

Semiparametric mixtures of symmetric distributions

Cristina Butucea[†] and Pierre Vandekerkhove^{†‡}

[†]*Université Paris-Est*

LAMA (UMR 8050), UPEMLV

F-77454, Marne-la-Vallée, France

and

[‡]*UMI Georgia Tech - CNRS 2958,*

George W. Woodruff School of Mechanical Engineering

Georgia Institute of Technology

January 17, 2013

Abstract

We consider in this paper the semiparametric mixture of two unknown distributions equal up to a shift parameter. The model is said to be semiparametric in the sense that the mixed distribution is not supposed to belong to a parametric family. In order to insure the identifiability of the model it is assumed that the mixed distribution is zero-symmetric, the model being then defined by the mixing proportion, two location parameters, and the probability density function of the mixed distribution. We propose a new class of M -estimators of these parameters based on a Fourier approach, and prove that they are \sqrt{n} -consistent under mild regularity conditions. Their finite-sample properties are illustrated by a Monte Carlo study and a benchmark real dataset is also studied with our method.

AMS 2000 subject classifications. Primary 62G05, 62G20; secondary 62E10.

Key words and phrases. Asymptotic normality, consistency, contrast estimators, Fourier transform, identifiability, inverse problem, semiparametric, two-component mixture model.

1 Introduction

The probability density functions (pdf) of d -variate multicomponent mixture models are defined by

$$g(x) = \sum_{i=1}^k \lambda_i f_i(x), \quad x \in \mathbb{R}^d, \quad (1)$$

where the unknown proportions λ_i ($\lambda_i \geq 0$ and $\sum_{i=1}^k \lambda_i = 1$) and the unknown pdf f_i are to be estimated. Generally the f_i 's are supposed to belong to a parametric family of density functions turning the inference problem for model (1) into a purely parametric estimation problem. There exists an extensive literature on this subject including the monographs of Everitt and Hand (1981), Titterington *et al.* (1985) or McLachlan and Peel (2000), which provide a good overview of the existing methods in this case such as maximum likelihood, minimum chi-square, moments method, Bayesian approaches etc. Note that the estimation of the number of components k in model (1) may also be a crucial issue leading to various rates of convergence for maximum likelihood estimators, as discussed by Chen (1995). In that case, the selection model is an important topic, see for example Dacunha-Castelle & Gassiat (1999), Lemdani & Pons (1999), and Leroux (1992). In addition the choice of a parametric family for the f_i 's may be difficult when few informations are known from each subpopulations. However, model (1) is generally nonparametrically nonidentifiable without additionnal assumptions. This is no longer true when training data are available from each subpopulation; see for example Cerrito (1992), Hall (1981), Lancaster & Imbens (1996), Murray & Titterington (1978), and Qin (1999). Hall & Zhou (2003) first considered the case where no parametric assumptions are made about the f_i 's involved in model (1). These authors looked at d -variate mixtures of two distributions, each having independent components, and proved that, under mild regularity conditions, their model is identifiable when $d \geq 3$. They propose in addition \sqrt{n} -consistent estimators of the $2d$ univariate marginal cumulative distribution functions and the mixing proportion. Even if model (1) is not nonparametrically identifiable, there exists for $d = 1$ and $k \geq 2$, many real data sets in the statistical literature for which such a model is used under parametric assumptions on the f_i 's, such as the Old Faithful dataset, see Azzalini & Bowman (1990), which corresponds to time measurement (in minute) between eruptions of the Old Faithful geyser in Yellowstone National Park, USA. Another famous example deals with average amounts of precipitation (rainfall) in inches for United States cities (from the Statistical

abstract of the United States, 1975; see McNeil (1977). These data sets are both included in the R statistical package.

To model from a semiparametric point of view this type of data ($d = 1$ and $k = 2, 3$), Bordes, Mottelet & Vandekerckhove (2006) (in abreviate BMV) and Hunter, Wang & Hettmansperger (2007) (in abreviate HWH) proposed jointly to consider i.i.d. sample data (X_1, \dots, X_n) drawn from a common pdf g satisfying

$$g(x) = \sum_{i=1}^k \lambda_i f(x - \mu_i), \quad x \in \mathbb{R}, \quad (2)$$

where $\mu_i \in \mathbb{R}$, $\lambda_i \geq 0$ for all $i \in \{1, \dots, k\}$ such that $\sum_{i=1}^k \lambda_i = 1$ and f is an unknown pdf. When f is supposed to be zero-symmetric, that is $f(x) = f(-x)$ for all $x \in \mathbb{R}$, the above authors proposed M -estimation methods based on the cumulative distribution function (cdf) in order to estimate separately the Euclidean and functional part of model (2). The crucial part of their work deals with the identifiability of model (2) under the simple symmetry assumption on f . Their basic results are established in BMV, Theorem 2.1 and HWH, Theorem 1, 2 and Corollary 1. The mixed density g in (2) can also be seen as the density of i.i.d. observations X_i in a convolution model:

$$X_i = Z_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (3)$$

where Z_i 's are i.i.d. with common pdf f and independent of i.i.d. errors ε_i 's with discrete law such that $P(\varepsilon = \mu_i) = \lambda_i$, for $i = 1, \dots, k$. Previous results mean that, if k is known and f is supposed to be zero-symmetric, then we can identify the law of the errors and estimate nonparametrically the pdf f . Let us notice that the mixture problem in (2) and the deconvolution problem in (3) are the same. They are both an inverse problem with unknown operator (i.e. convolution with an unknown law having support on k unknown points).

In particular when $k = 2$, $\lambda_1 := p_0$ and $(\mu_1, \mu_2) := (\alpha_0, \beta_0)$, according to Theorem 2.1 in BMV, such a model is identifiable if the Euclidean parameter $\theta_0 := (p_0, \alpha_0, \beta_0) \in [0, 1/2) \times \mathbb{R}^2 \setminus \Delta$, where $\Delta = \{(x, x); x \in \mathbb{R}\}$ and the mixed density f is zero-symmetric. When $k = 2$, BMV prove, under mild conditions, that both the Euclidean parameter and the cumulative distribution function of f of model (2) are estimated almost surely at the rate $n^{-1/4+\alpha}$, for all $\alpha > 0$ (see Theorem 3.3 and 3.4). When $k = 2$ or 3, HWH prove under mild conditions, the strong consistency of their estimator, and establish, under very technical conditions, its asymptotic normality (see Theorems 3 and 4 therein).

In this paper we propose to investigate a new estimation method. Let us first recall that BMV propose an iterative procedure to invert the mixture operator and a contrast based on this inversion step which implies the cdf G and the symmetry of the underlying unknown pdf f . HWH introduce another contrast based on the cdf of the observations G and estimate the Euclidean parameter using the symmetry property of the unknown pdf f as well. Here, we use Fourier analysis to invert the mixture operator and see that, under identifiability assumptions, the inverse problem is well posed. Then we construct a contrast based on characteristic functions of our data which allows to estimate θ when f is zero-symmetric. This contrast is a functional of the pdf g of our observations which is estimated by a U -statistic of order 2 at parametric rate. Our procedure is easier to deal with and allows to get a central limit theorem for the estimator of θ under much simpler conditions than those of Theorem 4 in HWH. Moreover, we define a kernel estimator of the pdf f and prove that it attains the same nonparametric rate as in the direct problem of density estimation. The inverse problem does not affect the pointwise rate of convergence of the density estimator. Our estimators and convergence results generalize to the mixture model with $k \geq 3$ components, as soon as the model verifies identifiability assumptions. Such assumptions are known for $k = 3$ only, see Corollary 1 in HWH. On the other hand our estimating method can be adapted, as it will be explained in remark at the end of Section 3, to the 2-component multivariate symmetric case ($k = 2$, $d > 1$) with a rate of convergence for f depending on d .

The paper is organized as follows: in Section 2 we propose a contrast function based on a Fourier transform of the pdf g of our observations and derive our estimation method; in Section 3 we present our main asymptotic result which concern the \sqrt{n} -rate of convergence for the Euclidean part of the parameter and show that the classical nonparametric rate of convergence is achieved for our inverse Fourier nonparametric estimator; Section 4 is dedicated to auxiliary results and proofs; in Section 5 we propose a Monte Carlo study of our estimators on several simulated examples and implement our method on a real dataset which deals with the average amounts of precipitation (rainfall) for United States cities, see McNeil (1977).

2 Estimation procedure

We observe X_1, \dots, X_n independent, identically distributed random variables having common pdf g in the model

$$g(x) = p_0 f(x - \alpha_0) + (1 - p_0) f(x - \beta_0), \quad x \in \mathbb{R}, \quad (4)$$

where $\theta_0 := (p_0, \alpha_0, \beta_0)$ denotes the unknown value of the Euclidean parameter and $f \in \mathbb{L}_2$ is unknown, zero-symmetric pdf in a large nonparametric class of functions.

For identifiability reasons, let θ_0 belong to a compact set $\Theta \subset (0, 1/2) \times \mathbb{R}^2 \setminus \Delta$. Therefore, there are positive P_*, P , which are smaller than $1/2$, such that $p_0 \in [P_*, P]$.

Note that in case $p_0 = 0$ we can still identify β_0 but not α_0 . As this case reduces to the estimation of the location of an unknown zero-symmetric pdf f as in Beran (1978), we do not consider this case further on.

From now on, we denote by $f^*(u) = \int_{\mathbb{R}} e^{ixu} f(x) dx$ the Fourier transform and recall that if $f^* \in \mathbb{L}_1$ we have the inversion formula $f(x) = (2\pi)^{-1} \int_{\mathbb{R}} e^{-iux} f^*(u) du$.

Let us denote $M(\theta, u) := pe^{iu\alpha} + (1 - p)e^{iu\beta}$, for all $\theta \in \Theta$ and $u \in \mathbb{R}$, and see that it cannot be 0 as soon as $p \neq 1/2$. It is enough to notice that $(1 - 2P)^2 \leq |M(\theta, u)|^2 \leq 1$ for all $(u, \theta) \in \mathbb{R} \times \Theta$.

The contrast uses the symmetry of the underlying, unknown pdf f . For the first time in the literature of mixture models, we relate the symmetry of f to the fact that its Fourier transform has no imaginary part. More precisely, in model (4)

$$g^*(u) = (p_0 e^{iu\alpha_0} + (1 - p_0) e^{iu\beta_0}) f^*(u) = M(\theta_0, u) f^*(u), \quad u \in \mathbb{R}. \quad (5)$$

When f is supposed to be zero-symmetric, we should expect that $Im(g^*(u)/M(\theta, u)) = 0$, for all $u \in \mathbb{R}$, if and only if $\theta = \theta_0$. This basic result is formally stated in the following theorem.

Theorem 1 *Consider model (2) with f zero-symmetric such that $f^* \in \mathbb{L}_1$ and $\theta_0 \in \Theta$. Then we have $Im(g^*/M(\theta, \cdot)) = 0$ for some $\theta \in \Theta$ if and only if $\theta = \theta_0$.*

Proof. Notice that for all $\theta \in \Theta$ such that $Im(g^*/M(\theta, \cdot)) = 0$ we explicitly have

$$Im\left(\frac{g^*(u)}{M(\theta, u)}\right) = Im\left(f^*(u) \frac{M(\theta_0, u)}{M(\theta, u)}\right) = \frac{f^*(u)}{|M(\theta, u)|^2} Im((M(\theta_0, u) \bar{M}(\theta, u))) = 0,$$

for all $u \in \mathbb{R}$. As $f^*(0) = 1$, we get that $\text{Im}(M(\theta_0, \cdot)\bar{M}(\theta, \cdot))$ is null in a neighborhood of 0 which leads, following the proof of Theorem 2.1 in BMV, to the wanted result $\theta = \theta_0$.

■

From now on, we suppose that f is squared integrable. Assuming g^* is known we can recover the true value of the Euclidean parameter by minimizing the discrepancy measure S defined by

$$S(\theta) := \int_{\mathbb{R}} \left(\text{Im} \left(\frac{g^*(u)}{M(\theta, u)} \right) \right)^2 dW(u), \quad \theta \in \Theta, \quad (6)$$

where W is a Lebesgue-absolutely continuous probability measure supported by \mathbb{R} .

Note that we can also write

$$S(\theta) = \int_{\mathbb{R}} \left[\frac{1}{2i} \left(\frac{g^*(u)}{M(\theta, u)} - \frac{\bar{g}^*(u)}{\bar{M}(\theta, u)} \right) \right]^2 dW(u).$$

From now on, \bar{z} denotes the complex conjugate of z .

Proposition 1 *Consider model (2) with f zero-symmetric such that $f^* \in \mathbb{L}_1$ and $\theta_0 \in \Theta$. Then, the function S in (6) is a contrast function, i.e. for all $\theta \in \Theta$, $S(\theta) \geq 0$ and $S(\theta) = 0$ if and only if $\theta = \theta_0$.*

Proof. The Fourier transform f^* being continuous, the same holds for $\text{Im} \left(\frac{g^*}{M(\theta, \cdot)} \right)$. By Theorem 1, if $\theta \neq \theta_0$ there exists $u_0 \in \mathbb{R}$ such that $\text{Im} \left(\frac{g^*(u_0)}{M(\theta, u_0)} \right) \neq 0$, and there exists $\varepsilon > 0$ and $\gamma > 0$ such that $\text{Im} \left(\frac{g^*(u)}{M(\theta, u)} \right) > \varepsilon$ on $[u_0 - \gamma, u_0 + \gamma]$. It follows that

$$S(\theta) \geq \varepsilon^2 \int_{u_0 - \gamma}^{u_0 + \gamma} dW(u) > 0.$$

Otherwise if $\theta = \theta_0$ it is straightforward to check that $S(\theta) = 0$. ■

Discussion. We point out that basic results similar to Theorem 1 and Proposition 1, can be established for model (2) when $k = 3$ under sufficient identifiability conditions. Indeed, in that case, it is enough to replace θ by $(\lambda_1, \lambda_2, \mu_1, \mu_2, \mu_3)^T$ and $M(\theta, u)$ by $\sum_{j=1}^3 \lambda_j e^{iu\mu_j}$ and check that the analog of Theorem 1 can be established following the Proof of Lemma A. 1, under conditions provided in Corollary 1, in HWH. Finally, similar estimators to those in Sections 2.1 and 2.2 and asymptotic results like those established in Section 3 for $k = 2$, can be established with a little extra work for $k = 3$.

2.1 Contrast minimization for the Euclidean parameter

Let the estimator of θ_0 be the following M -estimator

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} S_n(\theta), \quad (7)$$

where $S_n(\theta)$ is the following estimator of $S(\theta)$

$$S_n(\theta) = \frac{-1}{4n(n-1)} \int \sum_{j \neq k, j, k=1}^n \left(\frac{e^{iuX_k}}{M(\theta, u)} - \frac{e^{-iuX_k}}{M(\theta, -u)} \right) \left(\frac{e^{iuX_j}}{M(\theta, u)} - \frac{e^{-iuX_j}}{M(\theta, -u)} \right) dW(u), \quad (8)$$

where W is a Lebesgue absolutely continuous cdf. The estimator $S_n(\theta)$ is inspired by kernel estimators of quadratic functional of the pdf f as previously studied in Butucea (2007). It is written here in the Fourier domain. It is known that by removing the diagonal terms in the double sum (*i.e.* taking $j \neq k$) the bias is reduced. The weight function W solves all integrability issues of the criterion and its derivatives, as soon as it has finite moments up to order 3. Moreover, as this function is a distribution function, we will be able to compute the previous integral via Monte-Carlo simulation.

Let us denote by

$$\begin{aligned} Z_k(\theta, u) &:= \frac{e^{iuX_k}}{M(\theta, u)} - \frac{e^{-iuX_k}}{M(\theta, -u)}, \\ J(\theta, u) &:= \frac{g^*(u)}{M(\theta, u)} - \frac{g^*(-u)}{M(\theta, -u)}. \end{aligned}$$

Then it is easy to see that

$$\begin{aligned} S_n(\theta) &= \frac{-1}{4n(n-1)} \sum_{j \neq k, j, k=1}^n \int Z_k(\theta, u) Z_j(\theta, u) dW(u), \\ S(\theta) &= -\frac{1}{4} \int_{\mathbb{R}} J^2(\theta, u) dW(u), \end{aligned}$$

and that $E[Z_k(\theta, u)] = J(\theta, u)$.

Note that, for numerical implementation, we can also consider equivalently the Monte Carlo estimate of S , based on the simulation of an i.i.d. sample Y_1, \dots, Y_n (independent of X_1, \dots, X_n) derived from the distribution W , as follows

$$\tilde{S}_n(\theta) = \frac{-1}{4n^2(n-1)} \sum_{\ell=1}^n \sum_{j \neq k, j, k=1}^n \left(\frac{e^{iY_\ell X_k}}{M(\theta, Y_\ell)} - \frac{e^{-iY_\ell X_k}}{M(\theta, -Y_\ell)} \right) \left(\frac{e^{iY_\ell X_j}}{M(\theta, Y_\ell)} - \frac{e^{-iY_\ell X_j}}{M(\theta, -Y_\ell)} \right). \quad (9)$$

The M -estimator associated to $\tilde{S}_n(\theta)$ is then defined by

$$\tilde{\theta}_n = \arg \min_{\theta \in \Theta} \tilde{S}_n(\theta), \quad (10)$$

and has equivalent asymptotic properties (see Lemma 4 in Appendix) to those stated for $\hat{\theta}_n$ in Theorems 2 and 3 .

2.2 Kernel based nonparametric estimator

After estimating the Euclidean parameter, we want to estimate the nonparametric pdf f which is assumed squared integrable. For technical reasons, we suggest to use cross-validation for a kernel estimator as follows. We denote by $\hat{\theta}_{n,-k}$ the leave-one-out estimator of θ_0 , which uses the sample without the k -th observation in the procedure defined by (7) and (8). Then, we plug this in the classical nonparametric kernel estimator, whenever the unknown θ_0 is required. This procedure gives, in Fourier domain,

$$f_n^*(u) = \frac{1}{n} \sum_{k=1}^n \frac{K^*(b_n u) e^{iuX_k}}{M(\hat{\theta}_{n,-k}, u)}, \quad (11)$$

where K the kernel ($\int K = 1$ and $K \in \mathbb{L}_2$) and b_n the bandwidth are properly chosen. Note that $G_n^*(u) := K^*(b_n u)/M(\hat{\theta}_{n,-k}, u)$ is in \mathbb{L}_1 and \mathbb{L}_2 and has an inverse Fourier transform which we denote by $G_n(u/b_n)/b_n$. Therefore, the estimator of f is

$$f_n(x) = \frac{1}{nb_n} \sum_{k=1}^n G_n \left(\frac{x - X_k}{b_n} \right). \quad (12)$$

In practice, we work with the estimator $\hat{\theta}_n$. Indeed, we use the leave-one-out estimator $\hat{\theta}_{n,-k}$ of θ_0 for technical reasons, as it makes the random variables under the sum in (11) uncorrelated. Nevertheless, these correlations should be negligible. Another way to proceed is to split the sample in two: one part estimates θ_0 and the other one reconstructs the underlying density function f . It is important to notice at this step, that the estimator f_n is obtained by inversion of a nonparametric kernel estimator

$$g_n(x) = \frac{1}{nb_n} \sum_{k=1}^n K \left(\frac{x - X_k}{b_n} \right),$$

with kernel K and bandwidth b_n . The inversion is done in Fourier domain with the estimated $\hat{\theta}_n$ instead of the true θ_0 : $f_n^*(u) = g_n^*(u)/M(\hat{\theta}_n, u)$.

When dealing with the rain fall dataset studied in Section 4, we propose to consider, as in BMV, the version \tilde{f}_n of the estimator $f_n(x)$ (which has a negative part due to the small number of observations) defined by

$$\tilde{f}_n(x) = \frac{f_n(x)\mathbb{I}_{f_n(x)\geq 0}}{\int_{\mathbb{R}} f_n(x)\mathbb{I}_{f_n(x)\geq 0}}. \quad (13)$$

3 Main results

Let us state first several assumptions.

Assumption A Let $W : \mathbb{R} \rightarrow \mathbb{R}^+$ be a cumulative distribution function of some random variable which admits finite absolute moments up to the third order:

$$\int_{\mathbb{R}} (1 + |u| + u^2 + |u|^3) dW(u) < \infty.$$

Assumption B We assume that the underlying probability density f belongs to a ball of radius $L > 0$ in the Sobolev space of functions having smoothness $\beta > 0$:

$$W(\beta, L) := \left\{ f : \mathbb{R} \rightarrow \mathbb{R}_+ : f \in \mathbb{L}_2, \int f = 1, \int |f^*(u)|^2 |u|^{2\beta} du \leq L \right\},$$

where f^* denotes the Fourier transform of the function f .

The weight function W has been introduced for integrability of our estimator $S_n(\theta)$ of the criterion $S(\theta)$ and its derivatives with respect to θ . It is completely arbitrary and it may help compute numerically the values of our integrals by Monte-Carlo simulation, but it slightly reduces the asymptotic efficiency of $\hat{\theta}_n$. We could have used integrals with respect to the Lebesgue measure for highest efficiency of $\hat{\theta}_n$, but this would require stronger assumptions of smoothness and moments for the unknown probability density function f .

Proposition 2 Consider model (2) with f zero-symmetric such that $f^* \in \mathbb{L}_2$ and $\theta_0 \in \Theta$. For each $\theta \in \Theta$, the empirical contrast function $S_n(\cdot)$ defined in (8), with W verifying assumption **A**, is such that

$$\sup_{f \in W(\beta, L)} \sup_{\theta \in \Theta} E \left[(S_n(\theta) - S(\theta))^2 \right] \leq \frac{1 + o(1)}{(1 - 2P)^2 n},$$

as $n \rightarrow \infty$.

An easy consequence of the Theorem is that for each $\theta \in \Theta$, $|S_n(\theta) - S(\theta)| = O_P(n^{-1/2})$ as $n \rightarrow \infty$.

We will denote respectively in the sequel by \xrightarrow{P} and \xrightarrow{d} the convergence in probability, resp. in distribution.

Theorem 2 Consider model (2) with f zero-symmetric such that $f^* \in \mathbb{L}_2$ and $\theta_0 \in \Theta$. Let W verify assumption **A**. Then, the estimator $\hat{\theta}_n$ defined in (7) converges in probability to the true value of the Euclidean parameter θ_0 as $n \rightarrow \infty$.

Theorem 3 Under the assumptions of Theorem 2, the estimator $\hat{\theta}_n$ defined in (7) is asymptotically normally distributed:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma), \text{ as } n \rightarrow \infty,$$

where $\Sigma = \mathcal{I}^{-1}V\mathcal{I}$, $\mathcal{I} = \mathcal{I}(\theta_0) = -\frac{1}{2} \int_{\mathbb{R}} \dot{J}(\theta_0, u) \dot{J}^\top(\theta_0, u) dW(u)$, $V = \frac{1}{4} E(U_1(\theta_0) U_1^\top(\theta_0))$ and $U_1(\theta_0) = \int_{\mathbb{R}} Z_1(\theta_0, u) \dot{J}(\theta_0, u) dW(u)$.

The next theorem gives the upper bounds for the rate of convergence of the nonparametric estimator f_n of f , at some fixed point x , over Sobolev classes of functions. The main message of the theorem is that, if $\beta > 1/2$ then the nonparametric rates for density estimation are reached, provided a correct choice of the parameter b_n . This might seem surprising, but it is again related to the fact that the inverse problem under consideration is well posed and the estimation of the Euclidean parameter θ_0 does not affect the nonparametric rate for estimating f .

Theorem 4 Consider model (2) with f zero-symmetric such that f verifies assumption **B** for some $\beta > 1/2$ and $\theta_0 \in \Theta$. Let W verify assumption **A** and the estimator $\hat{\theta}_n$ of θ be defined in (7). Let $f_n(x)$ be the estimator of $f(x)$ at some fixed point $x \in \mathbb{R}$ in (12), with $b_n = cn^{-1/(2\beta)}$ for some $c > 0$ and a kernel K in \mathbb{L}_1 and in \mathbb{L}_2 with Fourier transform K^* having support included in the set $\{u : |u| \geq 1\}$. Then

$$\limsup_{n \rightarrow \infty} \sup_{f \in W(\beta, L)} \sup_{\theta_0 \in \Theta} E_{\theta_0, f} \left[n^{-\frac{2\beta-1}{2\beta}} |f_n(x) - f(x)|^2 \right] \leq C, \quad (14)$$

for some constant $C < \infty$ which depends on β , L , P and on $\int K^2$. Moreover,

$$\liminf_{n \rightarrow \infty} \inf_{\tilde{f}_n, \tilde{\theta}_n} \sup_{f \in W(\beta, L)} \sup_{\theta_0 \in \Theta} E_{\theta_0, f} \left[n^{-\frac{2\beta-1}{2\beta}} |\tilde{f}_n(x) - f(x)|^2 \right] \geq C_*, \quad (15)$$

for some $C_* > 0$ depending only on β , L , P , where the infimum is taken over all estimators \tilde{f}_n and $\tilde{\theta}_n$ of f and θ_0 , respectively.

In the proof of this Theorem we discuss the upper bounds (14). The lower bounds (15) are briefly explained as follows. We can choose an arbitrary point $\theta \in \Theta$ and write

$$\sup_{f \in W(\beta, L)} \sup_{\theta_0 \in \Theta} E_{\theta_0, f} \left[n^{-\frac{2\beta-1}{2\beta}} |\tilde{f}_n(x) - f(x)|^2 \right] \geq \sup_{f \in W(\beta, L)} E_{\theta, f} \left[n^{-\frac{2\beta-1}{2\beta}} |\tilde{f}_n(x) - f(x)|^2 \right].$$

For fixed given θ , the infimum in (15) is bounded from below by

$$\inf_{\tilde{f}_n} \sup_{f \in W(\beta, L)} E_{\theta, f} \left[n^{-\frac{2\beta-1}{2\beta}} |\tilde{f}_n(x) - f(x)|^2 \right],$$

and this problem reduces to a nonparametric problem of pointwise lower bounds. The lower bounds associated to these rates are known in the case of density estimation from direct observations, see for example results for more general Besov classes of functions in Härdle *et al.* (1998).

Remark. Generalization to the multivariate version of model (4) is straightforward since considering a symmetric multivariate density function f on \mathbb{R}^d with $d > 1$, we have $f(\mathbf{x}) = f(-\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$, which leads to

$$g^*(\mathbf{u}) := \int_{\mathbb{R}^d} e^{i\langle \mathbf{u}, \mathbf{x} \rangle} g(\mathbf{x}) d\mathbf{x} = (p_0 e^{i\langle \mathbf{u}, \alpha_0 \rangle} + (1 - p_0) e^{i\langle \mathbf{u}, \beta_0 \rangle}) f^*(\mathbf{u}), \quad (16)$$

for all $\mathbf{u} = (u_1, \dots, u_d)^T \in \mathbb{R}^d$. Now, considering successively the situations where $u_i \in \mathbb{R}$ and $u_j = 0$ for $i = 1, \dots, d$ and $j \neq i$, we obtain the basic Fourier equation (5) which allows, when applying our method, to estimate successively (α_i, β_i) for $i = 1, \dots, d$ at the \sqrt{n} -rate established in Theorem 3. Nevertheless we obtain a new rate of convergence for f_n equal to $n^{-\frac{2\beta-1}{2\beta-1+d}}$, when taking $b_n = cn^{-\frac{1}{2\beta-1+d}}$ in Theorem 4.

Remark. Note that similar results can be stated for the mean integrated squared error, MISE. The minimax rates for estimating the multivariate probability density over Sobolev classes are $n^{-2\beta/(2\beta+d)}$, when taking $b_n = cn^{-\frac{1}{2\beta+d}}$ in Theorem 4.

4 Simulations

We implement our method and study its behaviour on samples of size $n = 100, 200$. The mean behaviour of our estimator $\tilde{\theta}_n$ defined in (10) is calculated by replicating $M = 100$ times the same experiment. We will consider three different types of mixed densities in model (4) in order to answer the following natural questions:

- i) How behaves our method compared to the method proposed in BMV ? For this purpose we will consider a standard Gaussian density under the BMV's set of parameter $p = 0.15, 0.25, 0.35$, with $\alpha = -1$, $\beta = 2$, and choose $V \sim \mathcal{N}(0, 3^2)$.
- ii) Does heavily tailed distribution functions influence badly the performances of our method ? For this purpose we will consider a Cauchy density under the set of parameter $p = 0.2$, $\alpha = 1$ and $\beta = 1.2, 1.5, 2, 5$, and choose $V \sim \mathcal{N}(0, 6^2)$.
- iii) On the contrary, does strongly peaked distribution functions help in detecting the location parameters ? For this purpose we will consider a Laplace density under the set of parameters $p = 0.15, 0.25, 0.35, 0.45$, and $\alpha = -1$ and $\beta = 2$, and choose $V \sim \mathcal{N}(0, 3^2)$.

Our simulation results obtained in the situations i–iii) are respectively summarized in Tables 1–3, where we give the mean value of the estimated parameter and its standard deviation. We also plot, for situations coming from i–iii), the nonparametric estimator of the underlying density as compared to the true, in Figure 2.

We are considering the bandwidth $b_n = n^{-1/4}$. The choice of the bandwidth b_n in practice is an issue in nonparametric estimation of functions. For a review of methods, some of which are nowadays implemented in commonly used software, we address the reader to the book by Scott [20]. We note that the previous estimation methods proposed by BMV and HWH requires usually, to derive asymptotic rates of convergence, finite moments up to some order. These methods cannot for example deal (excepted eventually on the consistency problem) with the Cauchy density that we consider here, see Table 2. Indeed, our method is based on Fourier transform, which is fast decreasing in this case. We also consider non smooth Laplace density (or double exponential), see Table 3.

Comments on Table 1. Let denote $\bar{\theta}_n = (\bar{p}_n, \bar{\alpha}_n, \bar{\beta}_n)$ the estimator considered in BMV. The results given in Table 1 clearly show that our method behaves quite differently from the method proposed in BMV. On the one hand, our estimation of the proportion p is more (negatively) biased when the bias for the location parameters α and β is quite similar to the bias observed in BMV. Note also that the bias for p diminishes strongly when n jumps from $n = 100$ to $n = 200$. Concerning the standard deviation of our estimator we also observe that the estimation of p is more unstable in our case for $p_0 = 0.15$ and 0.25 when for $p = 0.35$ this difference tends to vanish, as it is summarized in Fig. 1 (a). On the other hand, our results concerning the estimation of the location parameters are

much more stable than those obtained in BMV, as it is summarized in Fig. 1 (b) and (c), specially the estimation of α which belongs to $[0.079, 0.176]$ in our case when it belongs to $[0.176, 0.365]$ in BMV for $p_0 = 0.15, 0.25$ and $n = 100, 200$. In conclusion, it seems that the BMV method estimates better small values of p when our method performs better when considering the estimation of the location parameters (α, β) , the performance of both methods becoming almost equivalent for $p = 0.35$.

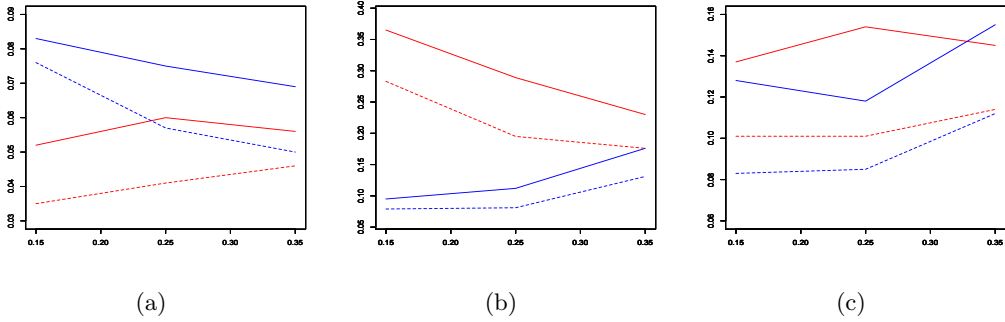


Figure 1: In blue, resp. in red, the standard deviation of $\tilde{\theta}_n = (\tilde{p}_n, \tilde{\alpha}_n, \tilde{\beta}_n)$, resp. $\bar{\theta}_n = (\bar{p}_n, \bar{\alpha}_n, \bar{\beta}_n)$, compared componentwise in (a), (b) and (c), for $p = 0.15, 0.25, 0.35$ where full line graphs, resp. dashed line graphs, correspond to $n = 100$, resp. $n = 200$.

Comments on Table 2. Let us observe first that our estimator has a very small bias over all the situations considered in this Table. On the other hand the componentwise standard deviation of $\tilde{\theta}_n = (\tilde{p}_n, \tilde{\alpha}_n, \tilde{\beta}_n)$ increases significantly when β_0 becomes close to α_0 , that is when the model becomes close to unidentifiable. Finally it is important to observe, comparing the rows 1–4 of Table 1 to the rows 1–2 in Table 2, corresponding respectively to $p_0 = 0.15, 0.25$ and $p_0 = 0.2$, both cases with $\beta_0 - \alpha_0 = 3$, that the results in the Cauchy case are slightly worse than in the Gaussian case (this last remark answering partially the influence of the distribution tails on our estimation method).

Comments on Table 3. The main fact to observe here is that the Laplace distribution introduces a bias on the estimation of p bigger than in the Gaussian case and diminishing even more slowly when n jumps from 100 to 200. Secondly the standard deviation of \tilde{p}_n is clearly smaller than in the Gaussian or Cauchy case, when the estimation of the location parameters is badly affected when p_0 is about 0.30. This phenomenon is probably due to the fact that the Laplace distribution has heavier tails than the Gaussian distribution

which implies that the two mixed populations are more overlapped (unclear design with one component under-represented). Nevertheless, as it is intuitively expected, we observe that for $p_0 = 0.45$ (two populations almost equally-represented) the two components of the mixture model are very well estimated as well as the mixture ratio.

n	(p_0, α_0, β_0)	Empirical means	Standard deviations
100	(0.15, -1, 2)	(0.126, -0.984, 1.993)	(0.083, 0.095, 0.128)
200	(0.15, -1, 2)	(0.141, -0.997, 2.001)	(0.076, 0.079, 0.083)
100	(0.25, -1, 2)	(0.219, -1.010, 1.994)	(0.075, 0.112, 0.118)
200	(0.25, -1, 2)	(0.243, -0.993, 2.081)	(0.057, 0.081, 0.085)
100	(0.35, -1, 2)	(0.312, -0.998, 1.948)	(0.069, 0.176, 0.155)
200	(0.35, -1, 2)	(0.344, -1.010, 1.997)	(0.050, 0.131, 0.112)

Table 1: Empirical means and standard deviations (from $M = 100$ samples of size n) of the estimator $\hat{\theta}_n = (\hat{p}_n, \hat{\alpha}_n, \hat{\beta}_n)$ of $\theta_0 = (p_0, \alpha_0, \beta_0)$ when f is standard Gaussian.

n	(p_0, α_0, β_0)	Empirical means	Standard deviations
100	(0.2, 1, 5)	(0.198, 0.994, 5.025)	(0.087, 0.113, 0.179)
200	(0.2, 1, 5)	(0.201, 0.996, 5.020)	(0.083, 0.094, 0.117)
100	(0.2, 1, 2)	(0.200, 0.980, 1.958)	(0.108, 0.254, 0.236)
200	(0.2, 1, 2)	(0.200, 1.991, 2.011)	(0.098, 0.153, 0.173)
100	(0.2, 1, 1.5)	(0.209, 1.019, 1.457)	(0.1426, 0.274, 0.242)
200	(0.2, 1, 1.5)	(0.198, 1.012, 1.511)	(0.096, 0.169, 0.171)

Table 2: Empirical means and standard deviations (from $M = 100$ samples of size n) of the estimator $\hat{\theta}_n = (\hat{p}_n, \hat{\alpha}_n, \hat{\beta}_n)$ of $\theta_0 = (p_0, \alpha_0, \beta_0)$ when f is standard Cauchy.

Rainfall dataset. In this paragraph we propose to study the performances of our method when compared to the results obtained in BMV. We have implemented the Gauss kernel estimator with bandwidth $b_n = 2n^{-1/4}$, $n = 70$, and used in (11), instead of $\hat{\theta}_{n,-k}$, the estimator $\hat{\theta}_n$. When K is the Gauss kernel, we explicitly have

$$f_n(x) = \frac{1}{n} \sum_{k=1}^n \int_{\mathbb{R}} Q(b_n, \hat{\theta}_n; u) [\hat{p}_n \cos(u(X_k - x - \hat{\alpha}_n)) + (1 - \hat{p}_n) \cos(u(X_k - x - \hat{\beta}_n))] du,$$

n	(p_0, α_0, β_0)	Empirical means	Standard deviations
100	(0.15, -1, 2)	(0.138, -1.001, 1.994)	(0.053, 0.093, 0.155)
200	(0.15, -1, 2)	(0.144, -1.002, 2.016)	(0.045, 0.092, 0.113)
100	(0.25, -1, 2)	(0.216, -1.021, 1.929)	(0.056, 0.172, 0.145)
200	(0.25, -1, 2)	(0.226, -1.019, 2.00)	(0.053, 0.119, 0.095)
100	(0.35, -1, 2)	(0.333, -0.964, 2.029)	(0.031, 0.225, 0.153)
200	(0.35, -1, 2)	(0.343, -0.992, 2.005)	(0.031, 0.132, 0.135)
100	(0.45, -1, 2)	(0.439, -0.994, 2.000)	(0.016, 0.093, 0.086)
200	(0.45, -1, 2)	(0.440, -1.000, 2.001)	(0.014, 0.049, 0.056)

Table 3: Empirical means and standard deviations (from $M = 100$ samples of size n) of the estimator $\hat{\theta}_n = (\hat{p}_n, \hat{\alpha}_n, \hat{\beta}_n)$ of $\theta_0 = (p_0, \alpha_0, \beta_0)$ when f is Laplace.

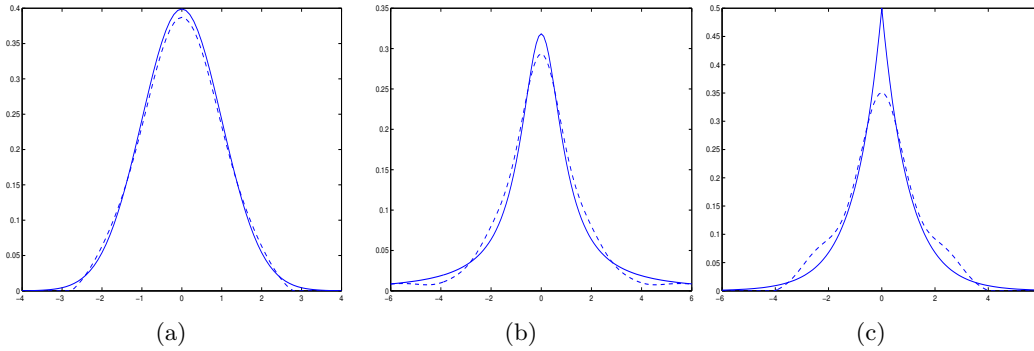


Figure 2: Underlying density (solid line) and kernel estimator (dashed line) for a) Gauss density, b) Cauchy density and c) Laplace density.

where

$$Q(\theta, b; u) := \frac{1}{2\pi} \times \frac{e^{-b^2 u^2 / 2}}{2p^2 - 2p + 1 + 2p(1-p) \cos(u(\alpha - \beta))}.$$

The results provided by our method are $\hat{p}_n = 0.15$, $\hat{\alpha}_n = 12.7$, $\hat{\beta}_n = 38.5$ and the behavior of the non parametric estimators of pdf's is summarized in Figure 2. Before commenting the good performances of our estimator $(\hat{\theta}_n, \tilde{f}_n)$ in Figure 2, it is crucial to notice that the reconstruction of the pdf g by $g_{\hat{\theta}_n, f_n}(\cdot) = \hat{p}_n f_n(\cdot - \hat{\alpha}_n) + (1 - \hat{p}_n) f_n(\cdot - \hat{\beta}_n)$ coincides with g_n itself, according to (11-13) and replacing $\hat{\theta}_{n,-k}$ by $\hat{\theta}_n$. This basic phenomenon is illustrated in Figure 3. As mentioned in Section 2.2, the function f_n is not necessarily a

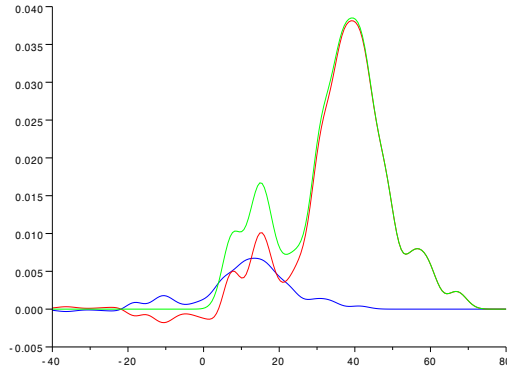


Figure 3: Rainfall dataset. In blue the graph of $\hat{p}_n f_n(\cdot - \hat{\alpha}_n)$, in red the graph of $(1 - \hat{p}_n) f_n(\cdot - \hat{\beta}_n)$, in green the graph of $g_{\hat{\theta}_n, f_n}(\cdot) = \hat{p}_n f_n(\cdot - \hat{\alpha}_n) + (1 - \hat{p}_n) f_n(\cdot - \hat{\beta}_n) = g_n$ obtained with $h_n = 2.5$.

pdf due to its negative part (coming from the small size of n and the fact that model (4) is not necessarily the true underlying model), hence it is needed to regularize f_n into \tilde{f}_n which leads to consider, on this real dataset, $\tilde{f}_n = 0.9644 \times f_n \mathbb{I}_{f_n \geq 0}$. This modification explains the fact the graph of $g_{\hat{\theta}_n, \tilde{f}_n} = \hat{p}_n \tilde{f}_n(\cdot - \hat{\alpha}_n) + (1 - \hat{p}_n) \tilde{f}_n(\cdot - \hat{\beta}_n)$ does not match exactly the graph of $g_{\hat{\theta}_n, f_n} = g_n$. Actually we observe that the graph of $g_{\hat{\theta}_n, \tilde{f}_n}(\cdot)$ fits almost perfectly the graph of \hat{g}_n in the interval $[0, 80]$, when it generates an extra bump in the interval $[-20, 0]$. Nethertheless when comparing our graphs to the graphs obtained in BMV (including a comparison with the two-component Gaussian mixture model), we observe that we both have the extra bump issue on the interval $[-20, 0]$, on the other hand we better estimate the two first bumps appearing on the graph of g_n within the interval

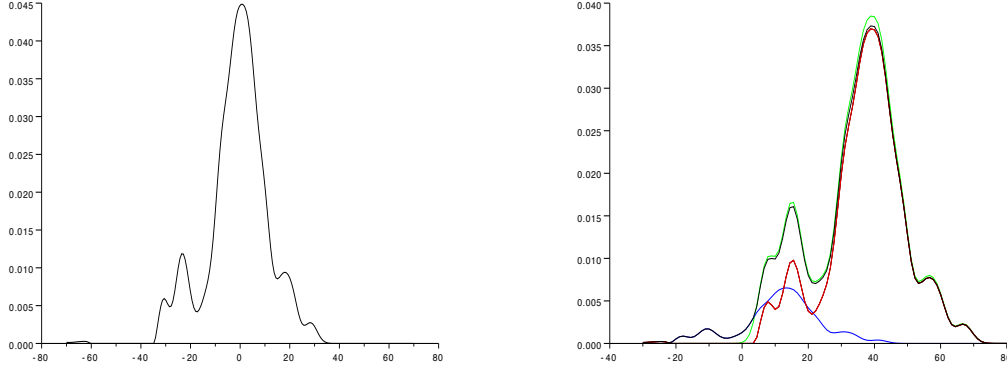


Figure 4: Rainfall dataset. a) Graph of $\tilde{f}_n f$; b) In blue the graph of $\hat{p}_n \tilde{f}_n(\cdot - \hat{\alpha}_n)$, in red the graph of $(1 - \hat{p}_n) \tilde{f}_n(\cdot - \hat{\beta}_n)$, in black the graph of $g_{\hat{\theta}_n, \tilde{f}_n}(\cdot) = \hat{p}_n \tilde{f}_n(\cdot - \hat{\alpha}_n) + (1 - \hat{p}_n) \tilde{f}_n(\cdot - \hat{\beta}_n)$, in green the graph of g_n obtained with $h_n = 2.5$.

[0, 20]. We think that our methodological approach performs better than the existing one, mainly because we do not symmetrize our estimator \tilde{f}_n in order to mimic as much as possible the shape of f_n (which shapeless is precisely the reason why $g_{\hat{\theta}_n, \tilde{f}_n} = g_n$, see Figure 4).

5 Auxiliary results and Proofs

Let us use the notation $\|v\|$ for the Euclidean norm of a vector $v \in \mathbb{R}^d$ and $\|A\|_2^2 = \text{tr}(A^\top A)$ for any matrix A in $\mathbb{R}^{d \times d}$. In this Section we assume that f is squared integrable and that W verifies assumption **A**.

Lemma 1 1. For all $u \in \mathbb{R}$, we have

$$\max\{\sup_{\theta \in \Theta} |Z_k(\theta, u)|, \sup_{\theta \in \Theta} |J(\theta, u)|\} \leq \frac{2}{1 - 2P},$$

for any $k = 1, \dots, n$.

2. For all $u \in \mathbb{R}$, we have

$$\max\{\sup_{\theta \in \Theta} \|\dot{Z}_k(\theta, u)\|, \sup_{\theta \in \Theta} \|\dot{J}(\theta, u)\|\} \leq \frac{4(1 + |u|)}{(1 - 2P)^2},$$

for any $k = 1, \dots, n$.

3. For all $u \in \mathbb{R}$, we have

$$\|\ddot{Z}_k(\theta, u)\|_2 \leq \frac{C(1 + |u| + u^2)}{(1 - 2P)^3},$$

for some absolute constant $C > 0$, for any $\theta \in \Theta$ and for any $k = 1, \dots, n$.

Proof. 1. It is easy to see that $|Z_j(\theta, u)| \leq 2/|M(\theta, u)| \leq 2/(1 - 2P)$ and that

$$|J(\theta, u)| \leq 2 \left| \frac{g^*(u)}{M(\theta, u)} \right| \leq \frac{2}{(1 - 2P)}.$$

2. We note that

$$\dot{Z}_k(\theta, u) = -\frac{e^{iuX_k}}{M^2(\theta, u)} \begin{pmatrix} e^{iu\alpha} - e^{iu\beta} \\ iupe^{iu\alpha} \\ iu(1-p)e^{iu\beta} \end{pmatrix} + \frac{e^{-iuX_k}}{M^2(\theta, -u)} \begin{pmatrix} e^{-iu\alpha} - e^{-iu\beta} \\ -iupe^{-iu\alpha} \\ -iu(1-p)e^{-iu\beta} \end{pmatrix},$$

and that

$$E[\dot{Z}_k(\theta, u)] = \dot{J}(\theta, u) = -\frac{g^*(u)}{M^2(\theta, u)} \begin{pmatrix} e^{iu\alpha} - e^{iu\beta} \\ iupe^{iu\alpha} \\ iu(1-p)e^{iu\beta} \end{pmatrix} + \frac{g^*(-u)}{M^2(\theta, -u)} \begin{pmatrix} e^{-iu\alpha} - e^{-iu\beta} \\ -iupe^{-iu\alpha} \\ -iu(1-p)e^{-iu\beta} \end{pmatrix}.$$

We have

$$\begin{aligned} \|\dot{J}(\theta, u)\| &= \left\| \frac{g^*(u)}{M^2(\theta, u)} \dot{M}(\theta, u) + \frac{g^*(-u)}{M^2(\theta, -u)} \dot{M}(\theta, -u) \right\| \\ &\leq \frac{1}{(1 - 2P)^2} (2(2^2 + p^2u^2 + (1-p)^2u^2))^{1/2} \leq \frac{4(1 + |u|)}{(1 - 2P)^2}, \end{aligned}$$

and the same goes for $\dot{Z}_k(\theta, u)$, which gives us the expected result.

3. We write briefly

$$\begin{aligned} \ddot{Z}_k(\theta, u) &= -\frac{e^{iuX_k}}{M^2(\theta, u)} \ddot{M}(\theta, u) + \frac{e^{-iuX_k}}{M^2(\theta, -u)} \ddot{M}(\theta, -u) \\ &\quad + 2\frac{e^{iuX_k}}{M^3(\theta, u)} \dot{M}(\theta, u) \dot{M}(\theta, u)^\top - 2\frac{e^{-iuX_k}}{M^3(\theta, -u)} \dot{M}(\theta, -u) \dot{M}(\theta, -u)^\top, \end{aligned}$$

and deduce our bound from the above expression. ■

Lemma 2 1. For all $u \in \mathbb{R}$, we have

$$\|\dot{Z}_k(\theta, u) - \dot{Z}_k(\theta', u)\| \leq \|\theta - \theta'\| \frac{C(1 + |u| + u^2)}{(1 - 2P)^3},$$

for any $\theta, \theta' \in \Theta$ and any $k = 1, \dots, n$.

2. For all $u \in \mathbb{R}$, we have

$$\|\ddot{Z}_k(\theta, u) - \ddot{Z}_k(\theta', u)\|_2 \leq \|\theta - \theta'\| \frac{C(1 + |u| + u^2 + |u|^3)}{(1 - 2P)^4},$$

for some absolute constant $C > 0$, for any $\theta, \theta' \in \Theta$ and for any $k = 1, \dots, n$.

Proof. The proof uses a Taylor expansion and bounds from above are obtained similarly to Lemma 1. ■

Proof of Proposition 2. It is easy to see that $E[Z_k(\theta, u)] = J(\theta, u)$. Therefore the estimator $S_n(\theta)$ of $S(\theta)$ is unbiased. We have for the variance

$$\begin{aligned} \text{Var}(S_n(\theta)) &= \frac{1}{16} E \left[\left(\frac{1}{n(n-1)} \sum_{j \neq k, j, k=1}^n \int (Z_j(\theta, u) Z_k(\theta, u) - J^2(\theta, u)) dW(u) \right)^2 \right]. \end{aligned}$$

It decomposes in $\text{Var}(S_n(\theta)) = \frac{1}{16}(T_n + V_n)$, where

$$\begin{aligned} T_n &= E \left[\left(\frac{1}{n(n-1)} \sum_{j \neq k, j, k=1}^n \int (Z_j(\theta, u) - J(\theta, u))(Z_k(\theta, u) - J(\theta, u)) dW(u) \right)^2 \right] \\ V_n &= E \left[\left(\frac{2}{n} \sum_{k=1}^n \int (Z_k(\theta, u) - J(\theta, u)) J(\theta, u) dW(u) \right)^2 \right] \end{aligned}$$

Indeed, random variables in the previous sums are uncorrelated. Let us study the asymptotic behavior of these terms. On the one hand,

$$\begin{aligned} T_n &= \frac{1}{n(n-1)} E \left[\left(\int (Z_1(\theta, u) - J(\theta, u))(Z_2(\theta, u) - J(\theta, u)) dW(u) \right)^2 \right] \\ &\leq \frac{1}{n(n-1)} E \left[\left(\int Z_1(\theta, u) Z_2(\theta, u) dW(u) \right)^2 \right] \leq \frac{16}{(1 - 2P)^4 n^2}, \end{aligned}$$

since from Lemma 1 we have $|Z_k(\theta, u)| \leq 2(1 - 2P)^{-1}$. On the other hand,

$$V_n = \frac{4}{n} E \left[\left(\int Z_1(\theta, u) J(\theta, u) dW(u) \right)^2 \right] - \frac{4}{n} \left(\int J^2(\theta, u) dW(u) \right)^2.$$

We have that $\int J^2(\theta, u) dW(u) = -4S(\theta)$. As for the first term, we use that $|J(\theta, u)| \leq 2(1 - 2P)^{-1}$, for all $(u, \theta) \in \mathbb{R} \times \Theta$, and write

$$E \left[\left(\int Z_1(\theta, u) J(\theta, u) dW(u) \right)^2 \right] \leq \frac{4}{(1 - 2P)^2},$$

which concludes the proof. ■

Lemma 3 i) The function S is Lipschitz over Θ .

ii) The empirical contrast S_n defined in (8) is Lipschitz over Θ .

iii) The empirical contrast S_n defined in (8) is such that \ddot{S}_n is Lipschitz over Θ .

Proof. i) According to the mean value theorem, we write

$$\begin{aligned} S(\theta) - S(\theta') &= -\frac{1}{4} \int_{\mathbb{R}} [J^2(\theta, u) - J^2(\theta', u)] dW(u) \\ &= -\frac{1}{4} \int_{\mathbb{R}} (\theta - \theta')^\top \dot{J}^2(\theta_u, u) dW(u) \\ &= -\frac{1}{2} \int_{\mathbb{R}} (\theta - \theta')^\top \dot{J}(\theta_u, u) J(\theta_u, u) dW(u), \end{aligned}$$

where for all $u \in \mathbb{R}$, θ_u lies in the line segment with extremities θ and θ' . By Cauchy-Schwarz inequality,

$$|S(\theta) - S(\theta')| \leq \frac{1}{2} \|\theta - \theta'\| \int_{\mathbb{R}} \|\dot{J}(\theta_u, u)\| \cdot |J(\theta_u, u)| dW(u).$$

By Lemma 1, $|S(\theta) - S(\theta')| \leq 4(1 - 2P)^{-3} \int (1 + |u|) dW(u) \|\theta - \theta'\|$, which establishes our Lipschitz property.

ii) Very similarly,

$$\begin{aligned} S_n(\theta) - S_n(\theta') &= -\frac{1}{4n(n-1)} \sum_{j \neq k, j, k=1}^n \int (\theta - \theta')^\top \nabla (Z_k(\theta, u) Z_j(\theta, u)) |_{\theta=\theta_u} dW(u) \\ &= -\frac{1}{2n(n-1)} \sum_{j \neq k, j, k=1}^n \int (\theta - \theta')^\top \dot{Z}_k(\theta_u, u) Z_j(\theta_u, u) dW(u), \end{aligned}$$

where for all $u \in \mathbb{R}$, θ_u lies in the line segment with extremities θ and θ' . Therefore

$$|S_n(\theta) - S_n(\theta')| \leq \frac{4}{(1 - 2P)^3} \|\theta - \theta'\| \int_{\mathbb{R}} (1 + |u|) dW(u).$$

Indeed, by Lemma 1, Z_j and \dot{Z}_k have the same upper bounds as J and \dot{J} , respectively.

iii) We have

$$\ddot{S}_n(\theta) = \frac{-1}{2n(n-1)} \sum_{k \neq j} \int \left[\ddot{Z}_k(\theta, u) Z_j(\theta, u) + \dot{Z}_k(\theta, u) \dot{Z}_j(\theta, u)^\top \right] dW(u).$$

We shall bound from above as follows

$$\begin{aligned} \|\ddot{S}_n(\theta, u) - \ddot{S}_n(\theta', u)\|_2 &\leq \frac{1}{2n(n-1)} \sum_{k \neq j} \left\{ \left\| \int (\ddot{Z}_k(\theta, u) - \ddot{Z}_k(\theta', u)) Z_j(\theta, u) dW(u) \right\|_2 \right. \\ &\quad + \left\| \int \ddot{Z}_k(\theta', u) (Z_j(\theta, u) - Z_j(\theta', u)) dW(u) \right\|_2 \\ &\quad + \left\| \int \dot{Z}_k(\theta, u) (\dot{Z}_j(\theta, u) - \dot{Z}_j(\theta', u))^\top dW(u) \right\|_2 \\ &\quad \left. + \left\| \int (\dot{Z}_k(\theta, u) - \dot{Z}_k(\theta', u)) \dot{Z}_j(\theta', u)^\top dW(u) \right\|_2 \right\}. \end{aligned}$$

For each term in the previous sum, we use Taylor expansion and Lemmas 1 and 2 to get

$$\left\| \ddot{S}_n(\theta, u) - \ddot{S}_n(\theta', u) \right\|_2 \leq \|\theta - \theta'\| \frac{C \int (1 + |u| + u^2 + |u|^3) dW(u)}{(1 - 2P)^5},$$

for some constant $C > 0$, which finishes the proof by our Assumption A. ■

Proof of Theorem 2. Our method is based on a consistency proof for minimum contrast estimators by Dacunha-Castelle and Duflo (1993, pp.94–96). Let us consider a countable dense set D in Θ , then $\inf_{\theta \in \Theta} S_n(\theta) = \inf_{\theta \in D} S_n(\theta)$, is a measurable random variable. We define in addition the random variable

$$W(n, \xi) = \sup \{ |S_n(\theta) - S_n(\theta')|; (\theta, \theta') \in D^2, \|\theta - \theta'\| \leq \xi \},$$

and recall that $S(\theta_0) = 0$. Let us consider a non-empty open ball B_0 centered on θ_0 such that S is bounded from below by a positive real number 2ε on $\Theta \setminus B_0$. Let us consider a sequence $(\xi_p)_{p \geq 1}$ decreasing to zero, and take p such that there exists a covering of $\Theta \setminus B_0$ by a finite number ℓ of balls $(B_i)_{1 \leq i \leq \ell}$ with centers $\theta_i \in \Theta$, $i = 1, \dots, \ell$, and radius less than ξ_p . Then, for all $\theta \in B_i$, we have

$$\begin{aligned} S_n(\theta) &\geq S_n(\theta_i) - |S_n(\theta) - S_n(\theta_i)| \\ &\geq S_n(\theta_i) - \sup_{\theta \in B_i} |S_n(\theta) - S_n(\theta_i)|, \end{aligned}$$

which leads to

$$\inf_{\theta \in \Theta \setminus B_0} S_n(\theta) \geq \inf_{1 \leq i \leq \ell} S_n(\theta_i) - W(n, \xi_p).$$

As a consequence we have the following events inclusions

$$\begin{aligned}
\{\hat{\theta}_n \notin B_0\} &\subseteq \left\{ \inf_{\theta \in \Theta \setminus B_0} S_n(\theta) < S_n(\theta_0) \right\} \\
&\subseteq \left\{ \inf_{1 \leq i \leq \ell} S_n(\theta_i) - W(n, \xi_p) < S_n(\theta_0) \right\} \\
&\subseteq \{W(n, \xi_p) > \varepsilon\} \cup \left\{ \inf_{1 \leq i \leq \ell} S_n(\theta_i) - S_n(\theta_0) \leq \varepsilon \right\}.
\end{aligned}$$

Thus we have

$$\{\hat{\theta}_n \notin B_0\} \subseteq \{W(n, \xi_p) > \varepsilon\} \cup \left\{ \inf_{1 \leq i \leq \ell} (S_n(\theta_i) - S_n(\theta_0)) \leq \varepsilon \right\}. \quad (17)$$

By the convergence given in Proposition 2 we have

$$\begin{aligned}
&P\left(\inf_{1 \leq i \leq \ell} (S_n(\theta_i) - S_n(\theta_0)) \leq \varepsilon\right) \\
&\leq 1 - \prod_{i=1}^{\ell} (1 - P(S_n(\theta_i) - S(\theta_0) \leq \varepsilon)) \\
&\leq 1 - \prod_{i=1}^{\ell} (1 - P(S_n(\theta_i) - S(\theta_i) + S_n(\theta_0) - S(\theta_0) \leq -\varepsilon)) \\
&\leq 1 - \prod_{i=1}^{\ell} (1 - [P(|S_n(\theta_i) - S(\theta_i)| \geq \varepsilon) + P(|S_n(\theta_0) - S(\theta_0)| \geq \varepsilon)]),
\end{aligned}$$

where the last term in the right hand side of the above inequality vanishes to zero according to Proposition 2. Because S_n is Lipschitz over Θ by Lemma 3, we have that for sufficiently large p , $|S_n(\theta) - S_n(\theta')| \leq \varepsilon/2$ for all (θ, θ') such that $|\theta - \theta'|_2 \leq \xi_p$, thus $P(W(n, \xi_p) > \varepsilon) = 0$. We just proved the consistency in probability of the contrast estimator $\hat{\theta}_n$ defined in (7). ■

Proof of Theorem 3. By a Taylor expansion of \dot{S}_n around θ_0 , we have

$$0 = \dot{S}_n(\hat{\theta}_n) = \dot{S}_n(\theta_0) + \ddot{S}_n(\theta_n^*)(\hat{\theta}_n - \theta_0), \quad (18)$$

where θ_n^* lies in the line segment with extremities $\hat{\theta}_n$ and θ_0 .

Step 1. Let us prove that

$$\dot{S}_n(\theta_0) = \frac{-1}{2n(n-1)} \sum_{j \neq k, j, k=1}^n \int \dot{Z}_k(\theta, u) Z_j(\theta, u) dW(u) \quad (19)$$

is asymptotically normal, *i.e.* $\sqrt{n}\dot{S}_n(\theta_0) \xrightarrow{d} N(0, V)$. Indeed noticing that $\dot{S}(\theta_0) = 0$ and $J(\theta_0, u) = 0$, for all $u \in \mathbb{R}$, we obtain

$$E[\dot{S}_n(\theta_0)] = -\frac{1}{2} \int \dot{J}(\theta_0, u)J(\theta_0, u)dW(u) = 0.$$

Therefore, we can decompose $\dot{S}_n(\theta_0)$ as follows:

$$\begin{aligned} \dot{S}_n(\theta_0) &= \frac{-1}{2n(n-1)} \sum_{j \neq k, j, k=1}^n \int \dot{Z}_k(\theta_0, u)Z_j(\theta_0, u)dW(u) \\ &= \frac{-1}{2n(n-1)} \sum_{j \neq k, j, k=1}^n \int [\dot{Z}_k(\theta_0, u) - \dot{J}(\theta_0, u)] Z_j(\theta_0, u)dW(u) \\ &\quad - \frac{1}{2n} \sum_{k=1}^n \int Z_k(\theta_0, u)\dot{J}(\theta_0, u)dW(u) =: A_n + B_n. \end{aligned}$$

We shall see that $\sqrt{n}B_n$ gives the dominant behaviour in the limit in distribution. Indeed, we can remark that

$$\begin{aligned} \|n\text{Var}(A_n)\| &\leq \frac{1}{4(n-1)} \left\| E \left[\left(\int \dot{Z}_1(\theta_0, u)Z_2(\theta_0, u)dW(u) \right) \left(\int \dot{Z}_1(\theta_0, u)Z_2(\theta_0, u)dW(u) \right)^\top \right] \right\| \\ &\leq \frac{C}{(1-2P)^6n} \left(\int (1+|u|)dW(u) \right)^2 = o(1), \end{aligned}$$

when the asymptotic behaviour of the distribution of $\sqrt{n}B_n$ is obtained by the central limit theorem. We write

$$\sqrt{n}B_n := -\frac{1}{2\sqrt{n}} \sum_{k=1}^n U_k(\theta_0),$$

where $U_k(\theta_0) = \int_{\mathbb{R}} Z_k(\theta_0, u)\dot{J}(\theta_0, u)dW(u)$, for $k = 1, \dots, n$, is a collection of i.i.d. and centered random variables not depending on n . Therefore,

$$\frac{1}{2\sqrt{n}} \sum_{k=1}^n U_k(\theta_0) \xrightarrow{d} \mathcal{N}(0, V), \quad n \rightarrow \infty,$$

where V denotes covariance matrix of $U_1(\theta_0)$ which is equal to $E(U_1(\theta_0)U_1(\theta_0)^T)/4$ (and cannot be explicitated due to the integral nature of the terms).

Step 2. Let us prove that

$$\ddot{S}_n(\theta_n^*) \xrightarrow{P} \mathcal{I}(\theta_0), \quad n \rightarrow \infty. \quad (20)$$

where $\mathcal{I} = \mathcal{I}(\theta_0) = -\frac{1}{2} \int \dot{J}(\theta_0, u) \dot{J}^\top(\theta_0, u) dW(u)$. We start by writing the triangular inequality

$$\|\ddot{S}_n(\theta_n^*) - \mathcal{I}\| \leq \|\ddot{S}_n(\theta_n^*) - \ddot{S}_n(\theta_0)\| + \|\ddot{S}_n(\theta_0) - \mathcal{I}\|.$$

Then we use the Lipschitz property of \ddot{S}_n stated in Lemma 2, and the convergence in probability of $\hat{\theta}_n$ to θ_0 established in Theorem 2. Finally, we compute the limit of $\ddot{S}_n(\theta_0)$ and check that

$$\begin{aligned} E(\ddot{S}_n(\theta_0)) &= -\frac{1}{2} \int (\ddot{J}(\theta_0, u) J(\theta_0, u) + \dot{J}(\theta_0, u) \dot{J}(\theta_0, u)^\top) dW(u) \\ &= -\frac{1}{2} \int \dot{J}(\theta_0, u) \dot{J}(\theta_0, u)^\top dW(u), \end{aligned}$$

as $J(\theta_0, u) = 0$. We then see that $E(\ddot{S}_n(\theta_0)) = \mathcal{I}(\theta_0)$. ■

The following Lemma establishes the asymptotic equivalence between $\hat{\theta}_n$ and $\tilde{\theta}_n$ respectively defined in (7) and (10).

Lemma 4 *i) The Monte Carlo evaluation $\tilde{S}_n(\theta)$ of $S_n(\theta)$ as given by (9) is such that*

$$\sup_{\theta \in \Theta} E((\tilde{S}_n(\theta) - S_n(\theta))^2) = O(n^{-1}).$$

$$ii) \sqrt{n} \|\dot{\tilde{S}}_n(\theta_0) - \dot{S}_n(\theta_0)\| = O_P(n^{-1}).$$

$$iii) \|\ddot{\tilde{S}}_n(\theta_0) - \ddot{S}_n(\theta_0)\| = O_P(n^{-1}) \text{ and } \ddot{\tilde{S}}_n \text{ is Lipschitz over } \Theta.$$

$$iv) \tilde{\theta}_n \text{ defined in (10) satisfies Theorems 2 and 3 when replacing } \hat{\theta}_n \text{ by } \tilde{\theta}_n.$$

Proof. i) Let denote for all $\ell = 1, \dots, n$, and all $\theta \in \Theta$

$$F(Y_\ell, X_1^n, \theta) := \frac{-1}{4n(n-1)} \sum_{j \neq k, j, k=1}^n Z_k(\theta, Y_\ell) Z_j(\theta, Y_\ell),$$

and remark that according to Lemma 1,

$$\begin{aligned} |F(Y_\ell, X_1^n, \theta)| &\leq \frac{1}{4n(n-1)} \sum_{j \neq k, j, k=1}^n \sup_{\theta \in \Theta, u \in \mathbb{R}} |Z_k(\theta, u)| \sup_{\theta \in \Theta, u \in \mathbb{R}} |Z_j(\theta, u)| \\ &\leq \frac{1}{(1-2P)^2}. \end{aligned}$$

We thus have

$$\begin{aligned}
E((\tilde{S}_n(\theta) - S_n(\theta))^2) &= E(E(\tilde{S}_n(\theta) - S_n(\theta))^2 | X_1^n) = E(\text{Var}(\tilde{S}_n(\theta) | X_1^n)) \\
&= \frac{1}{n^2} \sum_{\ell=1}^n E(E([F(Y_\ell, X_1^n, \theta) - E(F(Y_\ell, X_1^n, \theta))]^2 | X_1^n)) \\
&\leq \frac{2}{n^2} \sum_{\ell=1}^n E(E(F^2(Y_\ell, X_1^n, \theta) | X_1^n) + E((E(F(Y_\ell, X_1^n, \theta)))^2 | X_1^n)) \\
&\leq \frac{2}{n(1-2P)^4},
\end{aligned}$$

which concludes the proof.

ii) Consider \tilde{A}_n and \tilde{B}_n the Y -empirical versions of A_n and B_n defined in step 1 of the proof of Theorem 3. It is straightforward to prove, using arguments similar to those used in the proof of i), that $\|\text{Var}(\sqrt{n}(\tilde{A}_n - A_n))\| = O(n^{-1})$ and $\|\text{Var}(\sqrt{n}(\tilde{B}_n - B_n))\| = O(n^{-1})$ which leads to the wanted result.

iii) The first part, respectively the second part of iii), is obtained by arguments similar to those used in the proof of i), resp. the proof of Lemma 3 iii).

iv) Using i), respectively the approximations ii) and iii), we prove Theorem 2, resp. Theorem 3, when replacing $\hat{\theta}_n$ by $\tilde{\theta}_n$. ■

Proof of the Theorem 4. Note first that

$$\begin{aligned}
E(f_n(x)) &= E\left(\frac{1}{2\pi} \int e^{-iux} \frac{1}{n} \sum_{k=1}^n \frac{e^{iuX_k} K^*(b_n u)}{M(\hat{\theta}_{n,-k}, u)} du\right) \\
&= \frac{1}{2\pi} \int e^{-iux} g^*(u) K^*(b_n u) E\left(\frac{1}{M(\hat{\theta}_{n,-1}, u)}\right) du.
\end{aligned}$$

Recall that $\sup_{\theta \in \Theta} |M(\theta, u)| \geq 1 - 2P$, which means that $E(M^{-1}(\hat{\theta}_{n,-1}, u)) \leq (1 - 2P)^{-1}$.

Let us write the usual bias-variance decomposition. For the bias, we have

$$\begin{aligned}
E(f_n(x)) - f(x) &= \frac{1}{2\pi} \int e^{-iux} g^*(u) \left(K^*(b_n u) E\left(\frac{1}{M(\hat{\theta}_{n,-1}, u)}\right) - \frac{1}{M(\theta_0, u)} \right) du \\
&= \frac{1}{2\pi} \int e^{-iux} g^*(u) K^*(b_n u) \left(E\left(\frac{1}{M(\hat{\theta}_{n,-1}, u)}\right) - \frac{1}{M(\theta_0, u)} \right) du \\
&\quad + \frac{1}{2\pi} \int e^{-iux} \frac{g^*(u)}{M(\theta_0, u)} (K^*(b_n u) - 1) du.
\end{aligned}$$

Next, we use the facts that $|\sup_u K^*(u)| \leq 1$ and that the support of $K^*(b_n u)$ is included in the set $\{u : |u| \geq 1/b_n\}$. We then obtain

$$\begin{aligned} |E(f_n(x)) - f(x)| &\leq \frac{1}{2\pi} \left(\int |g^*(u)| |E(M^{-1}(\hat{\theta}_{n,-1}), u) - M^{-1}(\theta_0, u)| du \right. \\ &\quad \left. + \frac{1}{1-2P} \int_{|u| \geq 1/b_n} |g^*(u)| du \right) \\ &= O\left(\frac{1}{\sqrt{n}}\right) + O(1) \frac{b_n^{\beta-1/2}}{1-2P}, \end{aligned}$$

which gives us the wanted result.

For the variance term, we write

$$\begin{aligned} \text{Var}(f_n(x)) &= E \left[\left(\frac{1}{2\pi n} \sum_{k=1}^n \int e^{-iux} K^*(b_n u) \left(\frac{e^{iuX_k}}{M(\hat{\theta}_{n,-k}, u)} - g^*(u) E \left(\frac{1}{M(\hat{\theta}_{n,-1}, u)} \right) \right) du \right)^2 \right] \\ &\leq \frac{1}{4\pi^2 n} E \left[E \left[\left(\int K^*(b_n u) \frac{e^{iuX_1}}{M(\hat{\theta}_{n,-1}, u)} du \right)^2 \middle| X_2, \dots, X_n \right] \right] \\ &\leq \frac{1}{4\pi^2 n} E \left[\left(\int K^*(b_n u) \frac{g^*(u)}{M(\hat{\theta}_{n,-1}, u)} du \right)^2 \right] \leq \frac{\|K^*\|_2^2 \|g^*\|_2^2}{4\pi^2 (1-2P)^2 n b_n}. \end{aligned}$$

Therefore, by taking $b_n = cn^{-(\beta-1/2)/(2\beta)}$, we obtain the upper bound in our theorem. ■

References

- [1] AZZALINI, A. AND BOWMAN, A. W. (1990). A look at some data on the Old Faithful geyser. *Appl. Statist.* **39** 357–365.
- [2] BERAN, R. (1978) An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313.
- [3] BORDES, L., MOTTELET, S. and VANDEKERKHOVE, P. (2006a). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232.
- [4] BUTUCEA, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations; *Ann. Statist.* **35**, 5, 1907–1930.

- [5] CERRITO, P. B. (1992). Using stratification to estimate multimodal density functions with applications to regression. *Comm. Statist. Simulation Comput.* **21** 1149–1164.
- [6] CHEN, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233.
- [7] DACUNHA-CASTELLE, D. AND GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209.
- [8] EVERITT, B. S. AND HAND, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.
- [9] HALL, P. (1981). On the nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **43** 147–156.
- [10] HALL, P., AND ZHOU, X-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224.
- [11] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. AND TSYBAKOV, A.B. (1998). Wavelets, Approximation and Statistical Applications. *Lecture Notes in Statistics*. **129**. Springer
- [12] HUNTER, D. R., WANG, S. AND HETTMANSPEGER, T. P. (2004). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251.
- [13] LANCASTER, T., AND IMBENS, G. (1996). Case-control studies with contaminated controls. *J. Econometrics* **71** 145–160.
- [14] LEMDANI, M. AND PONS, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli* **5** 705–719.
- [15] LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360.
- [16] MCLACHLAN, G. J. AND PEEL, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

- [17] MCNEIL, D. R. (1977). *Interactive Data Analysis*. Wiley, New York.
- [18] MURRAY, G. D. AND TITTERINGTON, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist.* **27** 325–334.
- [19] QIN, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. Statist.* **27** 1368–1384.
- [20] SCOTT, D. W. (2008). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons
- [21] TITTERINGTON, D. M., SMITH, A. F. M. AND MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester.

Pierre Vandekerkhove, Département d'Analyse et de Mathématiques Appliquées, Université Paris-Est Marne-la-Vallée, 5 Bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2.

Email: pierre.vandek@univ-mlv.fr