Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2013-32

On Clustering Procedures and Nonparametric Mixture Estimation

S. AURAY[1] – N. KLUTCHNIKOFF[2]
L. ROUVIÈRE[3]

December, 2013

1 CREST-ENSAI, UCLO et CIRPEE, Canada. Email : stephane.auray@ensai.fr
2 CREST-ENSAI, IRMA et Université de Strasbourg. Email : nicolas.klutchnikoff@ensai.fr
3 CREST- ENSAI, IRMAR, UEB. Email : laurent.rouviere@ensai.fr (*corresponding author*)

On clustering procedures and nonparametric mixture estimation

Stéphane Auray [a], Nicolas Klutchnikoff[b] and Laurent Rouvière [c,*,†]

[a] CREST-Ensai
EQUIPPE (EA4018) – ULCO
and CIRPEE, Canada
stephane.auray@ensai.fr

[b] CREST-Ensai, IRMA (UMR CNRS 7501)
and Université de Strasbourg
nicolas.klutchnikoff@ensai.fr

[c] CREST-Ensai, IRMAR, UEB
laurent.rouviere@ensai.fr

[a,b,c] Campus de Ker-Lann, Rue Blaise Pascal, BP 37203
35172 Bruz cedex, France

Abstract

This paper deals with nonparametric estimation of conditional densities in mixture models. The proposed approach consists to perform a preliminary clustering algorithm 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 with regards to the performance of the clustering algorithm. In particular, we prove that these estimates achieve optimal rates over classical nonparametric density classes under mild assumptions on the clustering method used. Finally, we offer examples of clustering algorithms verifying the required assumptions.

Keywords: Nonparametric estimation, mixture models, clustering
AMS Subject Classification: 62G07, 62H30

*Corresponding author.

†Research partially supported by the French "Agence Nationale pour la Recherche" under grant ANR-09-BLAN-0051-02 "CLARA"

1 Introduction

Finite mixture models are widely used to account for population heterogeneities. In many fields such as biology, econometrics as well as 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 modeling 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

$$f(t) = \sum_{i=1}^{M} \alpha_i f_i(t), \quad t \in \mathbb{R}. \tag{1.1}$$

Here M is the number of subpopulation, 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), 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 (1.1), some issues arise. In some cases, the number of components M is unknown and therefore, needs to be estimated. To this end, some algorithms have been developed to provide consistent estimates of this parameter. For instance, Biau et al. (2007) and Cuevas et al. (2000) propose an estimator based on the level sets of f when M corresponds to the number of modes of f. Identifiability of the model is an additional issue that received some attention in the literature. Actually, model (1.1) is identifiable only with imposing restrictions on the vector $(\alpha_1, \ldots, \alpha_M, f_1, \ldots, f_M)$. In order to supply the minimal assumptions such that (1.1) becomes identifiable, Celeux and Govaert (1995), Bordes et al. (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) consider the case where appropriate covariates are available.

When the model (1.1) is identifiable, the statistical problem consists to find efficient estimates of the mixture proportions α_i and the density functions f_i. In the parametric case, some algorithms have been proposed such as maximum likelihood techniques (Redner and Walker (1984), Lindsay (1983a,b)) as well as Bayesian approaches (Diebolt and Robert (1994), Biernacki et al. (2000)). When the f_i's belong to some 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 Cerrito (1992), Titterington (1983), Hall and Titterington (1984, 1985)). 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 marginal distribution in a multivariate setting. We also refer to Bordes et al. (2006) who provide efficient estimators under the assumption that the unknown mixed distribution is symmetric. These estimates are extended by Benaglia et al. (2009, 2011) for multivariate mixture models.

The framework we consider lies 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 frow which Y is distributed. Our approach consists to perform 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 prediction of the clustering method.

Many authors have already proposed to perform a preliminary clustering step to improve density estimates. Ruzgas et al. (2006) conduct a comprehensive simulation study to conclude that a preliminary clustering using the EM algorithm allows to some extend 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 article proposes to fill this gap, studying the L_1-error of these estimates in terms of the performances of the clustering method used. In particular, we prove that these estimates achieve optimal rates over classical nonparametric density classes under mild assumptions on the clustering method used.

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. Simulations are performed in Section 4 to illustrate our theoretical results. Finally, proofs are gathered in Section 5.

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 $N_i = \#\{k \in [\![1, n]\!] : I_k = i\}$, we can estimate α_i using the empirical proportion $\bar{\alpha}_i = N_i/n$ and f_i using the kernel density estimate $\bar{f}_i$ defined for all $t \in \mathbb{R}$ by

$$\bar{f}_i(t) = \frac{1}{N_i} \sum_{k=1}^{n} K_h(t, Y_k) \mathbb{I}_i(I_k) \tag{2.1}$$

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

$$K_h(t, y) = \frac{1}{h} K\left(\frac{t - y}{h}\right) \tag{2.2}$$

is the classical convolution kernel located at point t (see Parzen (1962) and Rosenblatt (1956) for instance). Observe that $\bar{f}_i$ is just the classical 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, its 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 classes of densities such as Hölder or Lipschitz classes (see Devroye and Györfi (1985)).

Things turn out to be more complicated when the random variable I is not observed. This is typically the case in mixture density model estimation. 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 an other 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. The statistical problem with which we are faced is now to estimate both the components and the weights of the mixture based on the observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ extracted from a n-sample $(Y_1, X_1, I_1), \ldots, (Y_n, X_n, I_n)$ randomly drawn from the distribution of (Y, X, I). Our strategy consists to first guess the label I from the random variable X using a clustering algorithm. We then perform the kernel density estimate (2.1) where the true labels I_k are substituted by the predicted labels. It is described in the following section.

Remark 2.1 *Note that the distribution of (Y, X, I) is not entirely characterized. In particular, the dependence between Y and X is not specified. They could be either dependent ($Y = X$ for example) or independent. Moreover, we emphasize that only the conditional densities $g_{i,n}$ are allowed to depend on n (see Section 3 for some examples). It is not the case for α_i and f_i.*

Remark 2.2 *Observe that the clustering method is performed on the random variable X only. The underlying assumption is that one can find accurate estimates of the group I_k of Y_k using only the sample $X_1, \ldots, X_n$. Therefore, we do not need to suppose different behavior (in term of location of the distribution for instance) for the densities f_i. These kind of assumptions will be translated on the densities $g_{i,n}$ (see section 3).*

2.2 A kernel density estimate based on a clustering approach

To summarize the model so far, we observe a sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of random pairs extracted from $(Y_1, X_1, I_1), \ldots, (Y_n, X_n, I_n)$ and our goal is to estimate densities f_i of the conditional distribution $\mathcal{L}(Y|I = i)$ for all $i \in [\![1, M]\!]$. We propose the following two-step algorithm.

- First, we perform a clustering algorithm on the sample $X_1, \ldots, X_n$ to predict the label I_k of each observation X_k;
- Last, the conditional densities f_i's are estimated using kernel density estimates in each cluster.

Formally, we first 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 $i \in [\![1, M]\!]$ according to a given clustering method. The 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 but some examples are discussed in section 3. 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 $\widehat{I}_k,\ k \in [\![1, n]\!]$ as

$$\widehat{I}_k = \begin{cases} i & \text{if } X_k \in \mathcal{X}_i \\ 0 & \text{otherwise.} \end{cases}$$

Remark that if $I_k = i$ then X_k is not correctly affected to its group with probability

$$\mathbb{P}(X_k \notin \mathcal{X}_i | I_k = i) = \mathbb{P}(\widehat{I}_k \neq i | I_k = i).$$

So the maximal misclassification error

$$\varphi_n = \max_{1 \leq k \leq n} \max_{1 \leq i \leq M} \mathbb{P}(\widehat{I}_k \neq i | I_k = i) \tag{2.3}$$

measures the performance of the clustering procedure. This misclassification error φ_n will be evaluated in the examples discussed in section 3.

We are now in position to define the estimates of both the α_i's and the f_i's. The prior probabilities α_i are estimated by

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

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

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

where K_h is defined in (2.2). Observe that since for all $i \in [\![1, M]\!]$ the clusters $\mathcal{X}_i$ are nonempty, the estimates $\widehat{f}_i$ are well defined.

The 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 drawn from f_i. Under this assumption, φ_n is expected to be small and $\widehat{f}_i$ to be closed to the "ideal" estimates $\bar{f}_i$ defined by equation (2.1). The following Theorem compares the expected L_1-error of $\widehat{f}_i$ and $\bar{f}_i$. In other words, it allows to measure the effects of the clustering method with regards to the performances of $\widehat{f}_i$.

Theorem 2.1 *There exist positive constants $A_1 - A_4$ such that, for all $n \geq 1$ and $i \in [\![1, M]\!]$*

$$\mathbb{E}\left\|\widehat{f}_i - f_i\right\|_1 \leq \mathbb{E}\left\|\bar{f}_i - f_i\right\|_1 + A_1\varphi_n + A_2\exp(-n) \tag{2.4}$$

and

$$\mathbb{E}|\widehat{\alpha}_i - \alpha_i| \leq A_3\varphi_n + \frac{A_4}{\sqrt{n}}. \tag{2.5}$$

The constants $A_1 - A_4$ are specified in the proof of the theorem. We emphasize that inequalities (2.4) and (2.5) are non-asymptotic, that is, the bound are valid for all n. We can first remark that if we intend to prove any consistency results regarding $\widehat{f}_i$ and $\widehat{\alpha}_i$, the misclassification error φ_n should tend to zero. Moreover, inequality (2.4) says that if the misclassification error φ_n tends to zero much faster than the L_1-error of $\bar{f}_i$, then we have asymptotically a performance that is guaranteed to be equivalent to the performance of the "ideal" estimate $\bar{f}_i$. The L_1-error of $\bar{f}_i$, with properly chosen bandwidth h, is known to go to zero, under standard smoothness assumptions, at the rate $n^{-\frac{s}{2s+1}}$ where $s > 0$ is typically an index which represents the regularity of f_i. For example, when we consider Lipschitz or Hölder classes of functions, s corresponds to the number of absolutely continuous derivatives of the functions f_i. In this context, when $\varphi_n = \mathrm{o}(n^{-\frac{s}{2s+1}})$, the estimates $\widehat{f}_i$ achieve the rate

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

Remark 2.3 *As usual, the choice of the bandwidth h reveals crucial for the performance of the estimate. 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

Our procedure requires a preliminary clustering algorithm performed on the sample $X_1, \ldots, X_n$ drawn from the density $\sum_{i=1}^{M} \alpha_i g_{i,n}$. In this section, we study the misclassification error φ_n defined in (2.3) for two clustering methods.

3.1 A toy example

We first consider the simple case where the conditional densities $g_{i,n}$ are uniform univariate densities. Formally, we assume here that $M = 2$ and that, for all $x \in \mathbb{R}$,

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

where $(\lambda_n)_n$ 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. The predicted label $\widehat{I}_k$ is then defined according to

$$\widehat{I}_k = \begin{cases} 1 & \text{if} \quad X_k \leq 1 - \widehat{\lambda}_n \\ 0 & \text{if} \quad 1 - \widehat{\lambda}_n < X_k < 1 \\ 2 & \text{if} \quad X_k \geq 1, \end{cases} \tag{3.1}$$

see Figure 1. Many estimators of λ_n can be defined. Here we propose $\widehat{\lambda}_n = 2 - X_{(n)}$ 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).

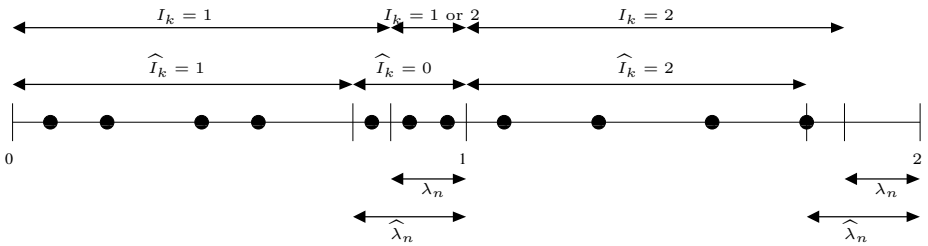

Figure 1: A sample of $n = 11$ points.

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

Proposition 3.1 *There exists a positive constant A_5 such that for all $n \geq 1$*

$$\varphi_n \leq \lambda_n + A_5 \frac{\log n}{n}.$$

Unsurprisingly, the misclassification error 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 $\lambda_n = \mathcal{O}(1/\sqrt{n})$.

3.2 A clustering procedure for support disjoint densities

In this section, we propose an automatic clustering procedure and we study its performances for support disjoint densities. More precisely, we assume that supports $S_{i,n} \subset \mathbb{R}^d$ of $g_{i,n}$ are disjoint connected compact sets and we denote by

$$\delta_n = \min_{1 \leq i \neq j \leq M} d(S_{i,n}, S_{j,n})$$

the minimum Euclidean distance between these supports (see Figure 2).

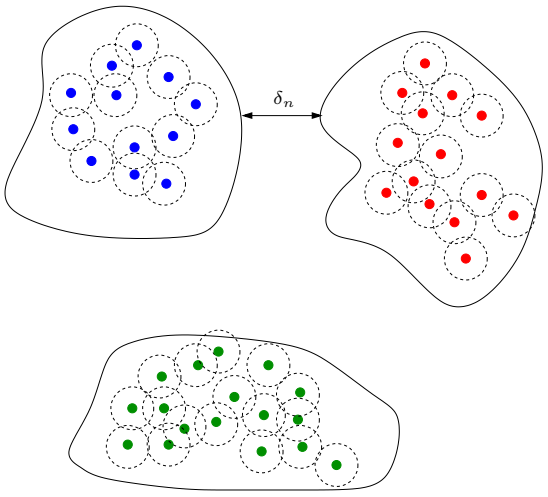

Figure 2: The support of the random variable X is divided into $M = 3$ disjoint connected compact sets. The three connected components of the graph induced by A^r are included in the connected components of the support of X.

Given n i.i.d. observations $X_1, \ldots, X_n$ drawn from $g_n(x) = \sum_{i=1}^{M} \alpha_i g_{i,n}(x)$, the goal is to split $\{X_1, \ldots, X_n\}$ into M clusters $\mathcal{X}_1, \ldots, \mathcal{X}_M$ such that the probability of the events $\{\mathcal{X}_i \subset S_{i,n}\}$ is close to one as δ_n decreases to zero. In such a situation, one can expect the misclassification error associated with this clustering algorithm to be close to zero.

3.2.1 The clustering procedure

Devroye and Wise (1980) proposed to estimate density supports by union of balls centered at each observation. Biau et al. (2008) then obtained rates of convergence for this approach. In this section, we adapt this procedure to define a new clustering algorithm. We then study the misclassification error of this method. The main idea is to find a data-driven procedure to choose a radius $\widehat{r}_n > 0$ such that the set

$$\bigcup_{k=1}^{n} B(X_k, \widehat{r}_n) \tag{3.2}$$

has exactly M connected components (see Figure 2). Here $B(x, r)$ stands for the closed Euclidean ball with center $x \in \mathbb{R}^d$ and radius $r > 0$. Cluster $\mathcal{X}_i$ will then be naturally composed by observations X_k which belong to the i^{th} connected component of the set (3.2). We need to find on optimal way to select $\widehat{r}_n$ from the observations $X_1, \ldots, X_n$.

To this aim, we define for each positive real number r the $n \times n$ matrix $A^r = (A^r_{k,\ell})_{1 \leq k,\ell \leq n}$ by

$$A^r_{k,\ell} = \begin{cases} 1 & \text{if } \|X_k - X_\ell\|_2 \leq 2r \iff B(X_k, r) \cap B(X_\ell, r) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

This matrix induces a non-orientated graph on the set $[\![1, n]\!]$ and we say that 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 $\mathcal{X}_1(r), \ldots, \mathcal{X}_{\widehat{M}_r}(r)$ the associated clusters. The radius is selected as follows

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

Note that $\widehat{r}_n$ is well defined since the random set $\mathcal{R}_M = \inf\{r > 0 : \widehat{M}_r \leq M\}$ 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, thanks to Lemma 5.2 one can easily prove that $\widehat{r}_n = \min \mathcal{R}_M$ and $\widehat{M}_{\widehat{r}_n} = M$ almost surely when $n \geq M$. We denote by $\mathcal{X}_1(\widehat{r}_n), \ldots, \mathcal{X}_M(\widehat{r}_n)$ the M clusters induced by $A^{\widehat{r}_n}$ and the goal is now to study the misclassification error (2.3) of this clustering algorithm.

Remark 3.1 *For a given value $r > 0$, clusters $\mathcal{X}_1(r), \ldots, \mathcal{X}_{\widehat{M}_r}(r)$ are arbitrarily indexed. Even if the indices of the clusters are not really important in practice, things are getting more complicated when we study the misclassification error*

$$\varphi_n = \max_{i=1,\ldots,M} \max_{k=1,\ldots,n} \mathbb{P}(\widehat{I}_k \neq i | I_k = i)$$

of the predicted rule

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

The way to index the clusters is clearly important in this context. To circumvent this problem we will study the misclassification error up to a permutation of the indices.

Remark 3.2 *Our algorithm requires to compute the connected components of the graph induced by the $n \times n$ matrix A^r 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 et al. (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 Basic assumptions

To study the misclassification error of the proposed clustering algorithm, we need the following assumptions.

Assumption 1 The density $g_n(x) = \sum_{i=1}^n \alpha_i g_{i,n}(x)$ is bounded away from zero on its support. We define the sequence $(t_n)_n$ by

$$t_n = \inf_{x \in S_n} g_n(x) > 0 \quad \text{where} \quad S_n = \bigcup_{i=1}^M S_{i,n}.$$

Assumption 2 Let

$$r_n^d = \frac{(\log n)^2}{n t_n}.$$

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

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

where Leb denotes the Lebesgue measure on $\mathbb{R}^d$.

Assumption 1 implies that densities $g_{i,n}$ do not vanish on the interior of the supports $S_{i,n}$. This assumption ensures that, for n large enough and a safe choice of radius r, all points $X_k \in S_{i,n}$ belong to the same connected component of the graph induced by the $n \times n$ matrix A^r defined in (3.3). Assumption 2 is more technical and is concerned with the diameter and the regularity of supports $S_{i,n}$. Roughly speaking, recall that our approach identify the supports $S_{i,n}$ by the connected components of $\bigcup_{k=1}^n B(X_k, r)$. It means that when the diameter of $S_{i,n}$ increases, large values or 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

$$N \leq \frac{n}{c_1 c_2 (\log n)^2}. \tag{3.4}$$

Finally, inequality $\text{Leb}(S_n \cap B_\ell) \geq c_2 r_n^d$ in assumption 2 can be seen as a regularity constraint on S_n. In particular, it allows to avoid too sharp boundaries for the support S_n.

Remark 3.3 *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 larger 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 arbitrarily large, the diameter of S_n could be arbitrarily large and assumption 2 does not hold. This assumption constraints to some extent the shape of S_n. It is satisfied for regular supports such that the diameter does not increase too fast 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 supports S_n that do not depend on the sample size n with smooth boundaries (see Biau et al. (2007)).

3.2.3 The misclassification error

The algorithm described in section 3.2.1 gives a partition of $\{X_1, \ldots, X_n\}$ into M clusters $\mathcal{X}_1(\widehat{r}_n), \ldots, \mathcal{X}_M(\widehat{r}_n)$. The following theorem provides an upper bound for the misclassification error φ_n of this clustering procedure.

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

$$\delta_n > 2r_n = 2\left(\frac{(\log n)^2}{nt_n}\right)^{1/d} \tag{3.5}$$

then, after a possible rearrangement of the indices, we can define $\widehat{I}_k = i \Longleftrightarrow X_k \in \mathcal{X}_i(\widehat{r}_n)$ *such that for all* $a > 0$ *with* $\log n \geq (1+a)/c_2$*, we have*

$$\varphi_n = \max_{i=1,\ldots,M} \max_{k=1,\ldots,n} \mathbb{P}(\widehat{I}_k \neq i | I_k = i) \leq A_6 n^{-a} \tag{3.6}$$

where A_6 *is positive constant.*

This theorem establishes that the misclassification error of the proposed clustering procedure tends to zero at a polynomial rate. In particular, using Borel-Cantelli lemma, it implies that $\widehat{I}_k = I_k$ almost surely for n large enough, *i.e.*, all the predictions are correct for n large enough. Inequality (3.5) gives the minimum distance between the supports $S_{i,n}$ to make the clustering method efficient. Observe that this minimum distance could tend to zero as the sample size increases. For example, when $d = 1$ and $g_n(x) \geq c/n^\gamma$ with $\gamma \in]0, 1[$, inequality (3.6) holds provided δ_n tends to zero much slower than $(\log n)^2/n^{1-\gamma}$.

Remark 3.4 *Even if the rate of convergence (3.6) does not depend on the dimension d, the curse of dimensionality appears in Theorem 3.1. Indeed, the minimum distance* δ_n *clearly increases with the dimension d. Consequently, even if we can obtain fast rates in large dimension, the assumption (3.5) becomes stronger as the dimension d increases.*

4 Simulation study

In this section, we illustrate the theory with simulation results enlightening the efficiency of the proposed estimator. Theorem 2.1 says that the proposed estimates $\widehat{f}_i$ perform well provided the misclassification error of the clustering procedure is small with respect to the L_1-error of the ideal estimates $\bar{f}_i$. In practice, any clustering method could be used. In the simulation study, we choose the clustering procedure proposed in section 3.2.

4.1 Comparison with the ideal estimates $\bar{f}_i$

We first compare the performances of the two step estimates $\widehat{f}_i$ with the ideal estimates $\bar{f}_i$ defined by (2.1). The model is as follows. The 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 -1 and 1 and variance 1. The conditional densities $g_{i,n}, i = 1, 2$ are uniform densities on $\mathbb{R}^d$ defined as follows:

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

where $\delta_n > 0$ measures the distance between the supports of $g_{1,n}$ and $g_{2,n}$. We apply the clustering procedure proposed in section 3.2 and we denote by $\widehat{f}_1$ and $\widehat{f}_2$ the resulting estimates of f_1 and f_2. To compute the nonparametric estimates $\bar{f}_i$ and $\widehat{f}_i$, we use a gaussian kernel. Recall that this paper does not conduct any theory to select the bandwidth h in an optimal way (see remark 2.3). We propose to use the default data-driven procedure proposed in the GNU-R library **np** (see Hayfield and Racine (2008)).

We compare the L_1 performances of $\bar{f}_i$ with the L_1 performances of $\widehat{f}_i$ for many values of sample sizes n, distances δ_n and dimensions d. In this setting, Theorem 3.1 ensures that estimates $\widehat{f}_i$ perform well provided

$$\delta_n > 2\left(8\frac{(\log(n))^2}{n}\right)^{1/d}.$$

However, a deeper analysis of the proof of this Theorem shows that the term $(\log(n))^2$ in the lower bound can be replaced by $(\log(n))^\alpha$ with $\alpha > 1$. For the simulations, we set

$$\delta_n = 2\left(8\frac{(\log(n))^\alpha}{n}\right)^{1/d} \tag{4.1}$$

for 6 values of α equally spaced between 1 and 2. For the sake of clarity, we only present results regarding the density f_1 since conclusions are the same for f_2. Tables 1, 2, 3 and 4 presents the ratio

$$\frac{\mathbb{E}\|\widehat{f}_1 - f_1\|_1}{\mathbb{E}\|\bar{f}_1 - f_1\|_1} \tag{4.2}$$

for various dimensions d and sample sizes n. The expectations are evaluated over 200 Monte Carlo replications.

	$\alpha = 1$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$d = 1$	4.079	4.079	3.874	3.670	3.470	1.000
$d = 2$	5.015	4.920	5.095	5.000	1.020	1.000
$d = 3$	5.366	5.370	5.157	1.745	1.000	1.000
$d = 4$	5.273	5.206	4.502	1.353	1.000	1.000

Table 1: $n = 50$

	$\alpha = 1$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$d = 1$	3.001	2.791	2.599	1.594	1.000	1.000
$d = 2$	3.590	3.467	1.576	1.000	1.000	1.000
$d = 3$	3.693	2.791	1.255	1.000	1.000	1.000
$d = 4$	3.557	2.094	1.193	1.000	1.000	1.000

Table 2: $n = 100$

	$\alpha = 1$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$d = 1$	2.164	2.012	1.112	1.000	1.000	1.000
$d = 2$	2.508	1.192	1.000	1.000	1.000	1.000
$d = 3$	1.633	1.124	1.000	1.000	1.000	1.000
$d = 4$	1.587	1.060	1.000	1.000	1.000	1.000

Table 3: $n = 200$

	$\alpha = 1$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$d = 1$	1.498	1.032	1.002	1.000	1.000	1.000
$d = 2$	1.054	1.000	1.000	1.000	1.000	1.000
$d = 3$	1.054	1.000	1.000	1.000	1.000	1.000
$d = 4$	1.086	1.030	1.000	1.000	1.000	1.000

Table 4: $n = 500$

The more this ratio is close to 1, the better the estimator. These tables show that the performance of our procedure are significantly better when the number of observations increases: for $n = 500$ L_1-errors of $\widehat{f}_i$ and $\bar{f}_i$ are rouglhy similar. Moreover, for large enough separation distances ($\alpha = 2$), performances are exactly the same for all n. Even if, for fixed values of α,

the ratio (4.2) are approximately the same, observe that the proposed procedure is affected by the curse of dimensionality since the minimum distance between the supports of $g_{1,n}$ and $g_{2,n}$ clearly increases with the dimension (see remark 3.4). Finally, Figure 3 displays the boxplots of the L_1-error of $\bar{f}_1$ and $\widehat{f}_1$ for 200 replications of three models with the following values: $n = 100$, $d = 2$ and $\alpha \in \{1, 1.4, 1.8\}$. As expected, the performances of $\widehat{f}_1$ come closer to those of $\bar{f}_1$ as α increases. They are exactly the same for $\alpha = 1.8$ since in that case the predicted labels $\widehat{I}_k$ correspond with the true labels $I_k, k = 1, \ldots, n$.

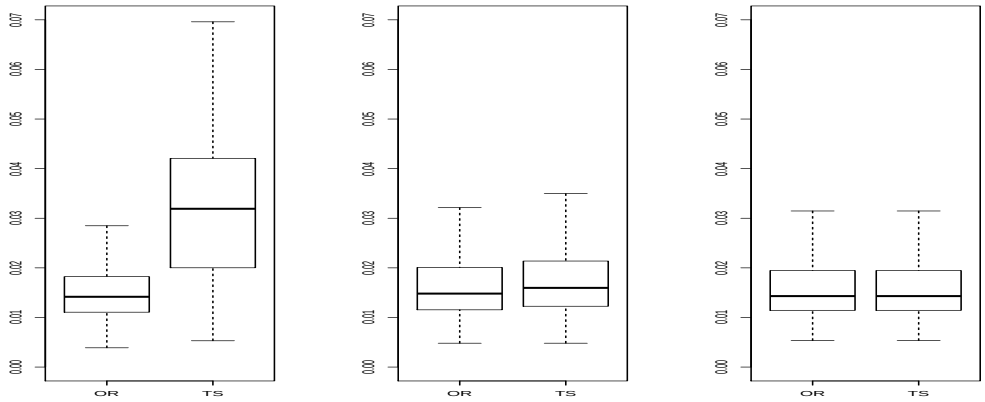

Figure 3: Boxplot of the L_1-error for the "oracle" estimator $\bar{f}_1$ (OR) and the two step estimator $\widehat{f}_1$ (TS). The separation distance between g_1 and g_2 corresponds to $\alpha = 1$ (left), 1.4 (middle) and 1.8 (right)

4.2 A comparison with the EM algorithm

We finally illustrate the performance of the proposed method with a comparison with the well known EM-algorithm (Dempster et al. (1977)). The density of Y is now 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. Thus, the parameter Δ measures the separation between the components f_1 and f_2 (see Figure 4).

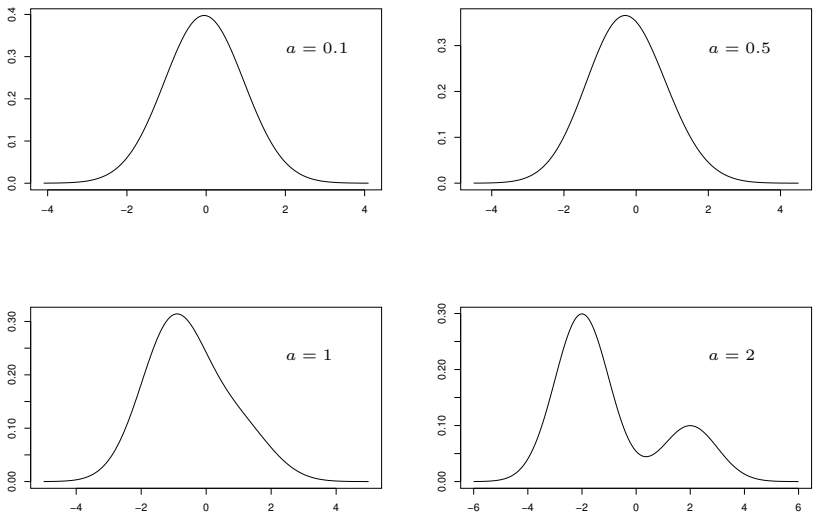

Figure 4: Density of Y for various values of Δ.

The 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 $\delta_n > 0$ still measures the distance between $g_{1,n}$ and $g_{2,n}$. Here $n = 200$ and δ_n takes the same values as in the previous section (see equation (4.1)). We compare the performance of our method with the EM algorithm performed on the sample $Y_1, \ldots, Y_n$. Formally, we run the EM algorithm (with equal variances) to estimate the gaussian parameters of the components 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. We keep the same setting as above to compute our estimates $\widehat{f}_1$ and $\widehat{f}_2$: clustering procedure of section 3.2, gaussian kernel and bandwidth selected with the library **np**. For the sake of clarity, we again only present results on $\widehat{f}_1$ and f_1^{em} since we observe the same conclusions for $\widehat{f}_2$ and f_2^{em}. Table 5 presents the ratio

$$\frac{\mathbb{E}\|\widehat{f}_1 - f_1\|_1}{\mathbb{E}\|f_1^{em} - f_1\|_1}, \tag{4.3}$$

where the expectations are evaluated over 200 monte carlo replications. Figure 5 displays the boxplots of the L_1-error for each estimate over the 200 replications.

	$\alpha = 1$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\Delta = 0.1$	0.276	0.272	0.299	0.285	0.275	0.272
$\Delta = 0.5$	0.423	0.413	0.429	0.416	0.423	0.407
$\Delta = 1$	0.863	0.828	0.853	0.900	0.840	0.888
$\Delta = 2$	1.725	1.611	1.675	1.778	1.589	1.645

Table 5: Ratio (4.3) evaluated over 200 replications.

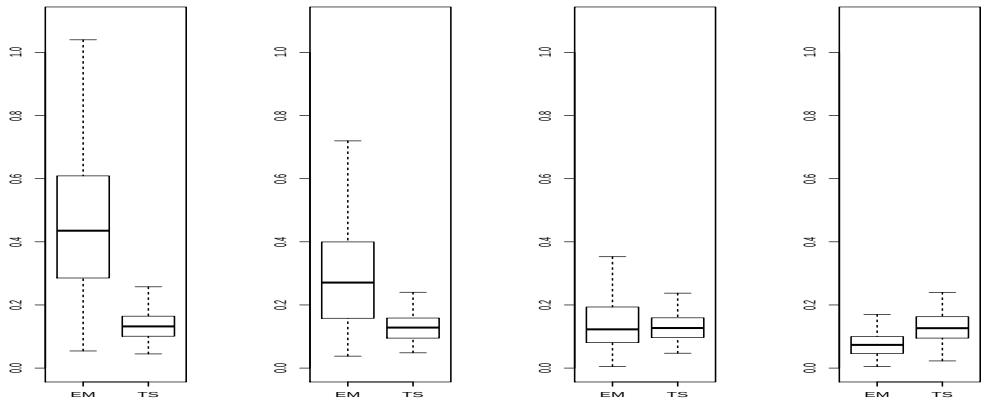

Figure 5: Boxplot of the L_1 error for the EM algorithm (EM) and the two step estimator proposed in the paper (TS). The separation distance Δ between f_1 and f_2 vary from 0.1 (left) to 2 (right) and $\alpha = 1.4$.

As expected, we first observe that the performances of the EM algorithm clearly depend on the separation distance between the target densities f_1 and f_2. For large values of Δ, parametric estimates resulting from the EM algorithm overperform the nonparametric estimates proposed in this paper (e.g. $\Delta = 2$ in Figure 5). This is no longer the case when f_1 is closed to f_2: the results clearly show the advantages of using an auxiliary information (represented by the covariate X) when the components f_1 and f_2 are not enough separated. Indeed, the L_1 performances of $\widehat{f}_1$ over f_1^{em} are significantly better for $\Delta = 0.1$ and $\Delta = 0.5$ and roughly similar for $\Delta = 1$. Finally, observe that Figure 5 shows that the L_1-error of $\widehat{f}_1$ does not depend on Δ.

5 Proofs

5.1 Proof of Theorem 2.1

First, let us prove inequality (2.4). Inserting the "ideal" estimator $\bar{f}_i$ and using triangle inequality, the L_1-risk of $\widehat{f}_i$ can be bounded as follows

$$\mathbb{E}\left\|\widehat{f}_i - f_i\right\|_1 \leq \mathbb{E}\left\|\bar{f}_i - f_i\right\|_1 + \mathbb{E}\left\|\widehat{f}_i - \bar{f}_i\right\|_1.$$

To prove our result, we just have to control the second term in the right-hand side of the previous inequality. Since $\widehat{f}_i = 0$ when $N_i = 0$ and noting that $\|\widehat{f}_i\|_1 = 1$ we obtain

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

For the sake of readability, let $\widetilde{\mathbb{E}}$ denotes 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 us define

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

Using these notations it is easily seen that

$$\mathbb{E}\left\|(\widehat{f}_i - \bar{f}_i)\mathbb{I}_{N_i>0}\right\|_1 = \mathbb{E}\widetilde{\mathbb{E}}\int_{\mathbb{R}} \widetilde{\widetilde{\mathbb{E}}}|A_i(t)|dt. \tag{5.1}$$

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

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

Thus

$$\widetilde{\mathbb{E}}\int_{\mathbb{R}} \widetilde{\widetilde{\mathbb{E}}}|A_i(t)|\, dt \leq \frac{\|K\|_1}{N_i}\widetilde{\mathbb{E}}\left[\frac{1}{\widehat{N}_i}\sum_{k=1}^{n}|N_i\mathbb{I}_{\{i\}}(\widehat{I}_k) - \widehat{N}_i\mathbb{I}_{\{i\}}(I_k)|\right]. \tag{5.2}$$

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

$$\begin{aligned}
&\widetilde{\mathbb{E}}\left[\frac{1}{\widehat{N}_i}\sum_{k=1}^{n}|N_i\mathbb{I}_{\{i\}}(\widehat{I}_k)-\widehat{N}_i\mathbb{I}_{\{i\}}(I_k)|\right]\\
&\qquad\leq \widetilde{\mathbb{E}}|N_i-\widehat{N}_i|+\widetilde{\mathbb{E}}\sum_{k=1}^{n}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\\
&\qquad\leq 2\widetilde{\mathbb{E}}\sum_{k=1}^{n}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|. \qquad (5.3)
\end{aligned}$$

Combining (5.1), (5.2) and (5.3) leads to

$$\begin{aligned}
\mathbb{E}\left\|(\widehat{f}_i-\bar{f}_i)\mathbb{I}_{N_i>0}\right\|_1 &\leq 2\|K\|_1\sum_{k=1}^{n}\mathbb{E}\left[\frac{\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\right]\\
&\leq \frac{2\|K\|_1}{n\alpha_i}\sum_{k=1}^{n}\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\right]. \qquad (5.4)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\right]\\
&\qquad\leq \mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}\leq 2}\right]\\
&\qquad\quad+\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right]. \qquad (5.5)
\end{aligned}$$

For the first term of this bound, we deduce from Lemma 5.1 that

$$\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}\leq 2}\right]\leq 2\varphi_n(1+(M-2)\alpha_i). \qquad (5.6)$$

For the second term, using Hölder inequality we obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right]\\
&\leq\sqrt{\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k)-\mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right]^2\mathbb{P}\left(\frac{n\alpha_i}{N_i}>2\right)}\\
&\leq\sqrt{\mathbb{E}\left(\frac{(n\alpha_i)^2}{N_i^2}\mathbb{I}_{N_i>0}\right)\mathbb{P}\left(N_i-n\alpha_i<-\frac{n\alpha_i}{2}\right)}. \qquad (5.7)
\end{aligned}$$

Now, it is easily seen that

$$\mathbb{E}\left(\frac{(n\alpha_i)^2}{N_i^2}\mathbb{I}_{N_i>0}\right) \leq 6\mathbb{E}\left(\frac{(n\alpha_i)^2}{(N_i+1)(N_i+2)}\right) \leq 6, \tag{5.8}$$

where last inequality follows from Hengartner and Matzner-Løber (2009). Using Hoeffding's inequality (see Hoeffding (1963)) we obtain for the second term in (5.5)

$$\mathbb{E}\left[\frac{n\alpha_i\mathbb{I}_{N_i>0}}{N_i}|\mathbb{I}_{\{i\}}(\widehat{I}_k) - \mathbb{I}_{\{i\}}(I_k)|\mathbb{I}_{\frac{n\alpha_i}{N_i}>2}\right] \leq \sqrt{6}\exp\left(-\frac{n{\alpha_i}^2}{4}\right). \tag{5.9}$$

From (5.4) – (5.9), we deduce that

$$\begin{aligned}\mathbb{E}\left\|(\widehat{f}_i - \bar{f}_i)\mathbb{I}_{N_i>0}\right\|_1 \leq & \frac{4\|K\|_1}{\alpha_i}(1+(M-2)\alpha_i)\varphi_n \\ & + \frac{2\sqrt{6}\|K\|_1}{\alpha_i}\exp\left(-\frac{n{\alpha_i}^2}{4}\right).\end{aligned}$$

Putting all pieces together, we finally obtain

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

which concludes the first part of the proof.
Inequality (2.5) is a direct consequence of Lemma 5.1:

$$\begin{aligned}\mathbb{E}|\widehat{\alpha}_i - \alpha_i| &\leq \mathbb{E}\left|\frac{\widehat{N}_i}{n} - \frac{N_i}{n}\right| + \mathbb{E}\left|\frac{N_i}{n} - \alpha_i\right| \\ &\leq \frac{1}{n}\sum_{k=1}^{n}|\mathbb{I}_{\{i\}}(\widehat{I}_k) - \mathbb{I}_{\{i\}}(I_k)| + \frac{1}{n}\sqrt{\mathbb{V}(N_i)} \\ &\leq (1+(M-2)\alpha_i)\varphi_n + \sqrt{\frac{\alpha_i(1-\alpha_i)}{n}}.\end{aligned}$$

5.2 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, let us first consider the case $i = 2$. Observe that

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

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

$$\mathbb{P}(\widehat{I}_k \neq 2 | I_k = 2) = \int_{1-\lambda_n}^{1} g_{2,n}(x)dx = \lambda_n. \tag{5.10}$$

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

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

This leads to the following inequality

$$\mathbb{P}(\widehat{I}_k \neq 1 | I_k = 1) \leq \mu_n + \mathbb{P}\left(X_{(n)} \leq 2 - \mu_n | I_k = 1\right). \tag{5.11}$$

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

$$\begin{aligned}\mathbb{P}\left(X_{(n)} \leq 2 - \mu_n | I_k = 1\right) &= \left(\mathbb{P}(X_1 \leq 2 - \mu_n)\right)^{n-1} \\ &= (1 - 2n^{-1}(\log n))^{n-1} \\ &\leq n^{-1},\end{aligned} \tag{5.12}$$

where the last inequality follows, for $n \geq 2$, from the fact that $1 - u \leq e^{-u}$ for all $u \geq 0$ and is still valid for $n = 1$.
Taking together equations (5.11) and (5.12), we finally obtain

$$\mathbb{P}(\widehat{I}_k \neq 1 | I_k = 1) \leq \lambda_n + n^{-1} + \frac{2}{\alpha_2} \cdot \frac{\log n}{n}. \tag{5.13}$$

Proposition follows from equations (5.10) and (5.13).

5.3 Proof of Theorem 3.1

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

$$\left(\bigcup_{k:I_k=i} B(X_k, r_n)\right) \cap \left(\bigcup_{k:I_k=j} B(X_k, r_n)\right) \subseteq (S_{i,n}+r_n) \cap (S_{j,n}+r_n) = \emptyset, \quad (5.14)$$

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

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

Inclusion (5.14) implies $\widehat{M}_{r_n} \geq M$. Remark that if

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

then $\widehat{M}_{r_n} = \widehat{M}_{\widehat{r}_n} = M$ and $\widehat{r}_n \leq r_n$. In that case, the matrices A^{r_n} and $A^{\widehat{r}_n}$ defined in (3.3) induce the same clusters $\mathcal{X}_1, \ldots, \mathcal{X}_M$. Moreover, it is easy to see that, up to a permutation of the indices of the clusters, we have $\widehat{I}_k = I_k$ almost surely for all $k \in [\![1,n]\!]$. It follows that

$$\max_{k=1,\ldots,n} \mathbb{P}(\widehat{I}_k \neq I_k) \leq \mathbb{P}(r_n \notin \mathcal{R}_M).$$

Thus to complete the proof of the result, we have to find an upper bound of the probability of the event $\{r_n \notin \mathcal{R}_M\}$. To this aim, observe that if

$$S_n \subseteq \bigcup_{k=1}^{n} B(X_k, r_n) \quad (5.15)$$

then $\widehat{M}_{r_n} = M$ and $r_n \in \mathcal{R}_M$. Moreover, inclusion (5.15) holds when for all $\ell \in [\![1,N]\!]$, the balls B_ℓ defined in assumption 2 contain at least one observation among $\{X_k, k = 1, \ldots, n\}$. Therefore

$$\begin{aligned}
\mathbb{P}\left(r_n \notin \mathcal{R}_M\right) &\leq \mathbb{P}\left(\exists \ell \in [\![1,N]\!], \forall k \in [\![1,n]\!], X_k \notin B_\ell\right) \\
&\leq \sum_{\ell=1}^{N} \left(\mathbb{P}\left(X_1 \notin B_\ell\right)\right)^n \\
&\leq \sum_{\ell=1}^{N} \left(1 - t_n \operatorname{Leb}(S_n \cap B_\ell)\right)^n
\end{aligned}$$

from assumption 1. According to assumption 2 and inequality (3.4), we obtain

$$\begin{aligned}
\mathbb{P}\left(r_n \notin \mathcal{R}_M\right) &\leq N\left(1 - c_2 t_n r_n^d)\right)^n \\
&\leq (c_1 c_2)^{-1} \frac{n}{(\log n)^2} \exp(-c_2 n t_n r_n^d) \\
&\leq (c_1 c_2)^{-1} \frac{n}{(\log n)^2} \exp(-c_2 (\log n)^2).
\end{aligned}$$

Thus, as soon as n is such that $c_2 \log n \geq 1 + a$ we have

$$\mathbb{P}\left(r_n \notin \mathcal{R}_M\right) = (c_1 c_2)^{-1} \frac{1}{n^a (\log n)^2}.$$

We finish the proof with

$$\varphi_n \leq \frac{1}{\min_{1 \leq i \leq M} \alpha_i} \max_{k=1,\ldots,n} \mathbb{P}(\widehat{I}_k \neq I_k) \leq \frac{1}{\min_{1 \leq i \leq M} \alpha_i} (c_1 c_2)^{-1} \frac{1}{n^a (\log n)^2}.$$

5.4 Technical lemmas

We have the following lemmas.

Lemma 5.1 *We have for $i = 1, \ldots, M$*

$$\mathbb{E}\left|\mathbb{I}_{\{i\}}(\widehat{I}_k) - \mathbb{I}_{\{i\}}(I_k)\right| \leq \varphi_n(1 + (M-2)\alpha_i).$$

Proof. The statement follows from the chain of inequalities

$$\begin{aligned}
\mathbb{E}\left|\mathbb{I}_{\{i\}}(\widehat{I}_k) - \mathbb{I}_{\{i\}}(I_k)\right| &\leq \sum_{j \neq i} \mathbb{E}(\mathbb{I}_{\{i\}}(\widehat{I}_k)\mathbb{I}_{\{j\}}(I_k) + \mathbb{I}_{\{j\}}(\widehat{I}_k)\mathbb{I}_{\{i\}}(I_k)) \\
&\leq \sum_{j \neq i} [\alpha_j \varphi_n + \alpha_i \varphi_n] \\
&\leq \varphi_n(1 + (M-2)\alpha_i).
\end{aligned}$$

Lemma 5.2 *Almost surely, the function $r \mapsto \widehat{M}_r$ is non-increasing and right-continuous.*

Proof. We first prove that $\widehat{M} : r \mapsto \widehat{M}_r$ is right-continuous. For $r > 0$ let

$$\mathcal{K}_r = \left\{(k, \ell) \in [\![1, n]\!]^2 : B(X_k, r) \cap B(X_\ell, r) = \emptyset\right\}.$$

For all (k, ℓ) in $\mathcal{K}_r$, there exists $h_{k,\ell}$ such that

$$B(X_k, r + h_{k,\ell}) \cap B(X_\ell, r + h_{k,\ell}) = \emptyset.$$

Moreover last inequality is still true when we replace $h_{k,\ell}$ by h such that $h \leq h_r^* = \min_{(k,\ell) \in \mathcal{K}_r} h_{k,\ell}$. This implies that the connected components induced by A^r and A^{r+h} are the same and thus $\widehat{M}_{r+h} = \widehat{M}_r$. To prove that $\widehat{M}$ is non-increasing it is sufficient to observe that $\mathcal{K}_r \subseteq \mathcal{K}_{r'}$ for $r \geq r'$.

References

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

T. Benaglia, D. Chauveau, and D. R. Hunter. Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger*, pages 15–27. World Scientific Publishing Co., 2011.

A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'ISUP*, 38(3), 1994.

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

G. Biau, B. Cadre, and B. Pelletier. Exact rates in density support estimation. *Journal of Multivariate Analysis*, 99(10):2185–2207, 2008.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725, 2000.

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

G. Celeux and G. Govaert. Parsimonous gaussian models in cluster analysis. *Pattern Recognition*, 28:781–793, 1995.

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

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

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

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

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

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

L. Devroye and G. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38: 480–488, 1980.

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

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

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

P. Hall and D. M. Titterington. 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, 1985.

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

T Hayfield and J. S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

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

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

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

Y. Kitamura. Nonparametric identifiability of finite mixtures. Yale University, 2004.

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

B. G. Lindsay. The geometry of mixture likelihoods. ii. the exponential family. *The Annals of Statistics*, 11:783–792, 1983b.

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

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

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

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

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

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

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