I).

2.2. Discussion on the model

Estimating components of a mixture model is a classical statistical problem. The new feature proposed here is to include covariates in the model which can potentially improve traditional algorithms. These covariates are represented by a random vector X which provides information on the unobserved group I. This model includes many practical situations. Three examples are provided in this section.

The classical mixture problem without covariates. A traditional problem in mixture models is the estimation of the components $f_i, i \in [\![1, M]\!]$ in (1) from (only) an i.i.d sample $Y_1, \ldots, Y_n$ drawn from f: no covariates are available. In this context, many parametric methods such as the EM algorithm (and its derivatives) as well as nonparametric procedures (under suitable identifiability constraints) can be used and are widely studied. Even if this model is formally a particular case of ours (we just have to take $X = Y$), the approach presented in this paper is not designed to be competitive in this situation with dedicated parametric or nonparametric methods. Indeed, our model focus on practical situations where covariates can be used to obtain useful information about the hidden variable I. Below, we offer two realistic situations where such covariates are naturally available.

Medical example. Many diseases evolve over time and exhibit different stages of development which can be represented by a variable I that takes a finite number of values. In many situations, the problem is not to study the stage I but some variables that can potentially have different behavior according to I. For instance, the survival time Y and its conditional distributions with respect to I are typically of interest in many situations. In practice, the stage I is generally not observed. It is assessed by the medical team from several items such as physiological data, medical examinations, interviews with the patient (and so on) that can be represented by covariates X in our model.

Electricity distribution. A distribution network may locally experience minor problems, due for example to bad weather, that may affect some customers

during a fixed period of time in a given geographical area. To better understand the origin and/or consequences of the dysfunctions, and thus better forecast network operations, electricity distributors are interested in the distribution behavior of several quantities Y for two different groups of customers: those affected by the malfunction and the others. Variables Y may for instance represent averages or variations of consumption after the disruption period. In this situation the group is represented by a variable I: $I = 1$ for the users affected by the disruption and 2 for the others. This binary variable I is not directly observed but it can be guessed from individuals curves of consumptions during the disruption period. In our framework, discrete versions of these curves correspond to the covariate X. This example is explained in-depth and analyzed in Section 5 using real data from ERDF, the main French distributor of electricity.

2.3. A kernel density estimate based on a clustering approach

To estimate densities f_i of the conditional distributions $\mathcal{L}(Y|I = i), i \in [\![1, M]\!]$, we propose a two-step algorithm that can be summarized as follows.

1. Apply a clustering algorithm on the sample $X_1, \ldots, X_n$ to predict the label I_k of each observation X_k;
2. Estimate conditional densities f_i by kernel density estimates (2) where unobserved labels are substituted by predicted labels.

Formally, we first perform a given clustering algorithm to split the sample $X_1, \ldots, X_n$ into $M + 1$ clusters $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_M$ such that $\mathcal{X}_i \neq \emptyset$ for all $i \in [\![1, M]\!]$. Clusters $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_M$ satisfy

$$\bigcup_{i=0}^{M} \mathcal{X}_i = \{X_1, \ldots, X_n\} \quad \text{and} \quad \forall i \neq j,\ \mathcal{X}_i \cap \mathcal{X}_j = \emptyset.$$

We do not specify the clustering method here, some examples are discussed in Sections 3 and 4. Observe that we define $M + 1$ clusters instead of M. The cluster $\mathcal{X}_0$ (which could be empty) contains the observations for which the clustering procedure is not able to predict the label. For example, if the clustering procedure reveals some outliers, they are collected in $\mathcal{X}_0$ and we do not use these outliers to estimate the f_i's.

Once the clustering step is performed, we define the predicted labels $\widehat{I}_k$ as

$$\widehat{I}_k = i \quad \text{if} \quad X_k \in \mathcal{X}_i, \qquad k \in [\![1, n]\!], \quad i \in [\![1, M]\!].$$

Observation X_k is not correctly assigned to its group with probability $\mathbb{P}(\widehat{I}_k \neq I_k)$. We measure the performance of the clustering algorithm by the maximal probability to not correctly attribute an observation:

$$\varphi_n = \max_{1 \leq k \leq n} \mathbb{P}(\widehat{I}_k \neq I_k). \tag{4}$$

We call this error term the *maximal misclassification error*. It will be studied for two clustering algorithms in Section 3.

To define our estimates, we just replace in (2) the true labels I_k by the predicted labels $\widehat{I}_k$. Formally, prior probabilities α_i are estimated by

$$\widehat{\alpha}_i = \frac{\widehat{N}_i}{n} \quad \text{where} \quad \widehat{N}_i = \#\{k \in [\![1, n]\!] : \widehat{I}_k = i\},$$

while for the conditional densities f_i, we consider the kernel density estimator with kernel $K : \mathbb{R} \to \mathbb{R}$ and bandwidth $h > 0$

$$\widehat{f}_i(t) = \frac{1}{\widehat{N}_i} \sum_{k : X_k \in \mathcal{X}_i} K_h(t, Y_k) = \frac{1}{\widehat{N}_i} \sum_{k=1}^{n} K_h(t, Y_k) \mathbb{I}_{\{i\}}(\widehat{I}_k), \tag{5}$$

where K_h is defined in (3). Observe that since for all $i \in [\![1, M]\!]$ the clusters $\mathcal{X}_i$ are nonempty, the estimates $\widehat{f}_i$ are well defined.

Kernel estimates $\widehat{f}_i$ are defined from observations in cluster $\mathcal{X}_i$. The underlying assumption is that, for all $i \in [\![1, M]\!]$, each cluster $\mathcal{X}_i$ collects almost all of the observations X_k such that Y_k is randomly drawn from f_i. Under this assumption, φ_n is expected to be small and $\widehat{f}_i$ to be closed to the oracle estimates $\bar{f}_i$ defined by equation (2). This closeness is measured in the following theorem which makes the connection between the expected L_1-errors of $\bar{f}_i$ and f_i.

Theorem 2.1. *There exist positive constants $A_1 - A_3$ such that, for all $n \geq 1$ and $i \in [\![1, M]\!]$*

$$\mathbb{E}\big\|\widehat{f}_i - f_i\big\|_1 \leq \mathbb{E}\big\|\bar{f}_i - f_i\big\|_1 + A_1 \varphi_n + A_2 \exp(-n) \tag{6}$$

and

$$\mathbb{E}|\widehat{\alpha}_i - \alpha_i| \leq \varphi_n + \frac{A_3}{\sqrt{n}}. \tag{7}$$

Constants $A_1 - A_3$ are specified in the proof of the theorem. We emphasize that inequalities (6) and (7) are non-asymptotic, that is, the bounds are valid for all n. If we intend to prove any consistency results regarding $\widehat{f}_i$ and $\widehat{\alpha}_i$, inequality (6) says that the maximal misclassification error φ_n should tend to zero. Moreover, if φ_n tends to zero much faster than the L_1-error of $\bar{f}_i$, then the asymptotic performance is guaranteed to be equivalent to the one of the oracle estimate $\bar{f}_i$. The L_1-error of $\bar{f}_i$, with properly chosen bandwidth h and kernel K, is known to go to zero, under standard smoothness assumptions, at rate $n^{-\frac{s}{2s+1}}$ where $s > 0$ is typically an index representing the regularity of f_i. For example, when we consider Lipschitz or Hölder classes of functions with compact supports, s corresponds to the number of absolutely continuous derivatives of the functions f_i. In this context, if $\varphi_n = \mathcal{O}(n^{-\frac{s}{2s+1}})$, then

$$\mathbb{E}\big\|\widehat{f}_i - f_i\big\|_1 = \mathcal{O}(n^{-\frac{s}{2s+1}}).$$

Remark 2.1. Note that even if clusters $\mathcal{X}_1, \ldots, \mathcal{X}_M$ are arbitrarily indexed, inequalities (6) and (7) are true whatever the choice of the indexes. However, when indexes are not chosen according to the true labels, φ_n could be large even if the clustering procedure performs well. In this situation there exists a permutation

of the indexes such that, after this permutation, the maximal misclassification error is small. More precisely it can be readily seen, using Theorem 2.1, that

$$\min_{\pi\in\Pi_M} \mathbb{E}\big\|\widehat{f}_{\pi(i)} - f_i\big\|_1 \le \mathbb{E}\big\|\bar{f}_i - f_i\big\|_1 + A_1 \min_{\pi\in\Pi_M} \varphi_n(\pi) + A_2\exp(-n) \tag{8}$$

where Π_M denotes the set of all permutations of $[\![1, M]\!]$ and $\varphi_n(\pi)$ is the maximal misclassification error of the clustering method after the permutation of the indexes:

$$\varphi_n(\pi) = \max_{k=1,\dots,n} \mathbb{P}(\pi(\widehat{I}_k) \neq I_k), \quad \pi \in \Pi_M. \tag{9}$$

Remark 2.2. As usual, the choice of the bandwidth h reveals crucial for the performance of the kernel density estimates. However, this paper does not provide any theory to select this parameter. If automatic or adaptive procedures are needed, they can be obtained by adjusting traditional automatic selection procedures for classical nonparametric estimators (see for example Berlinet and Devroye (1994) or Devroye and Lugosi (2001)).

3. Clustering procedures

The proposed procedure requires a preliminary clustering algorithm performed on the sample $X_1, \dots, X_n$. Even if any clustering algorithm could be applied in practice, it should be chosen according to the conditional distributions $\mathcal{L}(X|I = i), i \in [\![1, M]\!]$. More precisely, each cluster should match up with observations drawn from one of those conditional distributions. From a theoretical point of view, for a given clustering procedure, the problem is to find upper bounds for the maximal misclassification error φ_n to apply Theorem 2.1. In a parametric setting, *i.e.*, when conditional distributions are identified by unknown parameters, clustering algorithms are often based on efficient estimators of these unknown parameters. We provide an example in Section 3.1. Without parametric assumptions on the distribution, the problem is more complicated. Contrary to data analysis methods such as regression or classification, there are many ways to define clustering. One of the most popular approach consists of defining clusters as the connected components of the level sets of the density (see Hartigan (1975)). This amounts to saying that clusters represent high density regions of the data separated by low density regions. In this context, many authors have studied theoretical performances of clustering algorithms based on neighborhood graphs such as hierarchical or spectral clustering algorithms. In Section 3.2, we extend results of Maier, Hein and Von Luxburg (2009) and Arias-Castro (2011) to our framework for a hierarchical clustering algorithm based on pairwise distances. This procedure is challenged with other clustering methods in the simulation part.

3.1. A parametric example

We consider a mixture of two uniform univariate densities

$$g_{1,n}(x) = g_1(x) = \mathbb{I}_{[0,1]}(x) \quad \text{and} \quad g_{2,n}(x) = \mathbb{I}_{[1-\lambda_n, 2-\lambda_n]}(x),$$

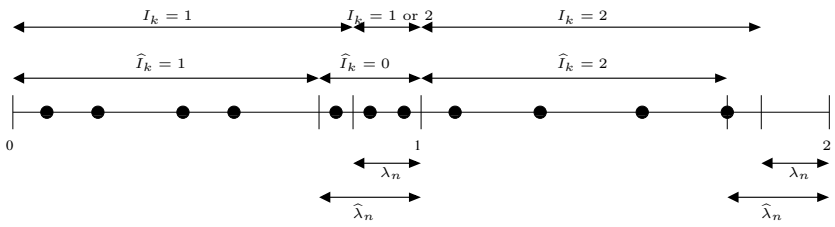

FIG 1. *A sample of* $n = 11$ *points.*

where we recall that $g_{i,n}$ is the density of the conditional distribution $\mathcal{L}(X|I = i), i = 1, 2$. Here $(\lambda_n)_n$ is a non-increasing sequence which tends to 0 as n goes to infinity. In this parametric situation, a natural way to guess the unobserved label I_k of the observation X_k is to find an estimator $\widehat{\lambda}_n$ of λ_n and to predict the labels (see Figure 1) according to

$$\widehat{I}_k = \begin{cases} 1 & \text{if} \quad X_k \leq 1 - \widehat{\lambda}_n \\ 0 & \text{if} \quad 1 - \widehat{\lambda}_n < X_k < 1 \\ 2 & \text{if} \quad X_k \geq 1. \end{cases} \tag{10}$$

The accuracy of these predictions depends on the choice of the estimator $\widehat{\lambda}_n$. Here we choose $\widehat{\lambda}_n = 2 - X_{(n)}$ where $X_{(n)} = \max_{1 \leq k \leq n} X_k$. Note that in this situation, we have for $i = 1, 2$

$$\widehat{I}_k = i \Longrightarrow I_k = i, \text{ a.s.}$$

It means that all classified observations (with non-zero estimated label) are well-classified and that misclassified observations are collected in $\mathcal{X}_0$ (see Figure 1).

The following proposition establishes a performance bound for the maximal misclassification error φ_n of this clustering procedure.

Proposition 3.1. *There exists a positive constant* A_4 *such that for all* $n \geq 1$

$$\varphi_n \leq \lambda_n + A_4 \frac{\log n}{n}.$$

Unsurprisingly, φ_n decreases as λ_n decreases. Moreover, since in most cases of interest, the expected L_1-error of $\bar{f}_i$ tends to zero much slower than $1/\sqrt{n}$, this property means that, asymptotically, the expected L_1-error of $\widehat{f}_i$ is of the same order as the expected L_1-error of $\bar{f}_i$ provided $\lambda_n = \mathcal{O}(1/\sqrt{n})$ (see (6)).

3.2. A hierarchical clustering algorithm

Assuming that clusters are defined as connected components of level sets of a density, many authors have studied theoretical properties of various clustering algorithms. For instance, Maier, Hein and Von Luxburg (2009) and Arias-Castro

(2011) prove that algorithms based on pairwise distances (k-nearest neighbor graph, spectral clustering...) are efficient as soon as these connected components are separated enough. In this section, we extend results of these authors to bound the maximal misclassification error φ_n for a hierarchical clustering algorithm.

3.2.1. The clustering algorithm

Given $X_1, \dots, X_n$, we consider a single linkage hierarchical clustering algorithm based on pairwise distances to extract exactly M disjoint clusters $\mathcal{X}_1, \dots, \mathcal{X}_M$ from the observations (see Arias-Castro (2011)). This algorithm consists of finding a data-driven radius $\widehat{r}_n > 0$ such that the set

$$\bigcup_{k=1}^{n} B(X_k, \widehat{r}_n) \tag{11}$$

has exactly M connected components. Here $B(x, r)$ stands for the closed Euclidean ball with center $x \in \mathbb{R}^d$ and radius $r > 0$. Cluster $\mathcal{X}_i$ is then naturally composed by observations X_k which belong to the i^{th} connected component of the set (11).

The radius $\widehat{r}_n$ can be defined in a formal way to derive statistical properties of the clustering procedure. To this end, we define for each positive real number r the $n \times n$ affinity matrix $A^r = (A^r_{k,\ell})_{1 \le k,\ell \le n}$ by

$$A^r_{k,\ell} = \begin{cases} 1 & \text{if } \|X_k - X_\ell\|_2 \le 2r \iff B(X_k, r) \cap B(X_\ell, r) \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $\|x\|_2$ stands for the Euclidean norm of $x \in \mathbb{R}^d$. This matrix induces a non-orientated graph on the set $[\![1, n]\!]$ and two different observations X_k and X_ℓ belong to the same cluster if k and ℓ belong to the same connected component of the graph. We let $\widehat{M}_r$ be the number of connected components of the graph and we denote by $\mathcal{X}_1(r), \dots, \mathcal{X}_{\widehat{M}_r}(r)$ the associated clusters. The radius is selected as follows

$$\widehat{r}_n = \inf\{r > 0 : \widehat{M}_r \le M\}.$$

Note that $\widehat{r}_n$ is well-defined since the random set $\mathcal{R}_M = \{r > 0 : \widehat{M}_r \le M\}$ is lower bounded (by 0) and non-empty since $r^* = \max_{k,\ell} \|X_k - X_\ell\|_2$ always belongs to this set ($\widehat{M}_{r^*} = 1$). Moreover, since $r \mapsto \widehat{M}_r$ is non-increasing and right-continuous, one can easily prove that $\widehat{r}_n = \min \mathcal{R}_M$ and $\widehat{M}_{\widehat{r}_n} = M$ almost surely when $n \ge M$. Let $\mathcal{X}_1(\widehat{r}_n), \dots, \mathcal{X}_M(\widehat{r}_n)$ be the M clusters induced by $A^{\widehat{r}_n}$, the aim is to study the maximal misclassification error (4) of this clustering algorithm.

Remark 3.1. The algorithm requires that the connected components of the graph induced by the $n \times n$ matrix A^r be computed for different values of r. Some algorithms can be performed to obtain these connected components. For instance, we can use the Depth-First search algorithm (see Cormen, Leiserson

and Rivest (1990)) which can be performed efficiently in $\mathcal{O}(V_n + E_n)$ operations, where V_n and E_n denote respectively the number of vertices and edges of the graph.

3.2.2. The clustering model

Recall that the clustering algorithm is performed on the sample $X_1, \ldots, X_n$. To study the maximal misclassification error, some assumptions on the distribution of X are needed.

Assumption 1. Let g_n denotes the probability density of X. We assume that there exists a positive sequence $(t_n)_n$ such that the set

$$\{x \in \mathbb{R}^d : g_n(x) \geq t_n\} \tag{13}$$

has exactly M disjoint connected compact sets $S_{1,n}, \ldots, S_{M,n}$ satisfying, for all $i \in [\![1, M]\!]$,

$$\mathbb{P}(X_1 \in S_{i,n} | I_1 = i) = \int_{S_{i,n}} g_{i,n}(x)\, \mathrm{d}x > 1/2, \tag{14}$$

where we recall that $g_{i,n}$ stands for the density of the conditional distribution $\mathcal{L}(X|I = i), i \in [\![1, M]\!]$. We note $S_n = \bigcup_{i=1}^{M} S_{i,n}$ and

$$\delta_n = \inf_{1 \leq i \neq j \leq M} \mathrm{dist}(S_{i,n}, S_{j,n}),$$

where

$$\mathrm{dist}(S_{i,n}, S_{j,n}) = \inf_{x \in S_{i,n}} \inf_{y \in S_{j,n}} \|x - y\|_2.$$

Assumption 2. There exist two positive constants c_1 and c_2, and a family of $N \in \mathbb{N}^\star$ Euclidean balls $\{B_\ell\}_{\ell=1,\ldots,N}$ with radius $r_n/2$ such that

$$\begin{cases} S_n \subset \bigcup_{\ell=1}^{N} B_\ell \\ \mathrm{Leb}(S_n) \geq c_1 \sum_{\ell=1}^{N} \mathrm{Leb}(S_n \cap B_\ell) \\ \forall \ell = 1, \ldots, N, \quad \mathrm{Leb}(S_n \cap B_\ell) \geq c_2 r_n^d, \end{cases}$$

where Leb denotes the Lebesgue measure on $\mathbb{R}^d$ and r_n is defined by

$$r_n^d = \frac{\tau \log n}{n t_n} \qquad \text{with} \quad \tau > 1/c_2.$$

Assumption 1 is classical to study performances of clustering algorithm (see Maier, Hein and Von Luxburg (2009)) or to estimate the number of clusters (see Biau, Cadre and Pelletier (2007)). It implies that clusters reflect high-density regions separated by low-density regions. Condition (14) is required to be sure that the connected components of (13) are correctly indexed. It makes it possible to avoid that most of the observation in $S_{i,n}$ are drawn from $g_{j,n}$ with $j \neq i$. Assumption 2 is more technical and pertains to the diameter and

regularity of the sets $S_{i,n}$. Our approach consists of identifying sets $S_{i,n}$ with the connected components of $\bigcup_{k=1}^{n} B(X_k, r)$. Thus, when diameter of $S_{i,n}$ increases, large values of radius r are necessary to connect observations in $S_{i,n}$. However for too large values of r, the number of connected components of $\bigcup_{k=1}^{n} B(X_k, r)$ becomes smaller than M and the method fails. Consequently, we need to constraint the diameter of $S_{i,n}$. This is ensured by assumption 2 since it implies that S_n can be covered by N Euclidean balls such that

$$N \leq \frac{n}{c_1 c_2 \tau \log n}. \tag{15}$$

Finally, inequality $\mathrm{Leb}(S_n \cap B_\ell) \geq c_2 r_n^d$ in assumption 2 can be seen as a smoothness assumption on the boundaries of S_n (see Biau, Cadre and Pelletier (2008)).

Remark 3.2. In dimension 1, since each $S_{i,n}$ is connected, it is a segment of the real line. Thus, under assumption 1, its diameter is bounded by $1/t_n$ and assumption 2 is satisfied. For higher dimensions, things turn out to be more complicated. Indeed, even if the measure of the compact set S_n is upper bounded by $1/t_n$, its diameter can be as large as we want. Consider for example the density

$$h_n(x, y) = \mathbb{I}_{[1-1/a_n, a_n]}(x) \mathbb{I}_{[0,1/x^2]}(y), \quad (x, y) \in \mathbb{R}^{+\star} \times \mathbb{R}^{+},$$

where $a_n > 1$. Since a_n could be chosen to be arbitrarily large, the diameter of S_n could also be arbitrarily large and assumption 2 does not hold. This assumption restricts to some extent the shape of S_n. It is satisfied for regular sets such that the diameter does not increase too quickly as n goes to infinity. For example, consider the two dimensional situation where S_n is a rectangle with length u_n and width v_n. In such a scenario, one can easily prove that if there exist two positive constants a_1 and a_2 such that $u_n \geq a_1 r_n$ and $v_n \geq a_2 r_n$, then assumption 2 holds. Note also that this assumption is verified for sets S_n that do not depend on the sample size n with smooth boundaries (see Biau, Cadre and Pelletier (2007) and Maier, Hein and Von Luxburg (2009)).

Remark 3.3. Assumption 1 is clearly satisfied when supports of conditional densities $g_{i,n}$ are disjoint. This assumption could also be verified when these supports overlap. As an example, consider the Laplace mixture model:

$$g_{i,n}(x) = \frac{1}{2\sigma_n} \exp\left(-\frac{|x - \mu_{i,n}|}{\sigma_n}\right), \qquad i = 1, 2,$$

where $\sigma_n > 0$ and $\mu_{i,n} \in \mathbb{R}$ (see Figure 2). Let $\ell_n = |\mu_{1,n} - \mu_{2,n}|$ be the distance between the two location parameters $\mu_{1,n}$ and $\mu_{2,n}$ and define

$$t_{*,n} = \frac{\sqrt{\alpha_1 \alpha_2}}{\sigma_n} \exp\left(-\frac{\ell_n}{2\sigma_n}\right)$$

and

$$t_{i,n}^{*} = \frac{1}{2\sigma_n}\left(\alpha_i + (1 - \alpha_i) \exp\left(-\frac{\ell_n}{\sigma_n}\right)\right), \quad i = 1, 2.$$

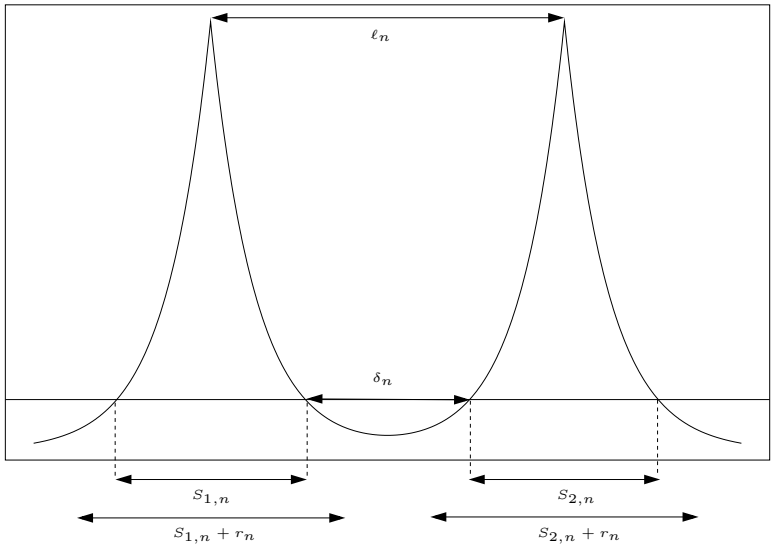

FIG 2. *Connected components of level sets for a mixture of Laplace distributions.*

Then direct calculations yield that for any $t_n \in (t_{*,n}, t^*_{1,n} \wedge t^*_{2,n})$, the level set $\{\alpha_1 g_{1,n} + \alpha_2 g_{2,n} \geq t_n\}$ has exactly $M = 2$ connected components provided $\log(\alpha_1/(1-\alpha_1)) \in (-\ell_n/\sigma_n, \ell_n/\sigma_n)$.

3.2.3. The maximal misclassification error

The algorithm described in Section 3.2.1 provides a partition of $\{X_1, \ldots, X_n\}$ into M clusters $\mathcal{X}_1(\widehat{r}_n), \ldots, \mathcal{X}_M(\widehat{r}_n)$. To apply Theorem 2.1, we have to find an upper bound of the maximal misclassification error for the predicted rule

$$\widehat{I}_k = i \iff X_k \in \mathcal{X}_i(\widehat{r}_n).$$

Observe that, for this clustering algorithm, clusters $\mathcal{X}_1(\widehat{r}_n), \ldots, \mathcal{X}_M(\widehat{r}_n)$ defined in Section 3.2.1 are arbitrarily indexed. Thus there is no guarantee that the predicted labels are correctly indexed. To circumvent this problem, as suggested in Remark 2.1, we study the maximal misclassification error up to a permutation of the indexes.

The proposed clustering algorithm has been studied by Maier, Hein and Von Luxburg (2009) and Arias-Castro (2011). They prove that each cluster corresponds to one of the connected components of (13) with high probability in a model similar to ours. In other words, clusters make it possible to identify each connected components of (13). Even if the identification of these connected components is important in our setting, it is not sufficient since our goal is to find an upper bound of the misclassification error (9). Moreover, since supports of

conditional densities $g_{i,n}$ can overlap, observations in the connected components $S_{i,n}$ of (13) are not guaranteed to emerge from the distribution of $\mathcal{L}(X|I=i)$. This leads us to define

$$\psi_n = \max_{i=1,\dots,M} \mathbb{P}(X_1 \notin (S_{i,n} + r_n)|I_1 = i)$$

where for $S \subset \mathbb{R}^d$ and $r > 0$

$$S + r = \{x \in \mathbb{R}^d : \exists y \in S \text{ such that } \|x - y\|_2 \leq r\}.$$

Observe that ψ_n is the maximal probability that an observation from the i^{th} group does not belong to $S_{i,n}+r_n$. This parameter reflects the degree of difficulty for the model to correctly predict the label of the observations: the larger ψ_n, the more difficult it is. We can now set forth the main result of this section.

Theorem 3.1. *Suppose that Assumption 1 and Assumption 2 hold. Moreover, if*

$$\delta_n > 2r_n = 2\left(\frac{\tau \log n}{nt_n}\right)^{1/d}, \tag{16}$$

then for all $0 < a \leq c_2\tau - 1$*, we have*

$$\min_{\pi\in\Pi_M} \max_{1,\dots,n} \mathbb{P}(\pi(\widehat{I}_k) \neq I_k) \leq \frac{A_5}{n^a \log n} + (n+2)\psi_n, \tag{17}$$

where A_5 *is positive constant.*

This theorem provides minimal assumptions to make accurate predictions of the labels I_k. Inequality (16) gives the minimum distance between the connected components $S_{i,n}$ to make the clustering method efficient. When supports of the conditional densities $g_{i,n}$ are disjoints, it is easily seen that $\psi_n = 0$ and $\widehat{I}_k = I_k$ almost surely for n large enough provided inequality (16) is satisfied. When the supports overlap, inequality (17) ensures that the algorithm performs well provided the probability ψ_n tends to zero much faster than $1/n$. In the Laplace example presented in Remark 3.3, it can be easily seen that

$$\psi_n = \mathcal{O}\left(\exp\left(-\frac{\ell_n}{2\sigma_n}\right)\right).$$

It implies that as soon as $\ell_n/\sigma_n \geq 3\log(n)/2$, $n\psi(n) \leq n^{-1/2}$ and the kernel density estimates defined in (5) satisfy

$$\min_{\pi\in\Pi_M} \mathbb{E}\|\widehat{f}_{\pi(i)} - f_i\| \leq \mathbb{E}\|\bar{f}_i - f_i\| + \frac{A_6}{\sqrt{n}}.$$

Finally, note that when $\psi_n = 0$, inequality (17) implies that each cluster $\mathcal{X}_i(\widehat{r}_n)$ belong to one of the connected components of (13) with high probability. This result was obtained by Arias-Castro (2011) in a context similar to ours under assumption (16). Theorem 3.1 extends this result for $\psi_n > 0$. Note also that proof of this theorem (see Section 7) is different from Arias-Castro (2011) and rely on support density estimation tools proposed by Biau, Cadre and Pelletier (2008).

4. Simulation study

In this section, we provide simulation results enlightening the efficiency of the proposed estimator. To this end, Y is simulated from mixtures of univariate Gaussian laws whereas several scenarios on the distribution of X are considered. To illustrate Theorem 2.1 and Theorem 3.1, we compare the accuracy of our two-step estimate $\widehat{f}_i$ (see (5)) with the accuracy of the oracle estimate $\bar{f}_i$ (see (2)). Such comparisons are made in both Sections 4.1 and 4.2. However, each of these sections focus on special points.

In Section 4.1, the two-step estimate is also compared with the classical EM algorithm. Even if this algorithm is known to be efficient under the parametric assumption made on the distribution of Y, it does not take advantage of the presence of covariates X. It allows our method to outperform the EM algorithm in favorable situations.

In Section 4.2, different clustering procedures on X are considered on several classical data sets. In particular the behavior of the spectral clustering and the k-means algorithm are studied. Both of them are compared with the hierarchical method studied in Section 3.2.

4.1. Comparison with the EM algorithm

In this simulation section, density of Y is given by

$$f(t) = \frac{3}{4} f_1(t) + \frac{1}{4} f_2(t), \quad t \in \mathbb{R}$$

where f_1 and f_2 stand for the densities of the normal distribution with mean $-\Delta$ and Δ and variance 1. Parameter Δ measures the separation between the components f_1 and f_2 (see Figure 3).

Two scenarios are considered for the distribution of X. In the first one, conditional densities $g_{i,n}, i = 1, 2$ are uniform univariate densities:

$$g_{1,n}(x) = \mathbb{I}_{]0,1[}(x) \quad \text{and} \quad g_{2,n}(x) = \frac{1}{2}\mathbb{I}_{]1+\delta_n, 3+\delta_n[}(x), \quad x \in \mathbb{R}$$

where $\delta_n > 0$ measures the distance between the supports of $g_{1,n}$ and $g_{2,n}$. For the second one, we consider the mixture of Laplace distributions discussed in Section 3.2.2: conditional densities $g_{i,n}, i = 1, 2$ are given by

$$g_{i,n}(x) = \frac{1}{2\sigma_n} \exp\left(-\frac{|x - \mu_{i,n}|}{\sigma_n}\right), \quad i = 1, 2,$$

where $\sigma_n = 1, \mu_{1,n} = 1$ and $\mu_{2,n} = \mu_{1,n} + \ell_n$ where $\ell_n > 0$. Observe that supports of $g_{i,n}$ are disjoints in the uniform scenario while they overlap in the Laplace example. The separation between these conditional distributions is represented by the location parameters δ_n and ℓ_n.

For the two proposed scenarios, estimators $\widehat{f}_1$ and $\widehat{f}_2$ defined in (5) are computed using the hierarchical clustering procedure proposed in Section 3.2. These estimates are compared in terms of L_1-error with the oracle (but unobservable) estimates $\bar{f}_1$ and $\bar{f}_2$ defined in (2). Nonparametric kernel estimates $\bar{f}_i$ and $\widehat{f}_i$

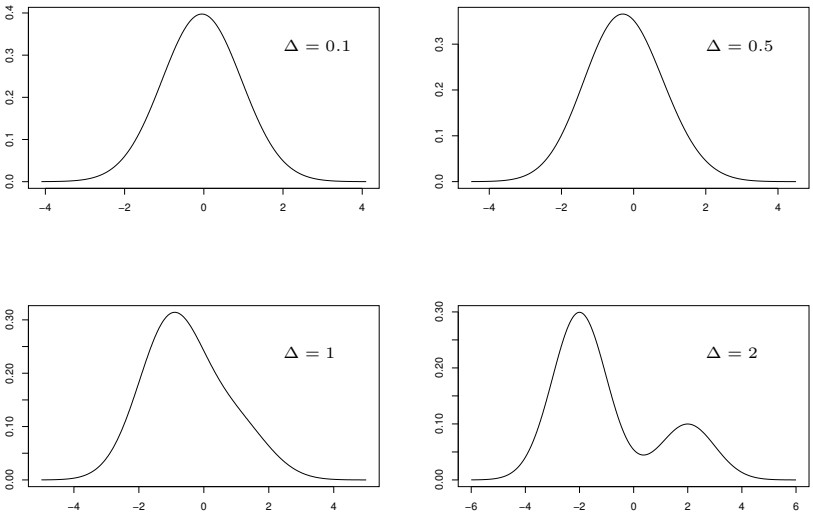

FIG 3. *Density of Y for various values of Δ.*

TABLE 1
L_1-ratio (18) evaluated over 500 replications

	Uniform: $\mathcal{R}(\widehat{f}_1)$ for $\delta_n = \cdots$			Laplace: $\mathcal{R}(\widehat{f}_1)$ for $\ell_n = \cdots$			$\mathcal{R}(\bar{f}_1)$
	0.03	0.05	0.1	4.5	5.5	6.5	
$\Delta = 0.1$	0.636	0.563	0.464	0.817	0.509	0.476	0.464
$\Delta = 0.5$	1.156	0.923	0.679	1.261	0.749	0.692	0.679
$\Delta = 1$	1.772	1.288	0.844	1.769	0.954	0.869	0.843
$\Delta = 2$	4.243	2.876	1.702	4.298	2.093	1.830	1.701

are computed with a Gaussian kernel. Recall that this paper does not put forth any theory for selecting the bandwidth h in an optimal way (see Remark 2.2). Here we use the default data-driven procedure proposed in the GNU-R library **np** (see Hayfield and Racine (2008)). In addition, these nonparametric density estimates are compared with the EM algorithm (Dempster, Laird and Rubin (1977)) known to perform well to estimate parameters in a Gaussian mixture model. Formally, we run this algorithm on the sample $Y_1, \ldots, Y_n$ to estimate Gaussian parameters of f_1 and f_2. We use the GNU-R library **mclust** and denote by f_1^{em} and f_2^{em} the resulting estimates. They are used as a benchmark. We set $n = 300$ and, for the sake of clarity, we present the results regarding f_1 only since conclusions are the same for f_2. Table 1 presents, for different values of Δ, δ_n and ℓ_n, the ratio

$$\mathcal{R}(\tilde{f}_1) = \frac{\mathbb{E}\|\tilde{f}_1 - f_1\|_1}{\mathbb{E}\|f_1^{em} - f_1\|_1} \tag{18}$$

where $\tilde{f}_1$ is either $\widehat{f}_1$ or $\bar{f}_1$. Expectations are evaluated over 500 Monte Carlo replications.

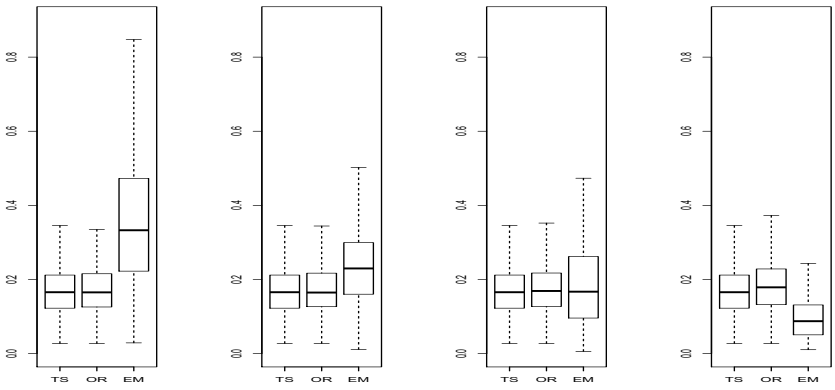

FIG 4. *Boxplot of the L_1-error for the estimate f_1^{em} (EM), the oracle estimate $\bar{f}_1$ (OR) and the two-step estimate $\widehat{f}_1$ (TS) for the Laplace example. The separation distance Δ between f_1 and f_2 vary from 0.1 (left) to 2 (right) and $\ell_n = 5.5$.*

As expected, the performances of the EM algorithm clearly depend on the separation distance between the target densities f_1 and f_2. For large Δ values, parametric estimates resulting from the EM algorithm outperform the nonparametric estimates proposed in this paper (e.g. $\Delta = 2$ in Figure 4). This is not the case when f_1 is closed to f_2: L_1-performance of $\widehat{f}_1$ over f_1^{em} is significantly better for $\Delta = 0.1$ and $\Delta = 0.5$ and roughly similar for $\Delta = 1$. Note also that the L_1-error of $\widehat{f}_1$ does not depend on Δ (see Figure 4). Figure 5 displays scatterplots of the L_1-error of $\widehat{f}_1$ versus those of the oracle $\bar{f}_1$ for $\Delta = 1$. As proved in Theorem 2.1, most points are above the diagonal. The distance from a point to the first bisector measures to some extent the distance between $\widehat{f}_1$ and $\bar{f}_1$ in terms of L_1-error. The closer to the bisector, the better $\widehat{f}_1$. In other words, this distance represents the performance of the clustering algorithm. We observe that points move closer to the first bisector as separation parameters δ_n and ℓ_n increase. As explained in Theorem 3.1, performances of the hierarchical clustering algorithm depend on the separation parameters δ_n and ℓ_n: when these parameters increase, performances of $\widehat{f}_1$ become similar to those of the oracle $\bar{f}_1$. Indeed, in our simulations, we observe that L_1-error of $\widehat{f}_1$ and $\bar{f}_1$ are quite the same for $\delta_n = 0.1$ (resp. $\ell_n = 6.5$) in the uniform case (resp. Laplace case).

4.2. A comparison of clustering algorithms

As discussed in Section 3, any clustering algorithm could be applied in practice. However, it is clear that L_1-performances of the proposed estimate depend largely on the performances of the clustering method. The problem is to find the appropriate clustering algorithm according to the covariates X. In this section, we propose to compare three standard clustering procedures: the hierarchical clustering algorithm presented in Section 3.2, the spectral clustering algorithm

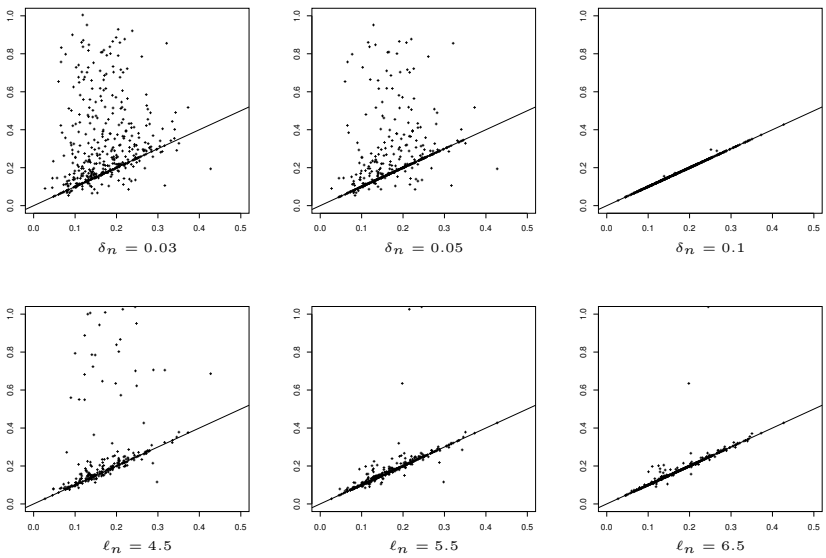

FIG 5. *L_1-error of $\bar{f}_1$ (x-axis) and $\widehat{f}_1$ (y-axis) for the uniform (up) and Laplace (down) example.*

performed with a Gaussian kernel (see Arias-Castro (2011)) and the k-means algorithm.

The model is as follows. The density of Y is now given by

$$f(t) = \frac{1}{2} f_1(t) + \frac{1}{2} f_2(t), \quad t \in \mathbb{R}$$

where f_1 and f_2 stand for the densities of the normal distribution with mean -1 and 1 and variance 1. Here, random variable X takes values in $\mathbb{R}^2$ and we again consider two scenarios for its distribution:

- "Circle-Square" model (see Baudry (2009)): $g_{1,n}$ is the density of the Gaussian distribution with mean $(a, 0)$ and identity variance covariance matrix; $g_{2,n}$ is the density of the uniform distribution over the square $[-1, 1]^2$ (see Figure 6).
- "Concentric circles" model (see Ng, Jordan and Weiss (2002)): $g_{1,n}$ is the density of the uniform distribution over $\mathcal{C}(0, r_1 + \varepsilon, r_1 - \varepsilon)$ and $g_{2,n}$ represents the uniform distribution over $\mathcal{C}(0, r_2 + \varepsilon, r_2 - \varepsilon)$, where for $r > 0$ and $\varepsilon > 0$ $\mathcal{C}(0, r + \varepsilon, r - \varepsilon)$ represents the set between circles with center 0 and radius $r + \varepsilon$ and $r - \varepsilon$ (see Figure 7). We fix $r_1 = 0.3$, $\varepsilon = 0.15$ and consider many values for r_2 such that $r_2 > r_1 + 2\varepsilon$.

The difficulty encountered in identifying each group depends on parameters a and r_2. The smaller a and r_2, the harder to identify the clusters.

For the two described examples, we use the two-step kernel density estimator for three clustering algorithms: hierarchical, spectral and k-means. The resulting

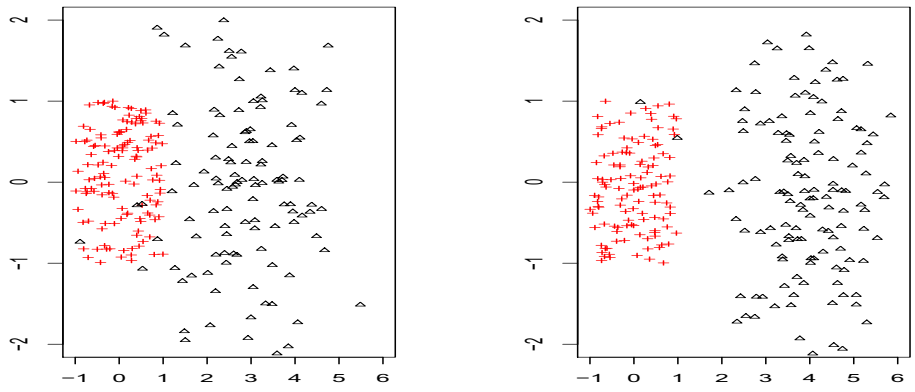

FIG 6. *A sample of* $n = 250$ *observations for the "Circle-Square" model with* $a = 3$ *(left) and* $a = 4$ *(right).*

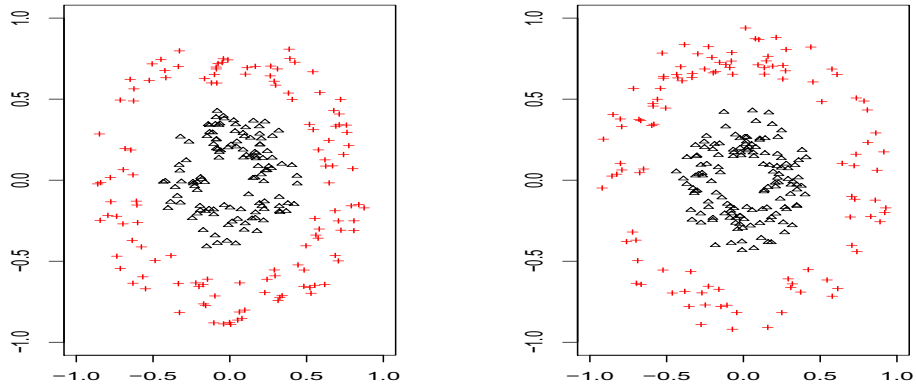

FIG 7. *A sample of* $n = 250$ *observations for the "Concentric circles" model with* $r_2 = 0.75$ *(left) and* $r_2 = 0.80$ *(right).*

estimates are compared with the oracle estimates $\bar{f}_1$ and $\bar{f}_2$. We keep the same setting as above to compute estimates $\widehat{f}_1$ and $\widehat{f}_2$: Gaussian kernel and bandwidth selected with the library **np**. For the sake of clarity, we again present only results on $\widehat{f}_1$ since we observe the same conclusions for $\widehat{f}_2$. Table 2 and Table 3 present the ratio

$$\mathcal{R}(\widehat{f}_1) = \frac{\mathbb{E}\|\widehat{f}_1 - f_1\|_1}{\mathbb{E}\|\bar{f}_1 - f_1\|_1}, \tag{19}$$

for many values of a, r_2 and n. Expectations are evaluated over 500 Monte-Carlo replications and Figure 8 presents boxplots of the L_1-error of the different estimates. For each replications, we also compute the error of the clustering procedure

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{\widehat{I}_k \neq I_k}$$

TABLE 2
Error ratio (19) evaluated over 500 Monte Carlo replications for the "Circle-Square" example

		Hier.		Spect.		k-means	
		$\mathcal{R}(\widehat{f}_1)$	err_n	$\mathcal{R}(\widehat{f}_1)$	err_n	$\mathcal{R}(\widehat{f}_1)$	err_n
$a=3$	$n=250$	4.680	0.475	1.748	0.121	1.047	0.043
	$n=500$	6.370	0.483	2.265	0.126	1.034	0.043
$a=4$	$n=250$	3.565	0.382	1.107	0.018	1.005	0.013
	$n=500$	5.688	0.449	1.190	0.023	1.000	0.013
$a=5$	$n=250$	1.285	0.067	0.999	0.001	0.997	0.003
	$n=500$	1.897	0.130	0.999	0.001	1.000	0.003

TABLE 3
Error ratio (19) evaluated over 500 Monte Carlo replications for the "Concentric circles" example

		Hier.		Spect.		k-means	
		$\mathcal{R}(\widehat{f}_1)$	err_n	$\mathcal{R}(\widehat{f}_1)$	err_n	$\mathcal{R}(\widehat{f}_1)$	err_n
$r_2=0.75$	$n=250$	4.040	0.349	2.776	0.195	4.568	0.468
	$n=500$	1.197	0.021	1.013	0.001	5.993	0.478
$r_2=0.80$	$n=250$	1.852	0.105	1.433	0.049	4.556	0.467
	$n=500$	1.010	0.001	1.000	0.000	5.986	0.477

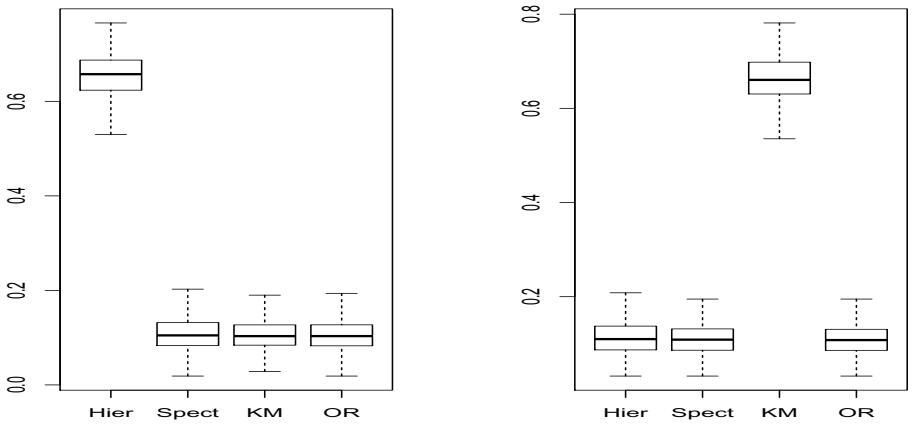

FIG 8. *Boxplot of the L_1-error for the oracle estimate $\bar{f}_1$ (OR) and two-step estimator $\widehat{f}_1$ using the hierarchical algorithm (Hier), spectral clustering algorithm (Spect) and k-means algorithm (KM). Results are for "Circle-Square" dataset with $a=4$ and $n=500$ (left) and "Concentric circles" dataset with $r_2=0.75$ and $n=500$ (right).*

and we display in Table 2 and Table 3 this error term averaged over the 500 replications (it is denoted err_n). Observe that this term is closely related to the maximal misclassification error φ_n.

As proved in Theorem 2.1, performances of $\widehat{f}_1$ depend on the accuracy of the clustering approach: the lower err_n, the better $\widehat{f}_1$. For the "Circle Square" dataset, unsurprisingly k-means algorithm overperforms the two other clustering methods. Indeed, k-means is well appropriate to this dataset since clusters can be

identified by their distances to two particular points (the centers of the uniform and Gaussian distributions). It is not the case for the "Concentric circle" dataset where estimates defined from hierarchical and spectral clustering algorithms achieve the best estimated L_1-error.

5. Application to electricity distribution

5.1. Context of the study

ERDF is the contract-holder of the public electricity distribution network in France. ERDF is in charge of operating, maintaining and developing the network. With 36,000 employees and 35 million customers served over 34,220 communes, ERDF is the largest electricity distributor in Europe. It operates more than 1.3 million km in power lines and runs more than 11 million operations per year. ERDF also plays an essential role in ensuring the proper functioning of the competitive electricity market by providing quality electricity supply among the best in Europe, and serving all network users without resorting to guaranteeing discriminatory practices.

In recent years, the electricity sector has entered a period of profound changes resulting from the emergence of decentralized and intermittent (wind, solar) means of production and new electricity uses (e.g. electric vehicle). The increasing integration of these new means of production and new uses has a major impact on ERDF's core business: connecting new users (producers, terminals electric vehicle), and adaptating rules of conduct and network planning/investment to meet the new specifications. ERDF has initiated its digital transformation plan so as to take advantage of new information technologies, and by meeting its new challenges, offer better public service. ERDF launched the "smart grid" experimental programs in order to run the network with more flexibility and efficiency. To do so, these programs use detailed network status and mine/produce information from different users. These more detailed data (including from a new generation of electricity meters, called smart meters) will accordingly be used to improve network monitoring (predictive maintenance).

In this section we focus on the detection of customers who experience a significant decrease in consumption, for a given period of time, *i.e.*, a period when overall malfunction of the network could be observed. This will make it possible to better understand the origin of dysfunctions and thus better forecast network operation. For this study, we have the benefit of a set of consumption curves for 226 customers with observations taken at regularly spaced instants. Based on the observation of the individual consumption curves, we can cluster individuals into two groups (those who have suffered an abnormal decline and the others) and estimate, in each group, distributions of many variables using the approach proposed in this paper.

5.2. Application of the two-step estimator

The consumption curves of $n = 226$ ERDF's customers are observed at 9 regularly spaced instants $t_1, \ldots, t_9$. The time interval $[t_1, t_9]$ covers a known period

of disruption between times t_4 and t_6. The observations consist of n vectors $\mathbf{Z}_k = (Z_{k1}, \ldots, Z_{k9}) \in \mathbb{R}^9$ where Z_{kj} stands for the consumption of user k at time t_j.

Since ERDF is interested in comparing the behavior of customers of both sub-populations (those who have suffered from the disruption and others) before and after the disruption period, we consider 6 different variables $Y^{(j)}$ in relation with the consumption around the disruption period. These variables, presented below, are observed for each customer and thus are defined for any $k \in [\![1, n]\!]$.

1. Average consumptions before, during and after the disruption period defined by:

$$Y_k^{(1)} = \frac{Z_{k1} + Z_{k2} + Z_{k3}}{3}, \quad Y_k^{(2)} = \frac{Z_{k4} + Z_{k5} + Z_{k6}}{3}$$

$$\text{and} \quad Y_k^{(3)} = \frac{Z_{k7} + Z_{k8} + Z_{k9}}{3};$$

2. Evolutions of consumption around the disruption period defined by:

$$Y_k^{(4)} = \frac{Y_k^{(2)} - Y_k^{(1)}}{Y_k^{(1)}}, \quad Y_k^{(5)} = \frac{Y_k^{(3)} - Y_k^{(1)}}{Y_k^{(1)}} \quad \text{and} \quad Y_k^{(6)} = \frac{Y_k^{(3)} - Y_k^{(2)}}{Y_k^{(2)}}.$$

Let I be the random variable taking value 1 if a customer has been affected by the disruption, 2 otherwise. If we denote by $f_1^{(j)}$ and $f_2^{(j)}$ the conditional densities of $\mathcal{L}(Y^{(j)}|I = 1)$ and $\mathcal{L}(Y^{(j)}|I = 2)$, the problem is to compare $f_1^{(j)}$ with $f_2^{(j)}$ for each $j \in [\![1, 6]\!]$. Even if ERDF can measure consumptions during the disruption period (between t_4 and t_6), it does not have the capacity to identify consumers affected by the perturbation. It means that random variables $I_k, k = 1, \ldots, n$ are not observed. However, we know that users impacted by the disruption posted a decline in consumption during t_4 and t_6. Figure 9 provides examples of customers potentially affected by the disruption (for confidentiality reasons, representations are anonymous and scales of power are not specified).

Using the approach developed in this paper, we first have to identify users impacted by the disruption with a clustering algorithm. As the disruption influences the consumptions of user k between t_4 and t_6 we define $X_k = (X_{k1}, X_{k2})$, $k = 1, \ldots, n$ with

$$X_{k1} = \min\left(v_{k,54}, v_{k,65}\right), \quad X_{k2} = v_{k,54} + v_{k,65}$$

where

$$v_{k,ij} = \frac{Z_{kj} - Z_{ki}}{Z_{ki}}, \quad 1 \leq i, j \leq 9.$$

Observe that $v_{k,ij}$ measures the relative variation of consumption for user k between t_i and t_j. It follows that $X_k = (X_{k1}, X_{k2})$ captures the development of consumption of user k during the disruption period. We use these covariates to cluster users into two groups: the first contains consumers assumed to be affected by the disruption, the second contains the others.

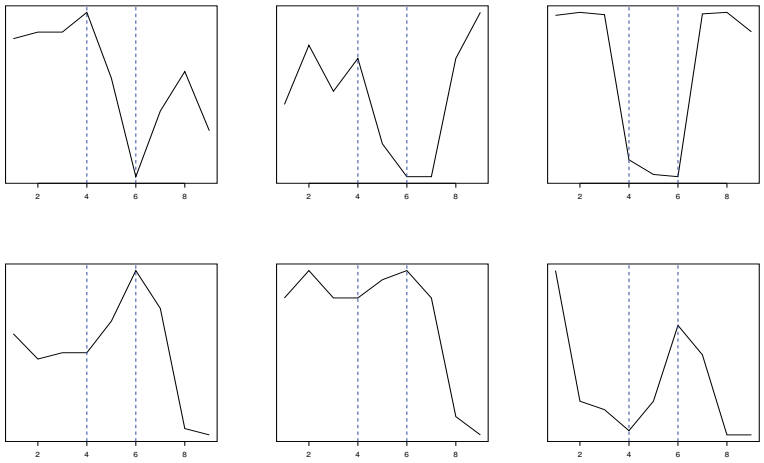

FIG 9. *Consumptions of users suspected to be affected (up) or not (down) by the perturbation.*

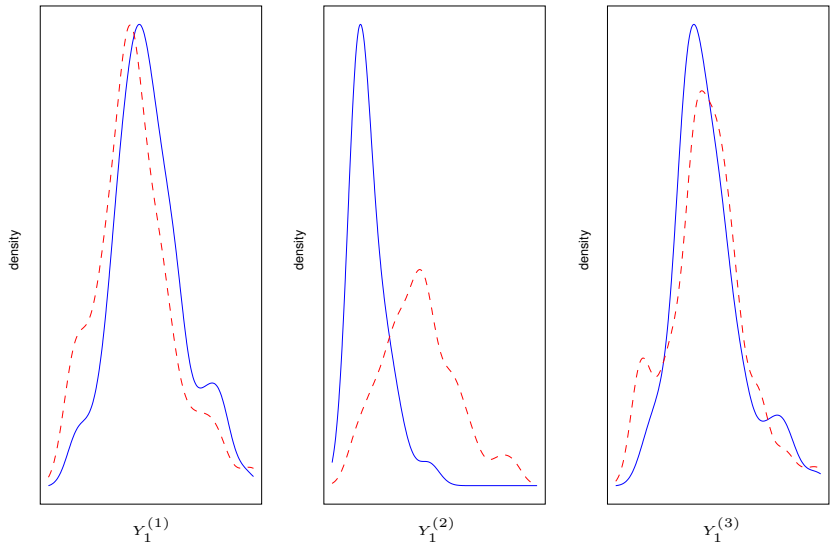

FIG 10. *Kernel estimates* $\widehat{f}_1^{(j)}$ *(solid lines) and* $\widehat{f}_2^{(j)}$ *(dashed lines) for* $j = 1$ *(left), 2 (center) and 3 (right).*

Two clustering algorithms have been tested: the hierarchical method studied in section 3.2 and the k-means algorithm. Since these methods lead to approximately the same clusters, we only present results for the hierarchical method. Figures 10 and 11 present kernel density estimates (5) of conditional densities $f_1^{(j)}$ and $f_2^{(j)}$ for $j \in [\![1, 6]\!]$. Parameters (bandwidth and kernel) of the kernel estimates are chosen as in the simulation part. For confidentiality reasons, scales of power are again not specified.

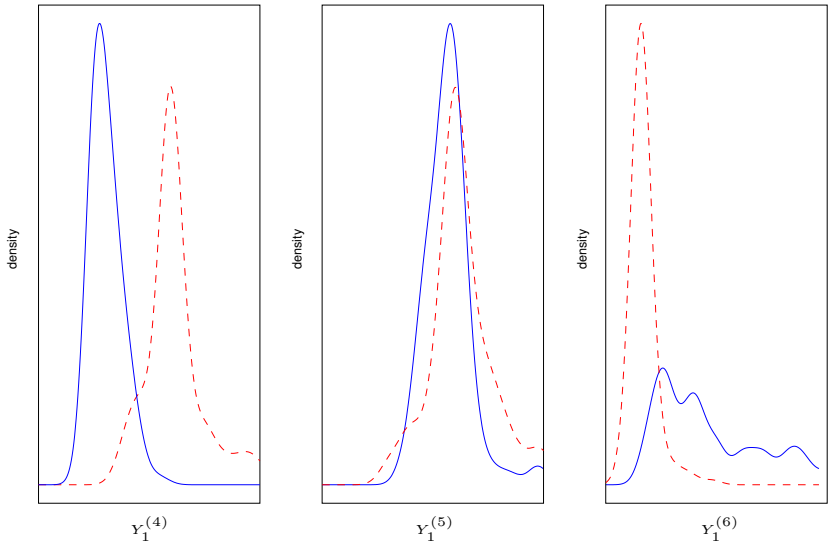

FIG 11. *Kernel estimates* $\widehat{f}_1^{(j)}$ *(solid lines) and* $\widehat{f}_2^{(j)}$ *(dashed lines) for* $j = 4$ *(left), 5 (center) and 6 (right).*

Figure 10 strongly supports the idea that the clustering procedure allows to correctly identify users impacted by the disruption. Indeed we observe that the average consumption during the disruption period is lower for consumers in the first group (second graph in Figure 10). We can also observe that average consumptions are quite the same for the two groups before and after the disruption period. It means that users impacted by the perturbation do not over-consume after the disruption period. This conclusion is also supported by the second graph in Figure 11: distributions representing the evolution of consumptions are similar for the two clusters.

6. Conclusion

This paper provides a new framework to estimate conditional densities in mixture models in the presence of covariates. To our knowledge, no clear probabilistic model has been proposed to take into account of the presence of covariates. The model we consider includes such covariates and Theorem 2.1 precisely describes the interest of a preliminary clustering step on these covariates to estimate components of the mixture model. It is shown that the performances of these estimates depend on the maximal misclassification error (4) of the clustering algorithm. This criterion is natural to measure performances of clustering algorithms but, as far as we know, it has not been addressed before. We obtain non-asymptotic upper bounds of this error term is section 3.2 for a particular hierarchical algorithm. This algorithm is not new but it has not been studied in this context. Results obtained for this algorithm could be extended to other clustering algorithms based on pairwise distances such as spectral clustering

(Arias-Castro (2011)) or on clustering methods based on neighborhoods graphs (Maier, Hein and Von Luxburg (2009)). Even if main contributions of this work are theoretical, both the simulation study and the application on real data enlighten the efficiency of the proposed estimator in the presence of covariates.

Acknowledgement

We would like to thank the editor, an associate editor as well as two anonymous referees for very thoughtful and detailed comments. We are also grateful to Datastorm and ERDF for providing us with the data-set used in the case study.

7. Proofs

7.1. Proof of Theorem 2.1

We first prove inequality (6). Since

$$\mathbb{E}\big\|\widehat{f}_i - f_i\big\|_1 \le \mathbb{E}\big\|\bar{f}_i - f_i\big\|_1 + \mathbb{E}\big\|\widehat{f}_i - \bar{f}_i\big\|_1,$$

we need only find an upper bound of the second term in the right-hand side of the previous inequality. Since $\bar{f}_i = 0$ when $N_i = 0$ and $\|\widehat{f}_i\|_1 = \|K\|_1$, we have

$$\begin{aligned}\mathbb{E}\big\|\widehat{f}_i - \bar{f}_i\big\|_1 &\le \mathbb{E}\left(\big\|\widehat{f}_i\big\|_1 \mathbb{I}_{N_i=0}\right) + \mathbb{E}\big\|(\widehat{f}_i - \bar{f}_i)\mathbb{I}_{N_i>0}\big\|_1 \\ &\le \|K\|_1(1-\alpha_i)^n + \mathbb{E}\big\|(\widehat{f}_i - \bar{f}_i)\mathbb{I}_{N_i>0}\big\|_1.\end{aligned}$$

For the sake of readability, let $\widetilde{\mathbb{E}}$ denote the conditional expectation with respect to $(I_1, \ldots, I_n)$ and $\widetilde{\widetilde{\mathbb{E}}}$ the conditional expectation with respect to $(I_1, \ldots, I_n, X_1, \ldots, X_n)$. Moreover, let

$$\begin{aligned}A_i(t) &= \big(\widehat{f}_i(t) - \bar{f}_i(t)\big)\mathbb{I}_{Ni>0} \\ &= \sum_{k=1}^{n} K_h(t, Y_k)\left(\frac{\mathbb{I}_{\{i\}}(\widehat{I}_k)}{\widehat{N}_i} - \frac{\mathbb{I}_{\{i\}}(I_k)}{N_i}\right)\mathbb{I}_{N_i>0}.\end{aligned}$$

Using these notations it is easily seen that

$$\mathbb{E}\Big\|(\widehat{f}_i - \bar{f}_i)\mathbb{I}_{N_i>0}\Big\|_1 = \mathbb{E}\widetilde{\mathbb{E}}\int_{\mathbb{R}} \widetilde{\widetilde{\mathbb{E}}}|A_i(t)|dt. \tag{20}$$

Since, for all $y \in \mathbb{R}$ we have $\int_{\mathbb{R}} |K_h(t,y)|dt = \|K\|_1$, we deduce that

$$\begin{aligned}\int_{\mathbb{R}} \widetilde{\widetilde{\mathbb{E}}}|A_i(t)|\,dt &\le \sum_{k=1}^{n} \widetilde{\widetilde{\mathbb{E}}}\left(\int_{\mathbb{R}} |K_h(t, Y_k)|\,dt\right)\left|\frac{\mathbb{I}_{\{i\}}(\widehat{I}_k)}{\widehat{N}_i} - \frac{\mathbb{I}_{\{i\}}(I_k)}{N_i}\right| \\ &\le \|K\|_1 \sum_{k=1}^{n}\left|\frac{\mathbb{I}_{\{i\}}(\widehat{I}_k)}{\widehat{N}_i} - \frac{\mathbb{I}_{\{i\}}(I_k)}{N_i}\right|.\end{aligned}$$

Thus

$$\widetilde{\mathbb{E}}\int_{\mathbb{R}}\widetilde{\widetilde{\mathbb{E}}}|A_i(t)|\,dt \le \frac{\|K\|_1}{N_i}\widetilde{\mathbb{E}}\left[\frac{1}{\widehat{N}_i}\sum_{k=1}^{n}|N_i\mathbb{I}_{\{i\}}(\widehat{I}_k)-\widehat{N}_i\mathbb{I}_{\{i\}}(I_k)|\right]. \tag{21}$$

Moreover, inserting $\widehat{N}_i\mathbb{I}_{\{i\}}(\widehat{I}_k)$ in the previous expectation, we obtain

$$\begin{aligned}\widetilde{\mathbb{E}}\left[\frac{1}{\widehat{N}_i}\sum_{k=1}^{n}|N_i\mathbb{I}_{\{i\}}(\widehat{I}_k)-\widehat{N}_i\mathbb{I}_{\{i\}}(I_k)|\right]&\\ \le \widetilde{\mathbb{E}}|N_i-\widehat{N}_i|+\widetilde{\mathbb{E}}\sum_{k=1}^{n}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|&\\ \le 2\widetilde{\mathbb{E}}\sum_{k=1}^{n}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|.&\end{aligned} \tag{22}$$

Combining (20), (21) and (22) leads to

$$\begin{aligned}\mathbb{E}\big\|(\widehat{f}_i-\bar{f}_i)\mathbb{I}_{N_i>0}\big\|_1 &\le 2\|K\|_1\sum_{k=1}^{n}\mathbb{E}\left[\frac{\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\right]\\ &\le \frac{2\|K\|_1}{n\alpha_i}\sum_{k=1}^{n}\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\right].\end{aligned} \tag{23}$$

The expectation on the right-hand side of this inequality can be bounded in the following way

$$\begin{aligned}\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\right] &\le \mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}\le 2}\right]\\ &\quad+\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right].\end{aligned} \tag{24}$$

For the first term of this bound, we have

$$\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}\le 2}\right]\le 2\varphi_n, \tag{25}$$

while for the second term, we obtain from Hölder inequality that

$$\begin{aligned}&\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right]\\ &\le\sqrt{\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right]^2\mathbb{P}\left(\frac{n\alpha_i}{N_i}>2\right)}\\ &\le\sqrt{\mathbb{E}\left(\frac{(n\alpha_i)^2}{N_i^2}\mathbb{I}_{N_i>0}\right)\mathbb{P}\left(N_i-n\alpha_i<-\frac{n\alpha_i}{2}\right)}.\end{aligned} \tag{26}$$

Now, it can be easily seen that

$$\mathbb{E}\left(\frac{(n\alpha_i)^2}{N_i^2}\mathbb{I}_{N_i>0}\right) \leq 6\mathbb{E}\left(\frac{(n\alpha_i)^2}{(N_i+1)(N_i+2)}\right) \leq 6, \tag{27}$$

where the last inequality follows from Hengartner and Matzner-Løber (2009). Using Hoeffding's inequality (see Hoeffding (1963)) we obtain for the second term in (24)

$$\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k) - \mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right] \leq \sqrt{6}\exp\left(-\frac{n{\alpha_i}^2}{4}\right). \tag{28}$$

From (23) – (28), we deduce that

$$\mathbb{E}\left\|(\widehat{f}_i - \bar{f}_i)\mathbb{I}_{N_i>0}\right\|_1 \leq \frac{4\|K\|_1}{\alpha_i}\varphi_n + \frac{2\sqrt{6}\|K\|_1}{\alpha_i}\exp\left(-\frac{n{\alpha_i}^2}{4}\right).$$

Putting all of the pieces together, we obtain

$$\mathbb{E}\left\|\widehat{f}_i - \bar{f}_i\right\|_1 \leq \frac{4\|K\|_1}{\alpha_i}\varphi_n + \frac{2\sqrt{6}\|K\|_1}{\alpha_i}\exp\left(-\frac{{\alpha_i}^2}{4}\cdot n\right) + \|K\|_1\exp\left(-n\log(1-\alpha_i)\right),$$

which concludes the first part of the proof.

Inequality (7) is proved as follows

$$\begin{aligned}\mathbb{E}|\widehat{\alpha}_i - \alpha_i| &\leq \mathbb{E}\left|\frac{\widehat{N}_i}{n} - \frac{N_i}{n}\right| + \mathbb{E}\left|\frac{N_i}{n} - \alpha_i\right| \\ &\leq \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}|\mathbb{I}_{\{i\}}(\widehat{I}_k) - \mathbb{I}_{\{i\}}(I_k)| + \frac{1}{n}\sqrt{\mathbb{V}(N_i)} \\ &\leq \varphi_n + \sqrt{\frac{\alpha_i(1-\alpha_i)}{n}}.\end{aligned}$$

7.2. *Proof of Proposition 3.1*

Let k be an arbitrary integer in $[\![1, n]\!]$. We have to bound $\mathbb{P}(\widehat{I}_k \neq i|I_k = i)$ for $i = 1, 2$. To do so, we first consider the case $i = 2$:

$$\begin{aligned}\mathbb{P}(\widehat{I}_k \neq 2|I_k = 2) &= \mathbb{P}(\widehat{I}_k \neq 2, 1-\lambda_n < X_k < 1|I_k = 2) \\ &\quad + \mathbb{P}(\widehat{I}_k \neq 2, X_k \geq 1|I_k = 2) \\ &= \mathbb{P}(1-\lambda_n < X_k < 1|I_k = 2)\end{aligned}$$

because, by definition, $\widehat{I}_k \neq 2 \iff X_k < 1$. Thus

$$\mathbb{P}(\widehat{I}_k \neq 2|I_k = 2) = \int_{1-\lambda_n}^{1} g_{2,n}(x)dx = \lambda_n. \tag{29}$$

Next, if $i = 1$ it is easy to see that $\mathbb{P}(\widehat{I}_k \neq 1 | I_k = 1) = \mathbb{P}(X_k \geq 1 - \widehat{\lambda}_n | I_k = 1)$. Let us consider

$$\mu_n = \lambda_n + \frac{2}{\alpha_2} \cdot \frac{\log n}{n} \quad \text{and} \quad A = \left\{ 1 - \widehat{\lambda}_n \geq 1 - \mu_n \right\}.$$

Using these notations we obtain

$$\begin{aligned} \{X_k \geq 1 - \lambda_n\} &= \left(\{X_k \geq 1 - \widehat{\lambda}_n\} \cap A \right) \cup \left\{ X_k \geq 1 - \widehat{\lambda}_n \right\} \cap \bar{A} \big) \\ &\subseteq \{X_k \geq 1 - \mu_n\} \cup \left\{ \widehat{\lambda}_n \geq \mu_n \right\}. \end{aligned}$$

This leads to the following inequality

$$\mathbb{P}(\widehat{I}_k \neq 1 | I_k = 1) \leq \mu_n + \mathbb{P}\left(X_{(n)} \leq 2 - \mu_n | I_k = 1 \right). \tag{30}$$

Since X_ℓ and I_k are independent for $k \neq \ell$, we obtain the following bound for the last probability

$$\begin{aligned} &\mathbb{P}\left(X_{(n)} \leq 2 - \mu_n | I_k = 1 \right) \\ &\qquad = \mathbb{P}\left(\forall \ell, X_\ell \leq 2 - \mu_n | I_k = 1 \right) \\ &\qquad = \left(\prod_{\ell \neq k} \mathbb{P}\left(X_\ell \leq 2 - \mu_n \right) \right) \mathbb{P}\left(X_k \leq 2 - \mu_n | I_k = 1 \right). \end{aligned}$$

The independence of the X_ℓ's and simple calculations lead to

$$\begin{aligned} \mathbb{P}\left(X_{(n)} \leq 2 - \mu_n | I_k = 1 \right) &= \left(\mathbb{P}(X_1 \leq 2 - \mu_n) \right)^{n-1} \\ &= \left(1 - 2n^{-1}(\log n) \right)^{n-1} \\ &\leq n^{-1}, \end{aligned} \tag{31}$$

where the last inequality follows, for $n \geq 2$, from the fact that $1 - u \leq e^{-u}$ for all $u \geq 0$. Taking together equations (30) and (31), we finally obtain

$$\mathbb{P}(\widehat{I}_k \neq 1 | I_k = 1) \leq \lambda_n + n^{-1} + \frac{2}{\alpha_2} \cdot \frac{\log n}{n}. \tag{32}$$

Proposition follows from equations (29) and (32).

7.3. Proof of Theorem 3.1

Since $\delta_n > 2r_n$ we have for all $(i, j) \in [\![1, M]\!]^2$ with $i \neq j$:

$$\left(\bigcup_{k : X_k \in S_{i,n}} B(X_k, r_n) \right) \cap \left(\bigcup_{k : X_k \in S_{j,n}} B(X_k, r_n) \right) \subseteq (S_{i,n} + r_n) \cap (S_{j,n} + r_n) = \emptyset, \tag{33}$$

where, for $S \subset \mathbb{R}^d$ and $r > 0$, we recall that

$$S + r = \{x \in \mathbb{R}^d : \exists y \in S \text{ such that } \|x - y\|_2 \leq r\}.$$

Inclusion (33) implies $\widehat{M}_{r_n} \geq M$. Moreover, observe that if

$$r_n \in \mathcal{R}_M = \{r > 0 : \widehat{M}_r \leq M\} \tag{34}$$

then $\widehat{M}_{r_n} = M$ and the affinity matrices A^{r_n} and $A^{\widehat{r}_n}$ defined in (12) induce the same clusters $\mathcal{X}_1(r_n), \ldots, \mathcal{X}_M(r_n)$. Furthermore, if (34) is verified, it is easily seen that $\forall i \in [\![1, M]\!], \exists j \in [\![1, M]\!]$ such that

$$\{X_k : X_k \in S_{i,n} + r_n\} \subseteq \mathcal{X}_j(r_n).$$

For simplicity, when (34) is satisfied, we index clusters $\mathcal{X}_1(r_n), \ldots, \mathcal{X}_M(r_n)$ such that

$$\{X_k : X_k \in S_{i,n} + r_n\} \subseteq \mathcal{X}_i(r_n), \quad i \in [\![1, M]\!].$$

We deduce that

$$\begin{aligned}
\mathbb{P}(\widehat{I}_k \neq I_k) &\leq \mathbb{P}(\{\widehat{I}_k \neq I_k\} \cap \{r_n \in \mathcal{R}_M\}) + \mathbb{P}(r_n \notin \mathcal{R}_M) \\
&\leq \mathbb{P}(\{\widehat{I}_k \neq I_k\} \cap \{r_n \in \mathcal{R}_M\} \cap \{X_k \in (S_n + r_n)\}) \\
&\qquad + \mathbb{P}(X_k \notin (S_n + r_n)) + \mathbb{P}(r_n \notin \mathcal{R}_M) \\
&\leq \sum_{i=1}^{M} \mathbb{P}(X_k \notin (S_{i,n} + r_n) | I_k = i) \mathbb{P}(I_k = i) + \psi_n + \mathbb{P}(r_n \notin \mathcal{R}_M) \\
&\leq 2\psi_n + \mathbb{P}(r_n \notin \mathcal{R}_M)
\end{aligned} \tag{35}$$

since $\mathbb{P}(X_k \notin (S_n + r_n)) \leq \psi_n$. To complete the proof, we have to find an upper bound for the probability of the event $\{r_n \notin \mathcal{R}_M\}$. Observe that

$$\begin{aligned}
\mathbb{P}(r_n \notin \mathcal{R}_M) \leq \mathbb{P}&\left(S_n \not\subseteq \bigcup_{k \in \kappa_n} B(X_k, r_n)\right) \\
&+ \mathbb{P}\left(\{r_n \notin \mathcal{R}_M\} \cap \left\{S_n \subseteq \bigcup_{k \in \kappa_n} B(X_k, r_n)\right\}\right)
\end{aligned} \tag{36}$$

where $\kappa_n = \{k \in [\![1, M]\!] : X_k \in S_n\}$. For the first term on the right hand side of the above equation, remark that inclusion

$$S_n \subseteq \bigcup_{k \in \kappa_n} B(X_k, r_n)$$

holds when for all $\ell \in [\![1, N]\!]$, the balls B_ℓ defined in assumption 2 contain at least one observation among $\{X_k, k \in \kappa_n\}$. Thus

$$\begin{aligned}
\mathbb{P}&\left(S_n \not\subseteq \bigcup_{k \in \kappa_n} B(X_k, r_n)\right) \\
&\leq \mathbb{P}\left(\exists \ell \in [\![1, N]\!], \forall k \in \kappa_n, X_k \notin B_\ell\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\ell=1}^{N} \mathbb{P}(\forall k \in \kappa_n, X_k \notin B_\ell) \\
&\leq \sum_{\ell=1}^{N} \mathbb{P}\left(\bigcap_{k=1}^{n} \{\{\{X_k \in S_n\} \cap \{X_k \notin B_\ell\}\} \cup \{X_k \notin S_n\}\} \right) \\
&\leq \sum_{\ell=1}^{N} (\mathbb{P}(\{X_k \in S_n\} \cap \{X_k \notin B_\ell\}) + \mathbb{P}(X_k \notin S_n))^n \\
&\leq \sum_{\ell=1}^{N} (1 - \mathbb{P}(X_k \in (B_\ell \cap S_n)) - \mathbb{P}(X_k \notin S_n) + \mathbb{P}(X_k \notin S_n))^n \\
&\leq \sum_{\ell=1}^{N} (1 - \mathbb{P}(X_k \in (B_\ell \cap S_n)))^n.
\end{aligned}
$$

According to assumption 2 and inequality (15), we obtain

$$
\begin{aligned}
\mathbb{P}\left(S_n \not\subseteq \bigcup_{k \in \kappa_n} B(X_k, r_n) \right) &\leq \sum_{\ell=1}^{N} (1 - t_n c_2 r_n^d)^n \\
&\leq N \left(1 - c_2 t_n r_n^d\right)^n \\
&\leq (\tau c_1 c_2)^{-1} \frac{n}{\log n} \exp(-c_2 \tau \log n).
\end{aligned}
$$

Since $c_2 \tau \geq 1 + a$ we have

$$
\mathbb{P}\left(S_n \not\subseteq \bigcup_{k \in \kappa_n} B(X_k, r_n) \right) \leq (\tau c_1 c_2)^{-1} \frac{1}{n^a \log n}. \tag{37}
$$

For the second term on the right hand side of (36), we have

$$
\mathbb{P}\left(\{r_n \notin \mathcal{R}_M\} \cap \left\{ \bigcup_{k \in \kappa_n} B(X_k, r_n) \right\} \right) \leq \mathbb{P}(\exists k \in [\![1, n]\!] : X_k \notin (S_n + r_n)) \leq n \psi_n. \tag{38}
$$

Taking (35), (37) and (38) together, result follows.

References

Arias-Castro, E. (2011). Clustering based on pairwise distances when the data is of mixed dimensions. *IEEE Transaction on Information Theory* **57** 1692–1706. MR2815843

Baudry, J. P. (2009). Sélection de modèle pour la classification non supervisée. Choix du nombre de classes. PhD thesis, Université Paris Sud 11.

Benaglia, T., Chauveau, D. and Hunter, D. R. (2009). An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics* **18** 505–526. MR2749842

Benaglia, T., Chauveau, D. and Hunter, D. R. (2011). Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In *Non-*

parametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger 15–27. World Scientific Publishing Co. MR2838717

Berlinet, A. and Devroye, L. (1994). A comparison of kernel density estimates. *Publications de l'ISUP* **38**. MR1743393

Biau, G., Cadre, B. and Pelletier, B. (2007). A graph-based estimator of the number of clusters. *ESAIM Probability and Statistics* **11** 272–280. MR2320821

Biau, G., Cadre, B. and Pelletier, B. (2008). Exact rates in density support estimation. *Journal of Multivariate Analysis* **99** 2185–2207. MR2463383

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** 719–725.

Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006). Estimation of a two-component mixture model. *The Annals of Statistics* **34** 1204–1232. MR2278356

Celeux, G. and Govaert, G. (1995). Parsimonous Gaussian models in cluster analysis. *Pattern Recognition* **28** 781–793.

Cerrito, P. B. (1992). Using stratification to estimate multimodal density functions with applications to regression. *Communications in Statistics – Simulation and Computation* **21** 1149–1164.

Cormen, T. H., Leiserson, C. E. and Rivest, R. L. (1990). *Introduction to Algorithms*. The MIT Press, Cambridge. MR1066870

Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. *Canadian Journal of Statistics* **28** 367–382. MR1792055

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39** 1–38. MR0501537

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley. MR0780746

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York. MR1843146

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56** 363–375. MR1281940

Everit, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Wiley, New York. MR0624267

Hall, P. and Titterington, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B* **46** 465–473. MR0790632

Hall, P. and Titterington, D. M. (1985). The use of uncategorized data to improve the performance of a nonparametric estimator of a mixture density. *Journal of the Royal Statistical Society, Series B* **47** 155–163. MR0805072

Hall, P. and Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics* **31** 201–224. MR1962504

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley. MR0405726

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software* **27**.

Hengartner, N. W. and Matzner-Løber, E. (2009). Asymptotic unbiased density estimators. *ESAIM. Probability and Statistics* **13** 1–14. MR2493852

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Society* **58** 13–30. MR0144363

Jeon, B. and Landgrebe, D. A. (1994). Fast Parzen density estimation using clustering-based branch and bound. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16** 950–954.

Kitamura, Y. (2004). Nonparametric identifiability of finite mixtures. Technical report – Yale University.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics* **11** 86–94. MR0684866

Lindsay, B. G. (1983b). The geometry of mixture likelihoods. II. The exponential family. *The Annals of Statistics* **11** 783–792. MR0707929

Maier, M., Hein, M. and Von Luxburg, U. (2009). Optimal construction of k-nearest-neigbor graphs for identifying noisy clusters. *Theoritical Computer Science* **410** 1749–1764. MR2514706

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and Applications to Clustering*. Dekker, New York. MR0926484

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* **14** 849–856.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33** 1065–1076. MR0143282

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26** 195–239. MR0738930

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* **27** 832–837. MR0079873

Ruzgas, T., Rudzkis, R. and Kavaliauskas, M. (2006). Application of clustering in the nonparametric estimation of distribution density. *Nonlinear Analysis: Modeling and Control* **11** 393–411.

Titterington, D. M. (1983). Minimum-distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B* **45** 37–46. MR0701074