Non-asymptotic detection of two-component mixtures with unknown means

Béatrice Laurent and Clément Marteau and Cathy Maugis-Rabusseau

Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse
INSA de Toulouse,
135, avenue de Rangueil,
31077 Toulouse Cedex 4, France.

Abstract: This work is concerned with the detection of a mixture distribution from a $\mathbb{R}$-valued sample. Given a sample $X_1, \ldots, X_n$ and an even density ϕ, our aim is to detect whether the sample distribution is $\phi(. - \mu)$ for some unknown mean μ, or is defined as a two-component mixture based on translations of ϕ. In a first time, a non-asymptotic testing procedure is proposed and we determine conditions under which the power of the test can be controlled. In a second time, the performances of our testing procedure are investigated in 'benchmark' asymptotic settings. A simulation study provides comparisons with classical procedures.

AMS 2000 subject classifications: 62G10,62G30.
Keywords and phrases: Non-asymptotic testing procedure, mixtures, order statistics, separation rates, Higher Criticism.

1. Introduction

In this paper, the detection problem of a mixture distribution from a $\mathbb{R}$-valued sample is considered. Let $(X_1, \ldots, X_n)$ be i.i.d. random variables from an unknown distribution F. All along the paper, F is assumed to admit a density f w.r.t. the Lebesgue measure on $\mathbb{R}$. The sample is said to be distributed from a mixture when f belongs to

$$\mathcal{F}_1 = \left\{ x \in \mathbb{R} \mapsto (1-\varepsilon)\phi(x-\mu_1) + \varepsilon\phi(x-\mu_2); \quad \varepsilon \in]0,1[, (\mu_1, \mu_2) \in \mathbb{R}^2, \ \mu_1 < \mu_2 \right\} \tag{1}$$

where $\phi(.)$ denotes a density. In this paper, $\phi(.)$ is assumed to be an even known density, and when Gaussian mixtures are considered, $\phi(.) = \phi_G(.)$ with

$$\phi_G(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \forall x \in \mathbb{R}.$$

For a complete introduction about mixtures, we refer to McLachlan and Peel (2000). The two-component mixtures are often encountered in practice, for instance in biology and medecine. They allow to model situations where a population can be discrimined into two different groups. The first subpopulation is then assumed to be distributed following the density $\phi(. - \mu_1)$ while the second one follows the density $\phi(. - \mu_2)$. The probability that an observation X_i arises from the first (resp. the second) subpopulation is then modeled by $1 - \varepsilon$ (resp. ε).

This model has been intensively studied and many pathes have been explored in order to provide a satisfying inference. In particular, the detection problem has attracted a lot of attention in the last two decades. The main goal is not to provide the best estimation of the parameters of interest

(ϵ, μ_1, μ_2) but rather to decide whether the incoming observations are following a mixture distribution or not. In other words, one wants to detect if the sample of interest comes from a homogeneous or heterogeneous population. Let $\mathcal{F}_0$ be the density set defined as

$$\mathcal{F}_0 = \{x \in \mathbb{R} \mapsto \phi(x-\mu);\ \mu \in \mathbb{R}\}. \tag{2}$$

Formally, one wants to test

$$\text{"}f \in \mathcal{F}_0\text{" against "}f \in \mathcal{F}_1\text{"}. \tag{3}$$

In various testing problems involving finite mixtures, the properties of the likelihood ratio test have been widely investigated. We can mention for instance Chernoff and Lander (1995), Dacunha-Castelle and Gassiat (1999), Azaïs et al. (2009) or Garel (2007) among others. In all these papers, the main challenge is to determine the asymptotic behaviour of the likelihood ratio under the alternative hypothesis in order to investigate the power of the related test. Alternative methods have also been considered: modified likelihood ratio test by Chen et al. (2001), estimation of the L^2 distance between density associated to null and alternative hypotheses by Charnigo and Sun (2004), EM approach in Chen and Li (2009) or tests based on the empirical characteristic function in Klar and Meintanis (2005).

The main challenge related to the problem (3) is to find (optimal) conditions on $(\varepsilon, \mu_1, \mu_2)$ for which a prescribed second kind error can be achieved. The first study in this way is due to Ingster (1999) in the particular case where the mean μ under the null hypothesis is known, $\mu_1 = \mu$ in the alternatives and $\phi(.)$ corresponds to a Gaussian density. Similar results have also been obtained in Donoho and Jin (2004). In this last paper, the so-called Higher Criticism has been investigated. This algorithm is very powerful in the sense that it is easy to implement, and provides similar power than the usual likelihood ratio test. The asymptotic detection regions have been carefully investigated in two different asymptotic regimes:

- the *sparse regime* where $\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta}$ and $\mu_2 - \mu_1 \underset{n\to+\infty}{\sim} \sqrt{2r\log(n)}$ with $\frac{1}{2} < \delta < 1$ and $0 < r < 1$. In such a case, it is proved that the two hypotheses can be asymptotically separated if
$$\begin{cases} r > \delta - \frac{1}{2} & \text{when } \frac{1}{2} < \delta \leq \frac{3}{4} \\ r > (1-\sqrt{1-\delta})^2 & \text{when } \frac{3}{4} < \delta < 1 \end{cases};$$
- the *dense regime* where $\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta}$ and $\mu_2 - \mu_1 \underset{n\to+\infty}{\sim} n^{-r}$ with $0 < \delta \leq \frac{1}{2}$ and $0 < r < \frac{1}{2}$. In this framework, the separation is asymptotically possible if $r < \frac{1}{2} - \delta$.

We refer for more details to Ingster (1999) and Donoho and Jin (2004). Jager and Wellner (2007) proposed a family of tests based on the Renyi divergences which generalizes the procedure based on the Higher Criticism.

We also mention that generalizations of this procedure to heteroscedastic mixtures have been proposed in Cai et al. (2011) while Cai et al. (2007) consider the problems of estimation and construction of confidence sets in sparse mixture models. Addario-Berry et al. (2010) determine non-asymptotic separation rates of testing for the contamination of a standard Gaussian vector in $\mathbb{R}^n$ by non-zero mean components when the alternatives have particular combinatorial and geometric structures. More recently, Cai and Wu (2012) consider the detection of sparse mixtures in the situation where the density of the observations under the null hypothesis is fixed, but not necessarily Gaussian.

In this paper, we consider a testing problem where the null hypothesis does not correspond to a fixed density but rather to the set of densities $\mathcal{F}_0$ defined by (2) which corresponds to a translation model. Thus the mean parameter μ under the null hypothesis is not assumed to be known. The considered alternative $\mathcal{F}_1$ corresponds to the set of densities that are mixtures of two densities of $\mathcal{F}_0$. Our aim is to decide whether the density f of the observations belongs to $\mathcal{F}_0$ or $\mathcal{F}_1$. To this end, we introduce a new testing procedure based on the ordered statistics. Contrary to the Higher Criticism algorithm (Donoho and Jin, 2004), the main advantage of this procedure is that the knowledge of the mean μ under H_0 is not required. Since one can find densities in $\mathcal{F}_1$ that are arbitrary close to $\mathcal{F}_0$, it is impossible to build a level-α test that achieves a prescribed power on the whole set $\mathcal{F}_1$. Hence, we introduce subsets of $\mathcal{F}_1$, denoted $\bar{\mathcal{F}}_1[n, \alpha, \beta]$, over which our level-$\alpha$ test has a power greater than $1-\beta$. The performances of our procedure are therefore non asymptotic. In the Gaussian case, this result is completed by a non-asymptotic lower bound, proving the optimality of our procedure (up to a logarithmic term).

Then, the asymptotic performances of our testing procedure will be investigated considering Gaussian mixtures and the so-called *sparse* and *dense* regimes as Donoho and Jin (2004) and Cai et al. (2011). In particular, we will see that the detection regions are slightly different than in Cai et al. (2011), due to the fact that the mean μ under H_0 is unknown.

The paper is organized as follows. In Section 2, a testing procedure based on the ordered statistics is introduced. Then the non-asymptotic behaviour of this test is investigated and we propose a general separation set $\bar{\mathcal{F}}_1[n, \alpha, \beta]$. In Section 3, we provide non-asymptotic lower and upper bounds in a Gaussian mixture model that enhance the quasi-optimality of our separation set. An asymptotic study is proposed in Section 4. Some numerical simulations, providing a comparison with existing procedures are displayed in Section 5. Proofs are gathered in Section 6 and technical lemmas in Appendix.

2. The testing procedure

Recall that given an i.i.d. sample $X_1, \dots, X_n$ having a common density f w.r.t. the Lebesgue measure on $\mathbb{R}$, our aim is to consider the testing problem $H_0 : f \in \mathcal{F}_0$ against $H_1 : f \in \mathcal{F}_1$, namely to decide whether f corresponds to a given even density function ϕ (up to a translation) or is defined as a two-component mixtures of translations of ϕ.

In this context, one of the most popular testing procedure is the Higher Criticism introduced in Donoho and Jin (2004), whose asymptotic behaviour has been widely investigated (see also references above). Nevertheless, there exists up to our knowledge no description of the non-asymptotic performances of this algorithm. Moreover this procedure heavily depends on the knowledge of the mean under H_0. In this paper, we work in a slightly different framework in the sense that a translation model under H_0 is considered.

In this section, a new testing procedure based on spacing of order statistics is proposed. The order statistics are denoted by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. The main underlying idea is that the spacing of these order statistics are free with respect to the mean under H_0: for some $k < l \in \{1, \dots, n\}$, the mean value affects the spatial position of a given $X_{(k)}$, but not $X_{(l)} - X_{(k)}$. Moreover, the distribution of the variables $X_{(l)} - X_{(k)}$ is known under H_0 and has a different behavior under H_1, provided k and l are well-chosen.

Let $\alpha \in]0, 1[$ be a fixed level. In the following, a level-α test function ψ_α denotes a measurable function of $(X_1, \dots, X_n)$ with value in $\{0, 1\}$, such that the null hypothesis is rejected if $\psi_\alpha = 1$

and $\sup_{f\in\mathcal{F}_0}\mathbb{P}_f(\psi_\alpha=1)\leq\alpha$. Assume that $n\geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1,2,\dots,n/2\}$ defined as

$$\mathcal{K}_n=\{2^j, 0\leq j\leq[\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k\in\mathcal{K}_n}\left\{\mathbb{1}_{X_{(n-k+1)}-X_{(k)}>q_{\alpha_n,k}}\right\}, \tag{4}$$

where, for all $u\in]0,1[$, $q_{u,k}$ is the $(1-u)$-quantile of $X_{(n-k+1)}-X_{(k)}$ under the null hypothesis and

$$\alpha_n=\sup\left\{u\in]0,1[,\mathbb{P}_{H_0}\left(\exists k\in\mathcal{K}_n, X_{(n-k+1)}-X_{(k)}>q_{u,k}\right)\leq\alpha\right\}.$$

Note that since the distribution of $X_{(n-k+1)}-X_{(k)}$ under the null hypothesis is independent of the mean value μ, $q_{\alpha_n,k}$ and α_n can be approximated (via Monte-Carlo simulations for instance) under the assumption that the X_i's have common density ϕ. In the following we also provide explicit upper bounds for the quantiles, which can be used instead of the true $q_{\alpha,k}$ if necessary.

By definition, the test statistics Ψ_α is exactly of level α. We now want to evaluate the power of the test. More precisely, we want to define subsets of the alternative $\mathcal{F}_1$ over which the test has a prescribed power. We first need some definitions.

Let $\bar{\Phi}(x)=1-\Phi(x)$, where Φ is the cumulative distribution function associated to the density function ϕ. For all $\alpha\in]0,1[$ and $k\in\{1,2,\dots,n/2\}$, let $t_{\alpha,k}$ be a positive real defined by

$$\bar{\Phi}\left(\frac{t_{\alpha,k}}{2}\right)=\frac{k}{n}\left[1-\sqrt{\frac{2\log(\frac{4}{\alpha})}{k}}\right] \tag{5}$$

if $k>2\log(\frac{4}{\alpha})$, and $t_{\alpha,k}=+\infty$ otherwise. For all $\alpha\in]0,1[$, $\rho>0$, and $k\in\{1,2,\dots,n/2\}$, we consider the subset $\bar{\mathcal{S}}(\alpha,\rho,k)$ of $\mathbb{R}^3$ defined by :

$$\bar{\mathcal{S}}(\alpha,\rho,k)=\left\{\begin{array}{l}(\varepsilon,\mu_1,\mu_2)\in]0,1[\times\mathbb{R}^2,\mu_2>\mu_1;\exists c\in\mathbb{R}\text{ such that :}\\ (1-\varepsilon)\bar{\Phi}\left(t_{\alpha,k}-c+\varepsilon(\mu_2-\mu_1)\right)+\varepsilon\bar{\Phi}\left(t_{\alpha,k}-c-(1-\varepsilon)(\mu_2-\mu_1)\right)>\rho\\ (1-\varepsilon)\bar{\Phi}\left(c-\varepsilon(\mu_2-\mu_1)\right)+\varepsilon\bar{\Phi}\left(c+(1-\varepsilon)(\mu_2-\mu_1)\right)>\rho\end{array}\right\}. \tag{6}$$

When $t_{\alpha,k}=+\infty$, we use the convention $\bar{\mathcal{S}}(\alpha,\rho,k)=\emptyset$ for all $\rho>0$.

The following proposition highlights the non-asymptotic performances of the test Ψ_α.

Theorem 1. *Let $\alpha\in]0,1[$ and $\beta\in]0,1-\alpha[$. Consider the test Ψ_α described in (4). Consider the alternative sets*

$$\bar{\mathcal{F}}_1[n,\alpha,\beta]=\left\{f(.)=(1-\varepsilon)\phi(.-\mu_1)+\varepsilon\phi(.-\mu_2);(\varepsilon,\mu_1,\mu_2)\in\bigcup_{k\in\mathcal{K}_n}\bar{\mathcal{S}}(\alpha_n,\rho(k,n),k)\right\}$$

where $\bar{\mathcal{S}}(\alpha_n,\rho(k,n),k)$ is defined by (6) with

$$\rho(k,n)=\frac{k}{n}+\frac{1+\sqrt{1+2k\beta}}{n\beta}.$$

Then Ψ_α *is a level-*α *test and*

$$\sup_{f\in\bar{\mathcal{F}}_1[n,\alpha,\beta]} \mathbb{P}_f(\Psi_\alpha = 0) \leq \beta.$$

In this theorem, we have defined a set $\bar{\mathcal{F}}_1[n,\alpha,\beta]$ over which the level-α test statistics Ψ_α has a power greater than $1-\beta$. This result holds for all n, it is non-asymptotic. The definition of the set $\bar{\mathcal{S}}(\alpha,\rho,k)$ is quite rough. Nevertheless, it will allow us to describe several situations for which the power of our testing procedure will be assessed, in both asymptotic and non-asymptotic cases. In the next section, explicit and sufficient conditions on $(\varepsilon,\mu_1,\mu_2)$ are given, ensuring that the mixture density $(1-\varepsilon)\phi(.-\mu_1)+\varepsilon\phi(.-\mu_2)$ belongs to $\bar{\mathcal{F}}_1[n,\alpha,\beta]$ when ϕ corresponds to the Gaussian density.

3. A non-asymptotic framework

The aim of this section is to provide sufficient and explicit conditions for which the triplet (ϵ,μ_1,μ_2) belongs to $\bar{\mathcal{S}}(\alpha,\rho,k)$. First, we introduce, for all $\rho>0$ and $M>0$, the separation set $\mathcal{F}_1[\rho,M]$ defined as

$$\mathcal{F}_1[\rho,M] = \{f(.) = (1-\varepsilon)\phi(.-\mu_1)+\varepsilon\phi(.-\mu_2), (\varepsilon,\mu_1,\mu_2)\in\mathcal{S}(\rho,M)\},$$

where

$$\mathcal{S}(\rho,M) = \left\{(\varepsilon,\mu_1,\mu_2)\in]0,1[\times\mathbb{R}^2, \varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2\geq\rho,\ 0<\mu_2-\mu_1\leq M\right\}.$$

When the density of the standard normal distribution is considered ($\phi=\phi_G$), the separation set is denoted $\mathcal{F}_{1,G}[\rho,M]$.

Remarks 1. *In this setting, we assume that the difference between the means* μ_1 *and* μ_2 *in the alternatives is bounded. This upper bound on* $\mu_2-\mu_1$ *is necessary in the proof of Theorem 3 to get a uniform upper bound for the second kind error of the test over the set* $\mathcal{F}_{1,G}[\rho,M]$ *for a suitable value of* ρ*.*

3.1. Lower bound for the detection of a Gaussian mixture model

In this section, we consider the same definitions of non-asymptotic lower bounds for hypotheses testing problems than the ones introduced by Baraud (2002) for signal detection in a Gaussian regression model or Gaussian sequence model. We provide a non-asymptotic lower bound for our testing problem in the case where ϕ corresponds to the standard Gaussian density. In particular, we exhibit values for ρ in $\mathcal{S}(\rho,M)$ for which the two hypotheses H_0 and H_1 cannot be separated with prescribed errors.

Theorem 2. *Let* $\alpha\in]0,1[$ *and* $\beta\in]0,1-\alpha[$*. Let*

$$\rho^\star = \frac{1}{C(M)}\left(\sqrt{\frac{-2\log[c(\alpha,\beta)]}{n}}\sqrt{1+\frac{\log[c(\alpha,\beta)]}{2n}}\right),$$

with $c(\alpha,\beta) = 1 - \frac{(1-\alpha-\beta)^2}{2}$ *and* $C(M) = \sqrt{\frac{1}{2} + \frac{M^2}{6}e^{M^2/4}}$. *Then for all* $\rho \leq \rho^\star$,

$$\beta(\mathcal{F}_{1,G}[\rho, M]) := \inf_{\psi_\alpha} \sup_{f \in \mathcal{F}_{1,G}[\rho,M]} \mathbb{P}_f(\psi_\alpha = 0) \geq \beta,$$

where the infimum is taken over all level-α *test* ψ_α.

Theorem 2 implies that whatever the level-α test ψ_α, if $\rho < \rho^\star$, there exists a density $f \in \mathcal{F}_{1,G}[\rho, M]$ for which $\mathbb{P}_f(\psi_\alpha = 0) \geq \beta$. In particular, testing is not possible if $\mu_2 - \mu_1$ is too small with respect to $\epsilon(1-\epsilon)$. We will show in the following that this condition on (ϵ, μ_1, μ_2) is optimal (up to constant), namely if $(\epsilon, \mu_1, \mu_2) \in \mathcal{S}(c\rho^\star, M)$, for some suitable constant c, then it is possible to construct a test and control the associated second kind error.

3.2. A Testing procedure based on the variance

In this paragraph, we are interested in a simple test based on the variance of the X_i's. We will prove that this test allows us to achieve the lower bound obtained in Theorem 2.

Remark that under H_0, $\mathrm{Var}(X_i) = 1$ while under H_1, $\mathrm{Var}(X_i) = 1 + \varepsilon(1-\varepsilon)(\mu_2 - \mu_1)^2$. Hence, we consider the test ψ_α defined by

$$\psi_\alpha = \mathbf{1}_{\{S_n^2 > v_{\alpha,n}\}}, \text{ where } S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \tag{7}$$

and $v_{\alpha,n}$ denotes the $(1-\alpha)$-quantile of the variable S_n^2 under H_0. Then the following proposition holds.

Proposition 1. *Let* $\alpha \in]0,1[$ *and* $\beta \in]0, 1-\alpha[$. *Assume that the density function* ϕ *has a finite fourth moment:* $\mathbb{E}_\phi[X^4] \leq B$. *Consider that* $f(.) = (1-\varepsilon)\phi(. - \mu_1) + \varepsilon\phi(. - \mu_2)$ *belongs to* $\mathcal{F}_1[\rho, M]$ *with* $M > 0$ *and*

$$\rho \geq C(\alpha, \beta, M, B)/\sqrt{n},$$

where $C(\alpha, \beta, M, B)$ *is a positive constant; namely* $0 < \mu_2 - \mu_1 \leq M$ *and* $\varepsilon(1-\varepsilon)(\mu_2 - \mu_1)^2 > C(\alpha, \beta, M, B)/\sqrt{n}$. *Then*

$$\mathbb{P}_f(\psi_\alpha = 0) \leq \beta.$$

In the Gaussian case, $\mathbb{E}_{\phi_G}[X^4] = 3$. Hence, Proposition 1 assesses the optimality of the lower bound given in Theorem 2. Note that the value of ρ proposed in (8) differs from $\rho^\star$ by constant. Finding optimal constant for our testing problem is a very difficult question that is out of the scope of this paper. For interested reader, we mention the work of Ingster (1999) in a slightly different (asymptotic) setting.

3.3. Upper bound for the testing procedure Ψ_α in the Gaussian case

The goal of this section is to give explicit conditions on $(\varepsilon, \mu_1, \mu_2)$ that will ensure a prescribed power for the test Ψ_α defined in (4), when ϕ is the standard Gaussian density. This will provide a better understanding of the behaviour of the set $\bar{\mathcal{S}}(\alpha, \rho, k)$ introduced in (6) in the Gaussian case.

Theorem 3. *Let $X_1, \dots, X_n$ be i.i.d real random variables with common density f. Let $\alpha \in]0,1[$ and consider the level-α test Ψ_α defined by (4). Let $\beta \in]0, 1-\alpha[$ and $M > 0$. Assume that n fulfills $n \geq 2$ and $\log(4\log_2(n)/\alpha)/n \leq \bar{\Phi}_G(M)/36$.*
Then, there exists a positive constant $C(\alpha, \beta, M)$ depending only on α, β and M, such that if

$$\rho \geq C(\alpha, \beta, M)\sqrt{\frac{\log\log(n)}{n}}, \tag{8}$$

$\mathcal{F}_{1,G}[\rho, M] \subset \bar{\mathcal{F}}_1[n, \alpha, \beta]$, which implies

$$\sup_{f \in \mathcal{F}_{1,G}[\rho,M]} \mathbb{P}_f(\Psi_\alpha = 0) \leq \beta.$$

Note that the value of ρ proposed in (8) differs from the lower bound $\rho^\star$ by a term of order $\sqrt{\log\log n}$. This log log term is due to the multiple (adaptive) testing procedure: the optimal value for $k \in \mathcal{K}_n$ in the test Ψ_α is chosen from the data. Hence this $\sqrt{\log\log n}$ term corresponds to the price to pay in such a setting. This kind of logarithmic loss is quite classical in test theory: see for instance Spokoiny (1996) or Fromont and Laurent (2006) in slightly different settings.

The result given in Proposition 1 above seems even better than the one stated in Theorem 3 since the condition to have a powerful test is $\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2 > C/\sqrt{n}$ instead of $C\sqrt{\log\log(n)}/\sqrt{n}$. Nevertheless, the test based on the variance would fail in the asymptotic sparse regime considered in the next section: this is not satisfying from a practical point of view since our aim is to provide a testing procedure which adapts to all possible situations.

4. Asymptotic results

The main conclusion of the previous (non-asymptotic) part, is that if $\epsilon(1-\epsilon)(\mu_2-\mu_1)^2$ is smaller than $c_{\alpha,\beta,M}/\sqrt{n}$ for a given constant $c_{\alpha,\beta,M}$ and $\mu_2 - \mu_1$ is bounded by M, then testing is impossible. In this part, an asymptotic point of view is adopted: we assume that $n \to +\infty$ and we would like to precise the allowed dependency of (ϵ, μ_1, μ_2) with respect to n. In such a setting, the conclusion of Section 3 can be understood as follows:

- if $\epsilon = o(1/\sqrt{n})$ as $n \to +\infty$, it is (at least) necessary that $\mu_2 - \mu_1 \to +\infty$ as $n \to +\infty$, so that the condition of Theorem 2 is not fulfilled,
- if $\epsilon >> 1/\sqrt{n}$ as $n \to +\infty$, then we can allow $\mu_2 - \mu_1$ to tend to 0.

This justifies partially the frameworks considered in Donoho and Jin (2004), namely *sparse* and *dense* regimes. Indeed, we recall for the sake of convenience that the *sparse* regime is characterized by

$$\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta} \text{ and } \mu_2 - \mu_1 \underset{n\to+\infty}{\sim} \sqrt{2r\log(n)} \text{ with } \frac{1}{2} < \delta < 1 \text{ and } 0 < r < 1, \tag{9}$$

while the *dense* regime corresponds to situations where

$$\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta} \text{ and } \mu_2 - \mu_1 \underset{n\to+\infty}{\sim} n^{-r} \text{ with } 0 < \delta \leq \frac{1}{2} \text{ and } 0 < r < \frac{1}{2}. \tag{10}$$

The notation $a_n \underset{n\to+\infty}{\sim} b_n$ means that $\lim_{n\to+\infty} a_n/b_n = 1$.

Below, we express conditions on δ and r for which the second kind error of the test (4) can be controlled.

4.1. The dense case

The results stated in Theorems 2 and 3 are non-asymptotic. Let us analyse the behaviour of the test Ψ_α defined by (4) from an asymptotic point of view in the dense regime.

Corollary 1. *The detection boundary in the dense regime (10) is* $r^*(\delta) = \frac{1}{4} - \frac{\delta}{2}$*: the detection is possible when* $r < r^*(\delta) = \frac{1}{4} - \frac{\delta}{2}$ *(for* n *large enough, the power of the test (4) is greater than* $1-\beta$*) and impossible if* $r > r^*(\delta)$*.*

The proof of Corollary 1 is omitted since it can be obviously deduced from Theorems 2 and 3.

Results stated in Corollary 1 are therefore different from the one obtained in a dense regime in a contamination framework where $H_0 : f = \phi_G(.)$ against $H_1 : f \in \{(1-\varepsilon)\phi_G(.) + \varepsilon\phi_G(.-\mu); \varepsilon \in]0,1[, \mu \in \mathbb{R}\}$. In this case, as mentioned in Introduction, the detection is possible in the dense regime for $r < \frac{1}{2} - \delta$ (see Ingster, 1999; Donoho and Jin, 2004). This difference is due to the fact that the mean under H_0 is unknown, which makes the testing problem harder.

4.2. Sparse case

Using the same methodology, we can now analyse the performances of our testing procedure in the so-called asymptotic *sparse* regime.

Theorem 4. *Let* $X_1, \dots, X_n$ *be i.i.d real random variables with common density* f*. Let* $\alpha \in]0,1[$ *and consider the level-*α *test* Ψ_α *defined by (4). We consider the case where* $\phi = \phi_G$*.*
We assume that the behaviour of (ϵ, μ_1, μ_2) *is governed by (9) and that* $r > r^*(\delta)$ *with*

$$r^*(\delta) = \begin{cases} \delta - \frac{1}{2} & \text{if } \frac{1}{2} < \delta < \frac{3}{4} \\ \\ (1 - \sqrt{1-\delta})^2 & \text{if } \frac{3}{4} \leq \delta < 1 \end{cases}.$$

Then, setting $f(.) = (1-\varepsilon)\phi_G(.-\mu_1) + \varepsilon\phi_G(.-\mu_2)$*, we have, for* n *large enough,*

$$\mathbb{P}_f(\Psi_\alpha = 0) \leq \beta.$$

In the sparse regime, we recover exactly the separation boundaries that are already known in the case where the null hypothesis is reduced to a standard normal density, and the alternative is the mixture $(1-\varepsilon)\phi_G(.) + \varepsilon\phi_G(.-\mu)$. Hence, the fact that the mean under H_0 is unknown does not affect the difficulty of the related testing problem in this sparse regime.

This proves the optimality of our procedure in the sparse regime. Indeed, the lower bounds established by Ingster (1999); Cai et al. (2011) in the case where the null hypothesis is reduced to the standard Gaussian density also provide lower bounds for our testing problem. This comes from the fact that a level-α test for our testing problem is also a level-α test for testing the null hypothesis "$f = \phi_G$".

5. Simulation study

In this section, we provide some numerical experiments in order to enhance the performances of our testing procedure Ψ_α. Comparisons with the Higher Criticism and the Kolmogorov-Smirnov test are provided. Since these both procedures are not designed for the considered framework (translated model with unknown mean), straightforward modifications are proposed.

5.1. Contamination of ϕ_G

In this section, the problem of detection of sparse heterogeneous mixtures as in Donoho and Jin (2004) is considered: Given $(X_1, \dots, X_n)$, i.i.d random variables with an unknown density function f, our aim is to test

$$H_0 : f(.) = \phi_G(.) \text{ against } H_1 : f \in \{x \mapsto (1-\varepsilon)\phi_G(x) + \varepsilon\phi_G(x-\mu); \mu \in \mathbb{R}, \varepsilon \in]0,1[\}. \tag{11}$$

In this case, our testing procedure Ψ_α described in (4) can be easy adapted as follows:

$$\tilde{\Psi}_\alpha = \sup_{k \in \mathcal{K}_n} \left\{ \mathbb{1}_{X_{(n-k+1)} > q_{\alpha_n,k}} \right\},$$

where $q_{\alpha,k}$ is the $(1-\alpha)$-quantile of $X_{(n-k+1)}$ under the null hypothesis, $\mathcal{K}_n = \{2^j; 0 \le j \le [\log_2(n/2)]\}$ and

$$\alpha_n = \sup\{u \in]0,1[, \mathbb{P}_{H_0}\left(\exists k \in \mathcal{K}_n, X_{(n-k+1)} > q_{u,k}\right) \le \alpha\}.$$

For the sake of brevity, we do not exhibit a theoretical study of the performances of this procedure for the testing problem (11). Indeed, the methodology is rather close to the one proposed in this paper, up to some technical modifications. It is possible to see that this procedure achieves the optimal asymptotic separation set in both the *dense* and *sparse* regimes, as described in Donoho and Jin (2004).

The power of our testing procedure is compared with the one of

- Kolmogorov-Smirnov test:
  The level-α test function is $\psi_{KS,\alpha} = \mathbb{1}_{T_{KS} > q_{KS,\alpha}}$ where
  $$T_{KS} = \sup_{x \in \mathbb{R}} \sqrt{n} |F_n(x) - \Phi_G(x)|$$
  with the empirical distribution function $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le x}$, and $q_{KS,\alpha}$ is the $(1-\alpha)$ quantile of T_{KS} under H_0.
- Higher Criticism (Donoho and Jin, 2004):
  Let $p_i = \mathbb{P}(Z > X_i)$ where $Z \sim \mathcal{N}(0,1)$ for all $i \in \{1, \dots, n\}$ and $p_{(1)} \le p_{(2)} \le \dots \le p_{(n)}$. This test is based on
  $$HC = \max_{1 \le i \le n} \frac{\sqrt{n}\left(\frac{i}{n} - p_{(i)}\right)}{\sqrt{p_{(i)}(1-p_{(i)})}}.$$
  The level-α test function is $\psi_{HC,\alpha} = \mathbb{1}_{HC > q_{HC,\alpha}}$ where $q_{HC,\alpha}$ is the $(1-\alpha)$ quantile of HC under H_0.

In order to study the power of these testing procedures, a Monte-Carlo procedure is considered with $N = 100000$ samples of size $n = 100$ from a mixture distribution $(1-\varepsilon)\phi_G(.) + \varepsilon\phi_G(.-\mu)$ with $\varepsilon \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$ and $\mu \in [0, 10]$. The power functions of these testing procedures in the different scenarios are reported in Figure 1.

It appears that our procedure performs as well as the Higher Criticism when ϵ is small w.r.t. the size of the sample, while the Kolmogorov-Smirnov test possesses a bad behavior. Such a setting is close to the *sparse* regime. Nevertheless, the performances of the Higher Criticism deteriorates as ϵ increases while the power of our test $\tilde{\Psi}_\alpha$ remains stable.

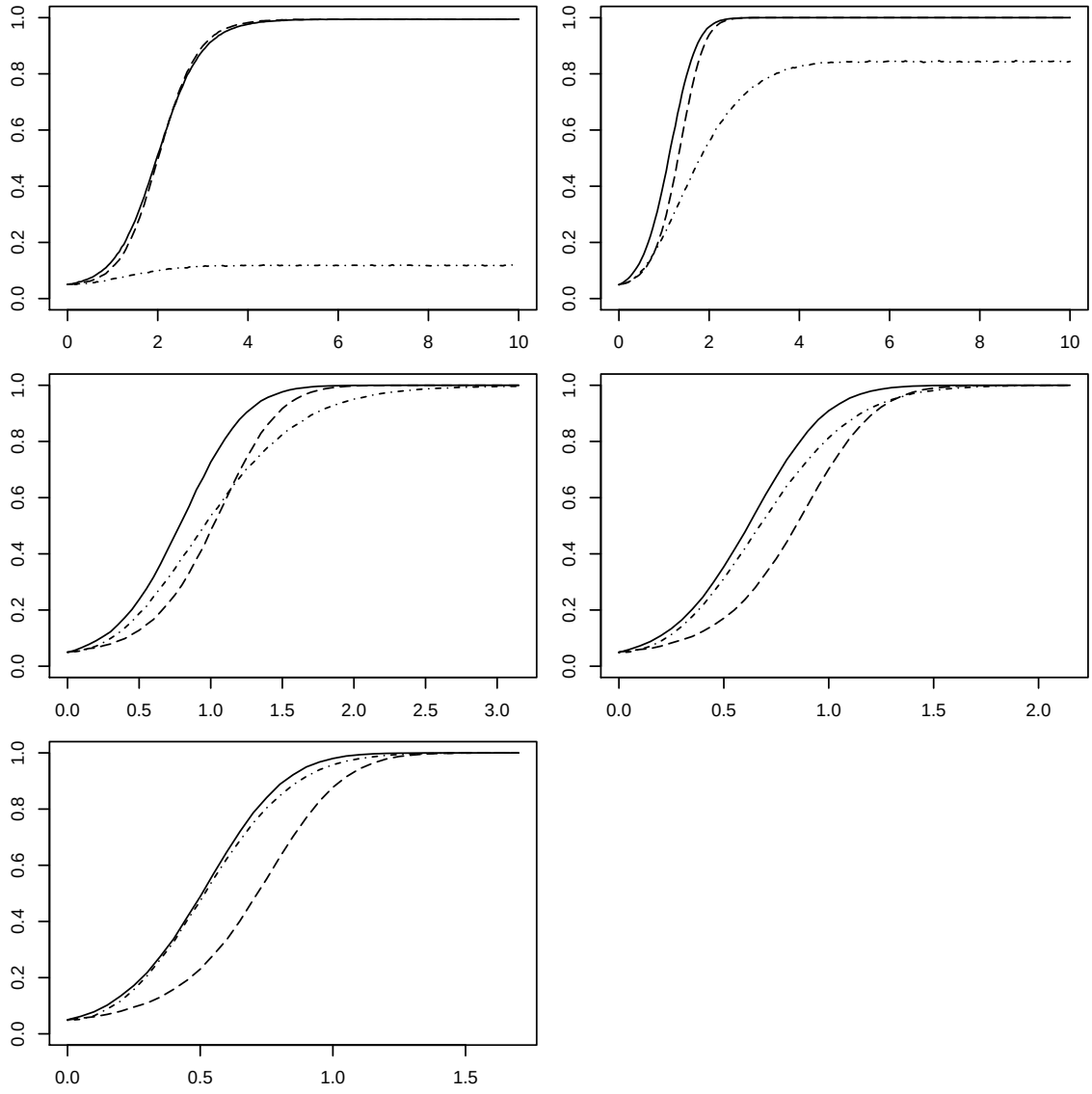

FIG 1. *Power function of the three considered testing procedures (continuous line for our test $\tilde{\Psi}_\alpha$, dashed line for Higher Criticism and dotted line for the Kolmogorov-Smirnov test) according to μ, for $\varepsilon = 0.05$ (top-left),* 0.15 *(top right),* 0.25 *(middle left),* 0.35 *(middle right) and* 0.45 *(bottom left) in a contamination framework.*

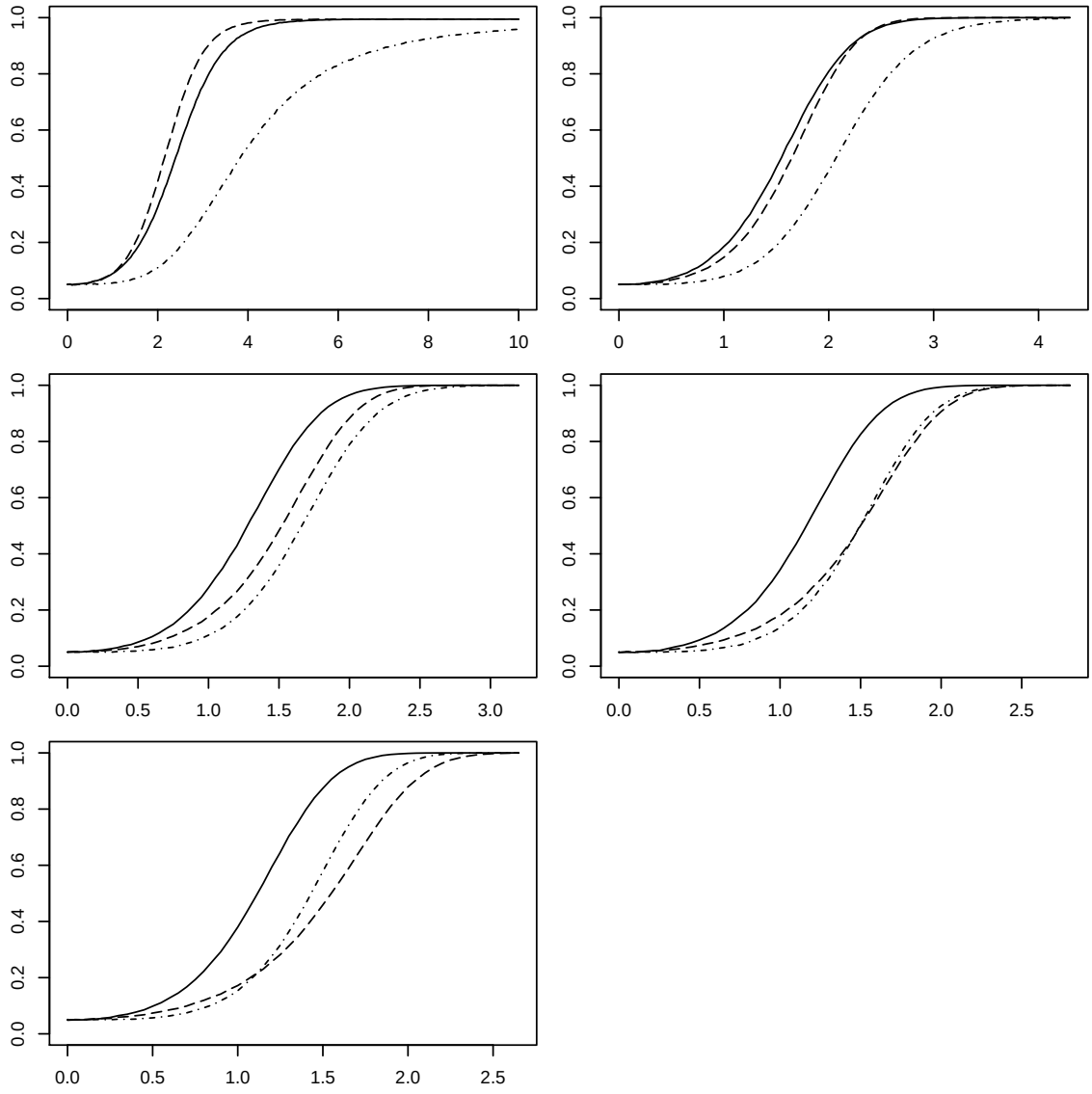

FIG 2. *Power function of the three considered testing procedures (continuous line for our test Ψ_α, dashed line for Higher Criticism and dotted line for the Kolmogorov-Smirnov test) according to μ, for $\varepsilon = 0.05$ (top-left),* 0.15 *(top right),* 0.25 *(middle left),* 0.35 *(middle right) and* 0.45 *(bottom left) in Gaussian mixture framework.*

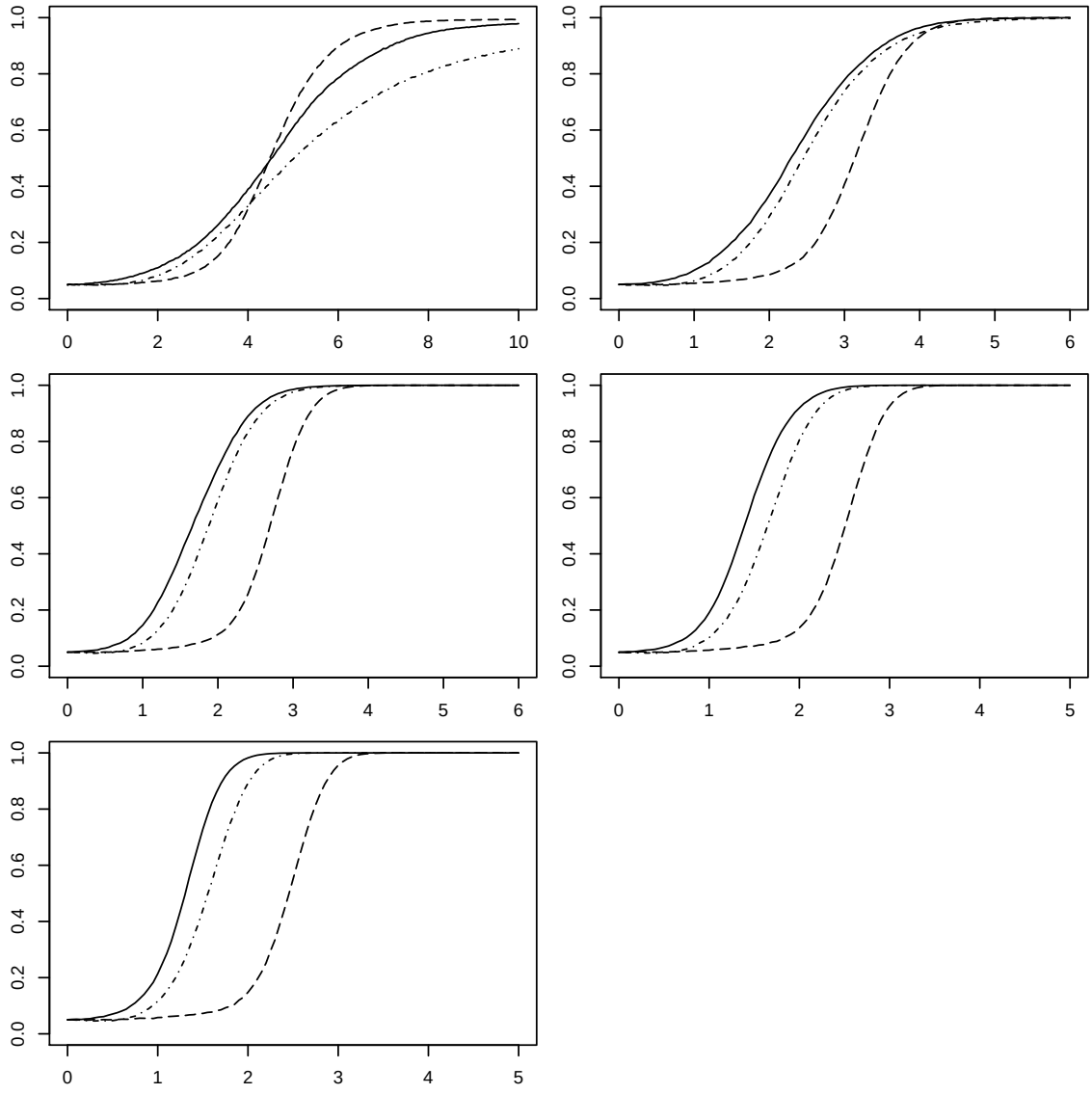

FIG 3. *Power function of the three considered testing procedures (continuous line for our test Ψ_α, dashed line for Higher Criticism and dotted line for the Kolmogorov-Smirnov test) according to μ, for $\varepsilon = 0.05$ (top-left),* 0.15 *(top right),* 0.25 *(middle left),* 0.35 *(middle right) and* 0.45 *(bottom left) in Laplace mixture framework.*

5.2. Gaussian mixtures with unknown means

In this section, we deal with our testing problem. A simulation study is proposed in order to investigate the power of our testing procedure Ψ_α described by (4). Our testing procedure is compared with the following adaptations of Kolmogorov-Smirnov test and Higher Criticism:

- Kolmogorov-Smirnov test:
  The level-α test function is $\widehat{\psi}_{KS,\alpha} = \mathbb{1}_{\hat{T}_{KS} > \hat{q}_{KS,\alpha}}$ where

$$\hat{T}_{KS} = \sup_{x \in \mathbb{R}} \sqrt{n} |F_n(x) - \Phi_G(x - \bar{X})|$$

  with the empirical mean $\bar{X}$, the empirical distribution function $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$, and $\hat{q}_{KS,\alpha}$ is the $(1-\alpha)$ quantile of $\hat{T}_{KS}$ under H_0.
- Higher Criticism (Donoho and Jin, 2004):
  Let $\hat{p}_i = \mathbb{P}(Z - \bar{X} > X_i)$ where $Z \sim \mathcal{N}(0,1)$ for all $i \in \{1, \ldots, n\}$ and $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \ldots \leq \hat{p}_{(n)}$. This test is based on

$$\widehat{HC} = \max_{1 \leq i \leq n} \frac{\sqrt{n}\left(\frac{i}{n} - \hat{p}_{(i)}\right)}{\sqrt{\hat{p}_{(i)}(1 - \hat{p}_{(i)})}}.$$

  The level-α test function is $\hat{\psi}_{HC,\alpha} = \mathbb{1}_{\widehat{HC} > \hat{q}_{HC,\alpha}}$ where $\hat{q}_{HC,\alpha}$ is the $(1-\alpha)$ quantile of $\widehat{HC}$ under H_0.

In order to study the power of these testing procedures, a Monte-Carlo procedure is considered with $N = 100000$ samples of size $n = 100$ from a mixture distribution $(1-\varepsilon)\phi_G(.) + \varepsilon\phi_G(. - \mu)$ with $\varepsilon \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$ and $\mu \in [0, 10]$. The power functions of these testing procedures in the different scenarios are reported in Figure 2.

Once again, our testing procedure appears to be competitive w.r.t. the existing procedures, and even offers better performances in some particular cases. As in the previous experiment, the behavior of the Higher Criticism deteriorates w.r.t. our procedure as ϵ increases, namely when we leave the *sparse* regime to the *dense* one.

5.3. Laplace mixtures with unknown means

Since our test Ψ_α is adapted for an even density function ϕ, a Laplace distribution is here considered: $\phi(x) = \frac{1}{2}\exp(-|x|)$. As in Section 5.2, the power of Ψ_α is compared with the one of Kolmogorov-Smirnov test and Higher Criticism. Note that these two last tests are adapted as in Section 5.2 but where Φ and Z are now associated to the Laplace distribution. A Monte-Carlo procedure is proposed with $N = 100000$ samples of size $n = 100$ from a mixture distribution $(1-\varepsilon)\phi(.) + \varepsilon\phi(. - \mu)$ with $\varepsilon \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$ and $\mu \in [0, 10]$. The power functions of these testing procedures in the different scenarios are reported in Figure 3. Apart in the case where $\varepsilon = 0.05$, our test outperforms Higher Criticism and Kolmogorov-Smirnov in all other conditions. The power of Higher Criticism is deteriorated as ε increases.

6. Proofs

6.1. Proof of Theorem 1

Proof. Following the definition of α_n, Ψ_α is ensured to be a level-α test.

In order to control the second kind error of the test Ψ_α, an upper bound for $q_{\alpha_n,k}$ is first given. Under the null hypothesis, there exists $\mu \in \mathbb{R}$ such that $f(.) = \phi(. - \mu)$. Thus $X_{(n-k+1)} - X_{(k)}$ is distributed as $Y_{(n-k+1)} - Y_{(k)}$ where $(Y_1, \ldots, Y_n)$ is a n sample from the density $\phi(.)$. Hence, if we find $c_{\alpha_n,k}$ such that $\mathbb{P}(Y_{(n-k+1)} - Y_{(k)} > c_{\alpha_n,k}) \leq \alpha_n$ then $q_{\alpha_n,k} \leq c_{\alpha_n,k}$. For all $d \in \mathbb{R}$,

$$\mathbb{P}(Y_{(n-k+1)} - Y_{(k)} > c_{\alpha_n,k}) \leq \mathbb{P}(Y_{(n-k+1)} > c_{\alpha_n,k} + d) + \mathbb{P}(Y_{(k)} \leq d).$$

According to Lemma 1, if d fulfills $\Phi(d) \leq \frac{k}{n}\left[1 - \sqrt{\frac{2\log(\frac{4}{\alpha_n})}{k}}\right]$ then $\mathbb{P}(Y_{(k)} \leq d) \leq \frac{\alpha_n}{2}$. Moreover, by the same lemma, if $c_{\alpha_n,k}$ is chosen such that $\bar{\Phi}(c_{\alpha_n,k} + d) \leq \frac{k}{n}\left[1 - \sqrt{\frac{2\log(\frac{4}{\alpha_n})}{k}}\right]$ then $\mathbb{P}(Y_{(n-k+1)} \geq c_{\alpha_n,k} + d) \leq \frac{\alpha_n}{2}$. Choosing d and $c_{\alpha_n,k}$ exactly such that

$$\Phi(d) = \bar{\Phi}(c_{\alpha_n,k} + d) = \frac{k}{n}\left[1 - \sqrt{\frac{2\log(\frac{4}{\alpha_n})}{k}}\right]$$

and since $\phi(.)$ is an even continuous function, we obtain that $d = -\frac{c_{\alpha_n,k}}{2}$. Finally, choosing $c_{\alpha_n,k} = t_{\alpha_n,k}$ where $\bar{\Phi}(\frac{t_{\alpha_n,k}}{2}) = \frac{k}{n}\left[1 - \sqrt{\frac{2\log(\frac{4}{\alpha_n})}{k}}\right]$, $\mathbb{P}_{H_0}(X_{(n-k+1)} - X_{(k)} > t_{\alpha_n,k}) \leq \alpha_n$ and thus $q_{\alpha_n,k} \leq t_{\alpha_n,k}$.

Considering $f \in \bar{\mathcal{F}}_1[n, \alpha, \beta]$, we want to control the second kind error of the test:

$$\begin{aligned} \mathbb{P}_f(\Psi_\alpha = 0) &= \mathbb{P}_f\left(\forall k \in \mathcal{K}_n, X_{(n-k+1)} - X_{(k)} \leq q_{\alpha_n,k}\right) \\ &\leq \inf_{k \in \mathcal{K}_n} \mathbb{P}_f\left(X_{(n-k+1)} - X_{(k)} \leq q_{\alpha_n,k}\right). \end{aligned} \tag{12}$$

Since $f \in \bar{\mathcal{F}}_1[n, \alpha, \beta]$, there exist $\varepsilon \in]0, 1[$ and $(\mu_1, \mu_2) \in \mathbb{R}^2$, $\mu_1 < \mu_2$ such that

$$\forall x \in \mathbb{R},\ f(x) = (1-\varepsilon)\phi(x - \mu_1) + \varepsilon\phi(x - \mu_2)$$

and for some $k \in \mathcal{K}_n$, there exists a real c such that $(\varepsilon, \mu_1, \mu_2)$ fulfills the two following conditions:

$$(1-\varepsilon)\bar{\Phi}\left(t_{\alpha_n,k} - c + \varepsilon(\mu_2 - \mu_1)\right) + \varepsilon\bar{\Phi}\left(t_{\alpha_n,k} - c - (1-\varepsilon)(\mu_2 - \mu_1)\right) > \rho(k, n), \tag{13}$$

$$(1-\varepsilon)\bar{\Phi}\left(c - \varepsilon(\mu_2 - \mu_1)\right) + \varepsilon\bar{\Phi}\left(c + (1-\varepsilon)(\mu_2 - \mu_1)\right) > \rho(k, n), \tag{14}$$

with $\rho(k, n) = \frac{k}{n} + \frac{1+\sqrt{1+2k\beta}}{n\beta}$. Using (12) and the fact that $q_{\alpha_n,k} \leq t_{\alpha_n,k}$,

$$\begin{aligned} \mathbb{P}_f\left(X_{(n-k+1)} - X_{(k)} \leq q_{\alpha_n,k}\right) &\leq \mathbb{P}_f(X_{(n-k+1)} - X_{(k)} \leq t_{\alpha_n,k}) \\ &\leq \mathbb{P}_f(X_{(n-k+1)} \leq t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c) \\ &\quad + \mathbb{P}_f(X_{(k)} > \mathbb{E}_f[X_1] - c). \end{aligned} \tag{15}$$

For the first term in the right-hand side of (15),

$$\begin{aligned}\mathbb{P}_f\left(X_{(n-k+1)} \leq t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c\right) &\leq \mathbb{P}_f\left(\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t_{\alpha_n,k}+\mathbb{E}_f[X_1]-c\}} > n-k\right) \\ &\leq \mathbb{P}_f\left(\sum_{i=1}^{n} \left\{\mathbb{1}_{\{X_i \leq t_{\alpha_n,k}+\mathbb{E}_f[X_1]-c\}} - q_1\right\} > n(1-q_1)-k\right)\end{aligned}$$

with

$$\begin{aligned}q_1 &= \mathbb{P}_f\left(X_1 \leq t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c\right) \\ &= (1-\varepsilon)\Phi\left(t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c - \mu_1\right) + \varepsilon\Phi\left(t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c - \mu_2\right) \\ &= (1-\varepsilon)\Phi\left(t_{\alpha_n,k} - c + \varepsilon(\mu_2-\mu_1)\right) + \varepsilon\Phi\left(t_{\alpha_n,k} - c - (1-\varepsilon)(\mu_2-\mu_1)\right)\end{aligned}$$

since $\mathbb{E}_f[X_1] = (1-\varepsilon)\mu_1 + \varepsilon\mu_2$. Condition (13) gives that $n(1-q_1)-k > 0$ and using Markov's inequality,

$$\mathbb{P}_f\left(X_{(n-k+1)} < t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c\right) \leq \frac{n(1-q_1)}{[n(1-q_1)-k]^2}.$$

Note that the inequality $\frac{nx}{(nx-k)^2} \leq \frac{\beta}{2}$ is fulfilled if and only if $x \notin \left[\frac{k}{n} + \frac{1}{\beta n} \pm \frac{\sqrt{1+2k\beta}}{\beta n}\right]$. Then, since Condition (13) ensures us that $1-q_1 \notin \left[\frac{k}{n} + \frac{1}{n\beta} \pm \frac{\sqrt{1+2k\beta}}{n\beta}\right]$,

$$\mathbb{P}_f\left(X_{(n-k+1)} < t_{\alpha_n,k} + \mathbb{E}_f[X_1] - c\right) \leq \frac{\beta}{2}.$$

For the second term in the right-hand side of (15),

$$\mathbb{P}_f\left(X_{(k)} > \mathbb{E}_f[X_1] - c\right) \leq \mathbb{P}_f\left(\sum_{i=1}^{n} \left\{\mathbb{1}_{\{X_i > \mathbb{E}_f[X_1]-c\}} - q_2\right\} > n(1-q_2)-k\right)$$

with

$$\begin{aligned}q_2 &= \mathbb{P}_f\left(X_1 > \mathbb{E}_f[X_1] - c\right) \\ &= (1-\varepsilon)\bar{\Phi}\left(\mathbb{E}_f[X_1] - c - \mu_1\right) + \varepsilon\bar{\Phi}\left(\mathbb{E}_f[X_1] - c - \mu_2\right) \\ &= (1-\varepsilon)\bar{\Phi}\left(-c + \varepsilon(\mu_2-\mu_1)\right) + \varepsilon\bar{\Phi}\left(-c - (1-\varepsilon)(\mu_2-\mu_1)\right) \\ &= (1-\varepsilon)\Phi\left(c - \varepsilon(\mu_2-\mu_1)\right) + \varepsilon\Phi\left(c + (1-\varepsilon)(\mu_2-\mu_1)\right).\end{aligned}$$

Condition (14) gives that $n(1-q_2)-k > 0$ and using Markov's inequality,

$$\mathbb{P}_f\left(X_{(k)} > \mathbb{E}_f[X_1] - c\right) \leq \frac{n(1-q_2)}{[n(1-q_2)-k]^2}.$$

According to Condition (14), $1-q_2 \notin \left[\frac{k}{n} + \frac{1}{n\beta} \pm \frac{\sqrt{1+2k\beta}}{n\beta}\right]$, thus

$$\mathbb{P}_f\left(X_{(k)} > \mathbb{E}_f[X_1] - c\right) \leq \frac{\beta}{2}.$$

Finally, $\mathbb{P}_f(\Psi_\alpha = 0) \leq \beta$. □

6.2. Proof of Theorem 2

Let ψ_α be a level-α test. For all $f \in \mathcal{F}_{1,G}[\rho, M]$,

$$\begin{aligned}\mathbb{P}_f(\psi_\alpha = 0) &= \mathbb{P}_\phi(\psi_\alpha = 0) + \mathbb{P}_f(\psi_\alpha = 0) - \mathbb{P}_\phi(\psi_\alpha = 0) \\ &\geq 1 - \alpha - [\mathbb{P}_\phi(\psi_\alpha = 0) - \mathbb{P}_f(\psi_\alpha = 0)].\end{aligned}$$

Thus for a density $\tilde{f} \in \mathcal{F}_{1,G}[\rho, M]$ which has to be specified after,

$$\begin{aligned}\sup_{f \in \mathcal{F}_{1,G}[\rho,M]} \mathbb{P}_f(\psi_\alpha = 0) &\geq 1 - \alpha - [\mathbb{P}_\phi(\psi_\alpha = 0) - \mathbb{P}_{\tilde{f}}(\psi_\alpha = 0)] \\ &\geq 1 - \alpha - \|\mathbb{P}_\phi - \mathbb{P}_{\tilde{f}}\|_{TV}\end{aligned}$$

where $\|P - Q\|_{TV}$ denotes the total variation distance between two probability distributions P and Q. Since $\|\mathbb{P}_\phi - \mathbb{P}_{\tilde{f}}\|_{TV} \leq \sqrt{2[1 - A(\phi, \tilde{f})^n]}$ where $A(\phi, \tilde{f}) = \int_{\mathbb{R}} \sqrt{\phi(x)\tilde{f}(x)}dx$ is the Hellinger affinity between the two density functions ϕ and $\tilde{f}$,

$$\beta(\mathcal{F}_{1,G}[\rho, M]) := \inf_{\psi_\alpha} \sup_{f \in \mathcal{F}_{1,G}[\rho,M]} \mathbb{P}_f(\psi_\alpha = 0) \geq 1 - \alpha - \sqrt{2[1 - A(\phi, \tilde{f})^n]}.$$

If we specify a density $\tilde{f} \in \mathcal{F}_{1,G}[\rho, M]$ such that $A(\phi, \tilde{f}) \geq c(\alpha, \beta)^{\frac{1}{n}}$ then $\beta(\mathcal{F}_{1,G}[\rho, M]) \geq 1 - \alpha - (1 - \alpha - \beta) = \beta$. And, since

$$A(\phi, \tilde{f}) \geq 1 - \frac{1}{2}\mathbb{E}_\phi \left[\left(\frac{\tilde{f}(X) - \phi(X)}{\phi(X)} \right)^2 \right],$$

$A(\phi, \tilde{f}) \geq c(\alpha, \beta)^{\frac{1}{n}}$ is obtained if $\mathbb{E}_\phi \left[\left(\frac{\tilde{f}(X) - \phi(X)}{\phi(X)} \right)^2 \right] \leq 2 \left[1 - c(\alpha, \beta)^{\frac{1}{n}} \right]$.

In the sequel, we consider the density $\tilde{f} = (1 - \varepsilon)\phi(. - \mu_1) + \varepsilon\phi(. - \mu_2)$, with

$$(1 - \varepsilon)\mu_1 = -\varepsilon\mu_2 \tag{16}$$

$$\max(\mu_1^2, \mu_2^2, |\mu_1\mu_2|) \leq \nu^2 = \frac{M^2}{4} \tag{17}$$

$$\varepsilon(1 - \varepsilon)(\mu_2 - \mu_1)^2 = \rho \tag{18}$$

In particular, $\tilde{f} \in \mathcal{F}_{1,G}[\rho, M]$ since $(\mu_2 - \mu_1)^2 \leq M^2$.

For this density choice,

$$\begin{aligned}\mathbb{E}_\phi \left[\left(\frac{\tilde{f}(X) - \phi(X)}{\phi(X)} \right)^2 \right] &= \int_{\mathbb{R}} \frac{[\tilde{f}(x) - \phi(x)]^2}{\phi(x)} dx \\ &= \int_{\mathbb{R}} \frac{\{(1 - \varepsilon)[\phi(x - \mu_1) - \phi(x)] + \varepsilon[\phi(x - \mu_2) - \phi(x)]\}^2}{\phi(x)} dx \\ &= (1 - \varepsilon)^2 \left[\int_{\mathbb{R}} \frac{\phi(x - \mu_1)^2}{\phi(x)} dx - 1 \right] + \varepsilon^2 \left[\int_{\mathbb{R}} \frac{\phi(x - \mu_2)^2}{\phi(x)} dx - 1 \right] \\ &\quad + 2\varepsilon(1 - \varepsilon) \left[\int_{\mathbb{R}} \frac{\phi(x - \mu_1)\phi(x - \mu_2)}{\phi(x)} dx - 1 \right].\end{aligned}$$

Considering the Gaussian density ($\phi(.) = \phi_G(.)$), we have $\int_{\mathbb{R}} \frac{\phi_G(x-\mu_1)\phi_G(x-\mu_2)}{\phi_G(x)}dx = \exp(\mu_1\mu_2)$,

$$\mathbb{E}_{\phi_G}\left[\left(\frac{\tilde{f}(X)-\phi_G(X)}{\phi_G(X)}\right)^2\right] = (1-\varepsilon)^2\left[e^{\mu_1^2}-1\right]+\varepsilon^2\left[e^{\mu_2^2}-1\right]+2\varepsilon(1-\varepsilon)\left[e^{\mu_1\mu_2}-1\right].$$

Next, using that $|e^u - 1 - u - \frac{1}{2}u^2| \leq \frac{e^{U^2}}{3!}|u|^3$ for all $|u| < U$ with Condition (17),

$$\begin{aligned}\mathbb{E}_{\phi_G}\left[\left(\frac{\tilde{f}(X)-\phi_G(X)}{\phi_G(X)}\right)^2\right] &\leq (1-\varepsilon)^2\left[\mu_1^2+\frac{1}{2}\mu_1^4+\frac{e^{\nu^2}}{3!}\mu_1^6\right]\\ &\quad+\varepsilon^2\left[\mu_2^2+\frac{1}{2}\mu_2^4+\frac{e^{\nu^2}}{3!}\mu_2^6\right]\\ &\quad+2\varepsilon(1-\varepsilon)\left[\mu_1\mu_2+\frac{1}{2}\mu_1^2\mu_2^2+\frac{e^{\nu^2}}{3!}|\mu_1\mu_2|^3\right]\\ &\leq \left[(1-\varepsilon)\mu_1+\varepsilon\mu_2\right]^2+\frac{1}{2}\left[(1-\varepsilon)\mu_1^2+\varepsilon\mu_2^2\right]^2\\ &\quad+\frac{e^{\nu^2}}{3!}\left[(1-\varepsilon)|\mu_1|^3+\varepsilon|\mu_2|^3\right]^2.\end{aligned}$$

The parameters of $\tilde{f}$ are constrained such that $(1-\varepsilon)\mu_1 + \varepsilon\mu_2 = 0$ thus

$$\begin{aligned}\mathbb{E}_{\phi_G}\left[\left(\frac{\tilde{f}(X)-\phi_G(X)}{\phi_G(X)}\right)^2\right] &\leq \frac{1}{2}\left[(1-\varepsilon)\varepsilon(\mu_2-\mu_1)^2\right]^2+\frac{e^{\nu^2}}{3!}\left\{(1-\varepsilon)\varepsilon|\mu_2-\mu_1|^3[\varepsilon^2+(1-\varepsilon)^2]\right\}^2\\ &\leq (1-\varepsilon)^2\varepsilon^2(\mu_2-\mu_1)^4\left[\frac{1}{2}+\frac{e^{\nu^2}}{3!}(\mu_2-\mu_1)^2\right]\\ &\leq C^2(M)\left[(1-\varepsilon)\varepsilon(\mu_2-\mu_1)^2\right]^2 = C^2(M)\rho^2\end{aligned}$$

with $C^2(M) = \frac{1}{2}+\frac{1}{6}M^2e^{M^2/4}$. Moreover, if $u < 0$, $1-e^u \geq -u-\frac{1}{2}u^2$ thus $1-c(\alpha,\beta)^{\frac{1}{n}} \geq -\frac{1}{n}\log c(\alpha,\beta) - \frac{1}{2}\left(\frac{\log c(\alpha,\beta)}{n}\right)^2$. Then, the condition

$$\rho = (1-\varepsilon)\varepsilon(\mu_2-\mu_1)^2 \leq \frac{1}{C(M)}\sqrt{-\frac{2}{n}\log c(\alpha,\beta)-\left(\frac{\log c(\alpha,\beta)}{n}\right)^2} := \rho^\star$$

implies that $\beta(\mathcal{F}_{1,G}[\rho, M]) > \beta$.

6.3. *Proof of Proposition 1*

Following the definition of the threshold $v_{\alpha,n}$, it is easy to see that ψ_α defined in (7) is a level-α test. Now, our aim is to upper bound the term

$$\mathbb{P}_f(\psi_\alpha = 0) = \mathbb{P}_f(S_n^2 \leq v_{\alpha,n})$$

when $f \in \mathcal{F}_1[\rho, M]$.

In a first time, a control of $v_{\alpha,n}$ is required. If a real $c_{\alpha,n}$ is determined such that $\mathbb{P}_{H_0}(S_n^2 > c_{\alpha,n}) \leq \alpha$, then $v_{\alpha,n} \leq c_{\alpha,n}$. According to Wilks (1962, page 200), if $Y_1, \ldots, Y_n$ are i.i.d. random variables such that $\mathbb{E}[(Y_1 - \mathbb{E}[Y_1])^4] < +\infty$, then

$$\mathrm{Var}\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2\right) \leq \frac{1}{n}\left\{\mathbb{E}[(Y_1 - \mathbb{E}[Y_1])^4] - \frac{n-3}{n-1}\mathrm{Var}(Y_1)^2\right\}. \tag{19}$$

Hence, since $\mathbb{E}_\phi[X_1^4] < B$,

$$\mathbb{P}_{H_0}(S_n^2 > c_{\alpha,n}) = \mathbb{P}_{H_0}(S_n^2 - 1 > c_{\alpha,n} - 1) \leq \frac{\mathrm{Var}(S_n^2)}{(c_{\alpha,n}-1)^2} \leq \frac{B}{n(c_{\alpha,n}-1)^2}.$$

In particular $\mathbb{P}_{H_0}(S_n^2 > c_{\alpha,n}) \leq \alpha$ with $c_{\alpha,n} = 1 + \sqrt{\frac{B}{n\alpha}}$, and thus

$$v_{\alpha,n} \leq 1 + \sqrt{\frac{B}{n\alpha}}.$$

Please note that $\mathbb{E}_f[S_n^2] = \mathrm{Var}(X_1) = 1 + \varepsilon(1-\varepsilon)(\mu_2 - \mu_1)^2$. Hence, for all $f \in \mathcal{F}_1[\rho, M]$,

$$\begin{aligned}
\mathbb{P}_f(\psi_\alpha = 0) &\leq \mathbb{P}_f\left(S_n^2 \leq 1 + \sqrt{\frac{B}{n\alpha}}\right), \\
&= \mathbb{P}_f\left(S_n^2 - \mathbb{E}[S_n^2] \leq 1 + \sqrt{\frac{B}{n\alpha}} - \mathbb{E}[S_n^2]\right), \\
&\leq \mathbb{P}_f\left(|S_n^2 - \mathbb{E}[S_n^2]| \geq \varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2 - \sqrt{\frac{B}{n\alpha}}\right), \\
&\leq \frac{\mathrm{Var}(S_n^2)}{\left[\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2 - \sqrt{\frac{B}{n\alpha}}\right]^2}
\end{aligned}$$

if $\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2 > \sqrt{\frac{B}{n\alpha}}$. Using Equation (19), we get

$$\mathbb{P}_f(\psi_\alpha = 0) \leq \frac{\mathbb{E}_f[(X_1 - \mathbb{E}[X_1])^4]}{n\left[\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2 - \sqrt{\frac{B}{n\alpha}}\right]^2}.$$

In order to conclude, just remark that

$$\begin{aligned}
\mathbb{E}_f[(X_1-\mathbb{E}[X_1])^4] &= (1-\varepsilon)\int_{\mathbb{R}}[x-(1-\varepsilon)\mu_1-\varepsilon\mu_2]^4\phi(x-\mu_1)dx \\
&\quad +\varepsilon\int_{\mathbb{R}}[x-(1-\varepsilon)\mu_1-\varepsilon\mu_2]^4\phi(x-\mu_2)dx \\
&= (1-\varepsilon)\int_{\mathbb{R}}[y-\varepsilon(\mu_2-\mu_1)]^4\phi(y)dy \\
&\quad +\varepsilon\int_{\mathbb{R}}[y+(1-\varepsilon)(\mu_2-\mu_1)]^4\phi(y)dy \\
&= \mathbb{E}_\phi[Z^4]+6\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2\mathbb{E}_\phi[Z^2]+[\varepsilon(1-\varepsilon)^4+\varepsilon^4(1-\varepsilon)](\mu_2-\mu_1)^4 \\
&\leq B+\frac{6}{4}\sqrt{B}M^2+M^4\leq(M^2+\sqrt{B})^2.
\end{aligned}$$

Thus

$$\mathbb{P}_f(\psi_\alpha=0)\leq\frac{(M^2+\sqrt{B})^2}{n\left[\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2-\sqrt{\frac{B}{n\alpha}}\right]^2}\leq\beta$$

as soon as

$$\varepsilon(1-\varepsilon)(\mu_2-\mu_1)^2\geq\frac{C(\alpha,\beta,M,B)}{\sqrt{n}},$$

for some positive constant $C(\alpha,\beta,M,B)$. This concludes the proof of Proposition 1.

6.4. Proof of Theorem 3

Proof. Let $f(.)=(1-\varepsilon)\phi_G(.-\mu_1)+\varepsilon\phi_G(.-\mu_2)$. We will prove that if (8) holds then $f\in\bar{\mathcal{F}}_1[n,\alpha,\beta]$ and the result will be a consequence of Theorem 1. Let $j\in\mathbb{N}^\star$ such that $2^{-j}<\bar{\Phi}_G(M)\leq 2^{-(j-1)}$. In the following, we consider $k\in\mathcal{K}_n$ such that $\frac{n}{2^{j+1}}<k\leq\frac{n}{2^j}$. Note that, under the assumptions of Theorem 3, $n\geq 2/\bar{\Phi}_G(M)$ and $n/2^j\geq 1$. Note also that $\sharp\mathcal{K}_n\leq\log_2(n)$, hence $\alpha_n\geq\alpha/\sharp\mathcal{K}_n\geq\alpha/\log_2(n)$. We will show that $(\varepsilon,\mu_1,\mu_2)\in\bar{\mathcal{S}}(\alpha_n,\rho(k,n),k)$: Considering $c=\frac{t_{\alpha_n,k}}{2}$ and denoting $\tau=\mu_2-\mu_1$, we want to prove that

$$(1-\varepsilon)\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}+\varepsilon\tau\right)+\varepsilon\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}-(1-\varepsilon)\tau\right) > \rho(k,n) \tag{20}$$

$$(1-\varepsilon)\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}-\varepsilon\tau\right)+\varepsilon\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}+(1-\varepsilon)\tau\right) > \rho(k,n) \tag{21}$$

hold, with $\rho(k,n)=\frac{k}{n}+\frac{1}{n\beta}+\frac{\sqrt{1+2k\beta}}{n\beta}$.
We use a Taylor expansion at the order 2, the terms of order 1 vanish and this leads to :

$$\begin{aligned}
&(1-\varepsilon)\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}+\varepsilon\tau\right)+\varepsilon\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}-(1-\varepsilon)\tau\right) \\
&=\bar{\Phi}_G\left(\frac{t_{\alpha,k}}{2}\right)+\frac{1}{2}(1-\varepsilon)\varepsilon\tau^2\left[\varepsilon(-\phi_G'(a))+(1-\varepsilon)(-\phi_G'(b))\right]
\end{aligned}$$

where a (resp. b) belongs to the interval $\left]\frac{t_{\alpha_n,k}}{2}, \frac{t_{\alpha_n,k}}{2}+\varepsilon\tau\right[$ (resp. $\left]\frac{t_{\alpha_n,k}}{2}-(1-\varepsilon)\tau, \frac{t_{\alpha_n,k}}{2}\right[$).
We recall that $\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) = \frac{k}{n}\left[1-\sqrt{\frac{2\log(4/\alpha_n)}{k}}\right]$. Since $k \le \frac{n}{2^j}$, $\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) - \frac{k}{n} \ge -\sqrt{\frac{\log(4/\alpha_n)}{2^{j-1}n}}$.
Moreover, it is easy to check that

$$\rho(k,n) - \frac{k}{n} \le \frac{2}{n\beta} + \frac{1}{\sqrt{2^{j-1}n\beta}}.$$

Hence, in order to prove that (20) holds, we just have to show that

$$(1-\varepsilon)\varepsilon\tau^2\left\{\varepsilon[-\phi_G'(a)] + (1-\varepsilon)[-\phi_G'(b)]\right\} \ge \sqrt{\frac{\log(4/\alpha_n)}{2^{j-3}n}} + \frac{4}{n\beta} + \sqrt{\frac{1}{2^{j-3}n\beta}}. \tag{22}$$

Next, we want to prove that $\left[\frac{t_{\alpha_n,k}}{2}-(1-\varepsilon)\tau, \frac{t_{\alpha_n,k}}{2}+\varepsilon\tau\right]$ remains included in a fixed interval $[c_1(M), c_2(M)]$ with $c_1(M) > 0$. First,

$$\frac{t_{\alpha_n,k}}{2} - (1-\varepsilon)\tau \ge \frac{t_{\alpha_n,k}}{2} - \tau \ge \bar{\Phi}_G^{-1}\left(2^{-j}\right) - M := c_1(M) > 0$$

since $\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) \le \frac{k}{n} \le 2^{-j}$ and $\bar{\Phi}_G(M) > 2^{-j}$. Second,

$$\begin{aligned}
\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) &= \frac{k}{n} - \sqrt{\frac{k}{n}}\sqrt{\frac{2\log(4/\alpha_n)}{n}} \\
&> \frac{1}{2^{j+1}} - \sqrt{\frac{1}{2^{j-1}}}\sqrt{\frac{\log(4\log_2(n)/\alpha)}{n}} \\
&> \frac{1}{4}\bar{\Phi}_G(M) - \sqrt{2\bar{\Phi}_G(M)}\sqrt{\frac{\bar{\Phi}_G(M)}{36}} \\
&> \bar{\Phi}_G(M)\left[\frac{1}{4} - \sqrt{\frac{1}{18}}\right].
\end{aligned}$$

Thus $\frac{t_{\alpha_n,k}}{2} + \tau < \bar{\Phi}_G^{-1}\left(\bar{\Phi}_G(M)\left[\frac{1}{4}-\sqrt{\frac{1}{18}}\right]\right) + M := c_2(M)$
Finally the function $-\phi_G'$ is bounded from below on this interval by some positive constant $C(M) = \min_{x\in[c_1(M),c_2(M)]}(-\phi_G'(x))$. This implies that (22) is satisfied if $\varepsilon(1-\varepsilon)\tau^2 \ge C(\alpha,\beta,M)\frac{\sqrt{\log\log(n)}}{\sqrt{n}}$ for some suitable constant $C(\alpha,\beta,M)$. This concludes the proof of (20). The proof of (21) follows the same arguments. □

6.5. *Proof of Theorem 4*

Proof. We will prove that, under the assumptions of Theorem 4, $f \in \bar{\mathcal{F}}_1[n,\alpha,\beta]$ and the result will be a consequence of Theorem 1. We recall that $\sharp\mathcal{K}_n \le \log_2(n)$, hence $\alpha \ge \alpha_n \ge \alpha/\sharp\mathcal{K}_n \ge \alpha/\log_2(n)$. We set $\tau = \mu_2 - \mu_1$ and we have to prove that there exists $k \in \mathcal{K}_n$ and $c \in \mathbb{R}$ such that

$$(1-\varepsilon)\bar{\Phi}_G\left(t_{\alpha_n,k} - c + \varepsilon\tau\right) + \varepsilon\bar{\Phi}_G\left(t_{\alpha_n,k} - c - (1-\varepsilon)\tau\right) > \rho(k,n) \tag{23}$$

$$(1-\varepsilon)\bar{\Phi}_G\left(c - \varepsilon\tau\right) + \varepsilon\bar{\Phi}_G\left(c + (1-\varepsilon)\tau\right) > \rho(k,n), \tag{24}$$

with $\rho(k,n) = \frac{k}{n} + \frac{1}{n\beta} + \frac{\sqrt{1+2k\beta}}{n\beta}$. Note that $\rho(k,n) \leq \frac{k}{n} + C_\beta \frac{\sqrt{k}}{n}$ with $C_\beta = \frac{2}{\beta} + \sqrt{\frac{2}{\beta}}$. We recall that $t_{\alpha_n,k}$ is defined by

$$\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) = \frac{k}{n}\left[1 - \sqrt{\frac{2\log(4/\alpha_n)}{k}}\right].$$

In the following, we set $C_{\alpha_n} = \sqrt{2\log(4/\alpha_n)}$. Since $\alpha_n \geq \alpha/\log_2(n)$, note that $0 < C_{\alpha_n} \leq C(\alpha)\sqrt{\log\log(n)}$ for some constant $C(\alpha)$ depending only on α. We choose $k \in \mathcal{K}_n$ such that

$$\lim_{n\to+\infty} \frac{k}{\log(n)\log\log(n)} = +\infty \text{ and } \lim_{n\to+\infty} \frac{n}{k} = +\infty \tag{25}$$

and we define

$$c = \frac{t_{\alpha_n,k}}{2} - \sqrt{\frac{2}{k}} C_{\alpha_n}. \tag{26}$$

For the sake of simplicity, we omit te dependency with respect to n in the notation of k and c.

Let us first show that (24) holds for n large enough. First note that

$$(1-\varepsilon)\bar{\Phi}_G\left(c - \varepsilon\tau\right) + \varepsilon\bar{\Phi}_G\left(c + (1-\varepsilon)\tau\right) > (1-\varepsilon)\bar{\Phi}_G(c).$$

With the assumptions on k, we have that $c > 0$ for n large enough since $t_{\alpha_n,k} \to +\infty$ and $C_{\alpha_n}/\sqrt{k} \to 0$ as $n \to +\infty$. Hence

$$\bar{\Phi}_G(c) \geq \bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) + \sqrt{\frac{2}{k}} C_{\alpha_n} \phi_G\left(\frac{t_{\alpha_n,k}}{2}\right).$$

Moreover, for all $u > 0$,

$$\bar{\Phi}_G(u) \leq \frac{1}{2}\exp(-u^2/2) = \sqrt{\frac{\pi}{2}}\phi_G(u),$$

hence

$$\phi_G\left(\frac{t_{\alpha_n,k}}{2}\right) \geq \sqrt{\frac{2}{\pi}}\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right).$$

This leads to

$$(1-\varepsilon)\bar{\Phi}_G(c) > (1-\varepsilon)\left(1 + \frac{2C_{\alpha_n}}{\sqrt{\pi k}}\right)\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right).$$

After some obvious computations, Condition (24) is satisfied as soon as

$$(1-\varepsilon)C_{\alpha_n}\left(\frac{2}{\sqrt{\pi}} - 1\right)\frac{\sqrt{k}}{n} > \varepsilon\frac{k}{n} + C_\beta\frac{\sqrt{k}}{n} + \frac{2C_{\alpha_n}^2}{\sqrt{\pi}n}.$$

Since $\varepsilon < 1/\sqrt{n}$ and $k \leq n$, we have $\varepsilon k < \sqrt{k}$. We recall that $C_{\alpha_n} \to +\infty$ as $n \to +\infty$ and with the assumptions on k, we have that $\sqrt{k}/C_{\alpha_n} \to +\infty$ as $n \to +\infty$, and the above inequality holds for n large enough.

It remains to prove that (23) is satisfied with the conditions on k imposed by (25) and the value of c defined by (26). Let Δ satisfy $0 < r < \Delta \leq 1$, we choose $k \in \mathcal{K}_n$ satisfying (25) and such that

$n^{1-\Delta} \leq k \leq 2n^{1-\Delta}\log^2(n)$. Note that such values of k exist for n large enough. It follows from Lemma 2 that $t_{\alpha_n,k}/2 \leq \sqrt{2\Delta\log(n)}$. First,

$$\begin{aligned}
\bar{\Phi}_G\left(t_{\alpha_n,k} - c + \varepsilon\tau\right) &= \bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2} + \sqrt{\frac{2}{k}}C_{\alpha_n} + \varepsilon\tau\right) \\
&\geq \bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) - \left(\sqrt{\frac{2}{k}}C_{\alpha_n} + \varepsilon\tau\right)\phi_G\left(\frac{t_{\alpha_n,k}}{2}\right) \\
&\geq \frac{k}{n}\left[1 - \frac{C_{\alpha_n}}{\sqrt{k}}\right] - \left(\sqrt{\frac{2}{k}}C_{\alpha_n} + \varepsilon\tau\right)\phi_G\left(\frac{t_{\alpha_n,k}}{2}\right).
\end{aligned}$$

We have to give an upper bound for $\phi_G\left(\frac{t_{\alpha_n,k}}{2}\right)$. We use the inequality

$$\forall u > 0,\ \bar{\Phi}_G(u) \geq \left(\frac{1}{u} - \frac{1}{u^3}\right)\phi_G(u),$$

this leads to

$$\forall u > 0, \phi_G(u) \leq \frac{u^3}{u^2-1}\bar{\Phi}_G(u) \leq u^3\bar{\Phi}_G(u),$$

provided that $u^2 - 1 \geq 1$. This is the case, for n large enough for $u = t_{\alpha_n,k}/2$, hence we have

$$\begin{aligned}
\phi_G\left(\frac{t_{\alpha_n,k}}{2}\right) &\leq \left[\frac{t_{\alpha_n,k}}{2}\right]^3 \bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) \\
&\leq \left[\sqrt{2\Delta\log(n)}\right]^3 \frac{k}{n} \\
&\leq 4\sqrt{2}\left[\log(n)\right]^{7/2} n^{-\Delta}.
\end{aligned}$$

Finally, we obtain that

$$\bar{\Phi}_G\left(t_{\alpha_n,k} - c + \varepsilon\tau\right) \geq \frac{k}{n} - C_{\alpha_n}\frac{\sqrt{k}}{n} - \left(\frac{\sqrt{2}C_{\alpha_n}}{\sqrt{k}} + \varepsilon\tau\right)4\sqrt{2}\left[\log(n)\right]^{7/2} n^{-\Delta}.$$

Second, we want to lower bound $\bar{\Phi}_G\left(t_{\alpha_n,k} - c - (1-\varepsilon)\tau\right)$. We have that

$$\begin{aligned}
\bar{\Phi}_G\left(t_{\alpha_n,k} - c - (1-\varepsilon)\tau\right) &= \bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2} + \sqrt{\frac{2}{k}}C_{\alpha_n} - (1-\varepsilon)\tau\right) \\
&\geq \bar{\Phi}_G\left(\sqrt{2\Delta\log(n)} - \tau + \sqrt{\frac{2}{k}}C_{\alpha_n} + \varepsilon\tau\right) \\
&\geq \bar{\Phi}_G\left(\sqrt{2\Delta\log(n)} - \sqrt{2r\log(n)}\right) - \left(\varepsilon\tau + \frac{\sqrt{2}C_{\alpha_n}}{\sqrt{k}}\right)\phi_G(\sqrt{2\Delta\log(n)} - \sqrt{2r\log(n)})
\end{aligned}$$

since $\tau = \sqrt{2r\log(n)}$. Moreover, since $\phi_G(\sqrt{2\Delta\log(n)} - \sqrt{2r\log(n)}) = (\sqrt{2\pi})^{-1}n^{-(\sqrt{\Delta}-\sqrt{r})^2}$, and using again the inequality $\bar{\Phi}_G(u) \geq (\frac{1}{u} - \frac{1}{u^3})\phi_G(u)$ which holds for all $u > 0$, we obtain that

$$\bar{\Phi}_G\left(t_{\alpha_n,k} - c - (1-\varepsilon)\tau\right) \geq Cn^{-(\sqrt{\Delta}-\sqrt{r})^2}\left(\frac{1}{\sqrt{\log(n)}} - \varepsilon\tau - \frac{\sqrt{2}C_{\alpha_n}}{\sqrt{k}}\right),$$

for some positive constant C depending on Δ and r. Condition (23) is thus fulfilled if

$$C\varepsilon n^{-(\sqrt{\Delta}-\sqrt{r})^2}\left(\frac{1}{\sqrt{\log(n)}}-\varepsilon\tau-\sqrt{\frac{2}{k}}C_{\alpha_n}\right)>\varepsilon\frac{k}{n}+(C_{\alpha_n}+C_\beta)\frac{\sqrt{k}}{n}+\left(\sqrt{\frac{2}{k}}C_{\alpha_n}+\varepsilon\tau\right)4\sqrt{2}\left[\log(n)\right]^{7/2}n^{-\Delta}.$$

By (25), $C_{\alpha_n}/\sqrt{k}=o(1/\sqrt{\log(n)})$, and the left hand side of this inequality is equivalent as $n\to+\infty$ to $C\varepsilon n^{-(\sqrt{\Delta}-\sqrt{r})^2}/\sqrt{\log(n)}$ and the right hand side is equivalent as $n\to+\infty$ to $8C_{\alpha_n}\left(\log(n)\right)^{7/2}n^{-\Delta}/\sqrt{k}$. Hence, the condition (23) will be satisfied asymptotically if for some $\Delta\in]0,1]$,

$$\delta+(\sqrt{\Delta}-\sqrt{r})^2<\frac{1+\Delta}{2}.$$

- If $\frac{1}{2}<\delta\leq\frac{3}{4}$ and $0<r\leq\frac{1}{4}$, we set $\Delta=4r$ and the above condition becomes $r>\delta-\frac{1}{2}$.
- If $\frac{1}{2}<\delta\leq\frac{3}{4}$ and $r>\frac{1}{4}$, the above condition is satisfied with $\Delta=1$ and no additional condition is required.
- If $\delta>\frac{3}{4}$, we set $\Delta=1$ and the above condition becomes $r>(1-\sqrt{1-\delta})^2$.

This concludes the proof of Theorem 4.

□

Appendix A: Lemmas for upper-bound

Lemma 1. *Let $Y_1,\ldots,Y_n$ be n random variables with a cumulative distribution function F and the order statistics are denoted $Y_{(1)}\leq Y_{(2)}\leq\ldots,Y_{(n)}$. Let $\alpha\in]0,1[$ and let $k\in\{0,\ldots,n-1\}$ such that $k>2\log\left(\frac{2}{\alpha}\right)$. Let c and d be two reals such that*

$$F(d)\vee(1-F(c))\leq\frac{k}{n}\left[1-\sqrt{\frac{2\log(\frac{2}{\alpha})}{k}}\right]. \tag{27}$$

Then $\mathbb{P}(Y_{(n-k+1)}\geq c)\leq\alpha$ and $\mathbb{P}(Y_{(k)}\leq d)\leq\alpha$.

Proof.

$$\begin{aligned}\mathbb{P}(Y_{(n-k+1)}\geq c) &= \mathbb{P}\left(\sum_{i=1}^n\mathbb{1}_{\{Y_i\geq c\}}\geq k\right)\\ &= \mathbb{P}\left(\sum_{i=1}^n\{\mathbb{1}_{\{Y_i\geq c\}}-[1-F(c)]\}\geq k-n[1-F(c)]\right).\end{aligned}$$

According to Condition (27),

$$k-n[1-F(c)]\geq k\sqrt{\frac{2\log(\frac{2}{\alpha})}{k}}>0.$$

Using a Bernstein's inequality, we get

$$\mathbb{P}(Y_{(n-k+1)}\geq c)\leq 2\exp\left[-\frac{1}{2}\frac{(k-n[1-F(c)])^2}{v+\frac{1}{3}(k-n[1-F(c)])}\right]$$

with $v = \sum_{i=1}^n \mathbb{E}[(\mathbb{1}_{\{Y_i \geq c\}} - [1-F(c)])^2] = \sum_{i=1}^n \mathrm{Var}(\mathbb{1}_{Y_i \geq c}) = nF(c)[1-F(c)] \leq n[1-F(c)]$. Thus, $3v + k - n[1-F(c)] \leq 2n[1-F(c)] + k \leq 3k - 2k\sqrt{\frac{2\log(\frac{2}{\alpha})}{k}} \leq 3k$. This implies that

$$\mathbb{P}(Y_{(n-k+1)} \geq c) \leq 2\exp\left[-\frac{3}{2}\frac{(k-n[1-F(c)])^2}{3k}\right] \leq 2\exp\left[-\log\left(\frac{2}{\alpha}\right)\right] = \alpha.$$

In the same way,

$$\begin{aligned}\mathbb{P}(Y_{(k)} \leq d) &= \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{\{Y_i \geq d\}} \leq n-k\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n \{\mathbb{1}_{\{Y_i \geq d\}} - [1-F(d)]\} \leq nF(d) - k\right).\end{aligned}$$

Since $nF(d) - k < 0$ according to Condition (27), a Bernstein's inequality implies that

$$\mathbb{P}(Y_{(k)} \leq d) \leq \mathbb{P}\left(\left|\sum_{i=1}^n \{\mathbb{1}_{\{Y_i \geq d\}} - [1-F(d)]\}\right| \geq k - nF(d)\right) \leq 2\exp\left[-\frac{1}{2}\frac{[nF(d)-k]^2}{v + \frac{1}{3}[k-nF(d)]}\right]$$

with $v = \sum_{i=1}^n \mathbb{E}[(\mathbb{1}_{\{Y_i \geq d\}} - [1-F(d)])^2] = \sum_{i=1}^n \mathrm{Var}(Y_i \geq d) = nF(d)[1-F(d)] \leq nF(d)$. Thus, $3v + k - nF(d) \leq 2nF(d) + k \leq 3k - 2k\sqrt{\frac{2\log(\frac{2}{\alpha})}{k}} \leq 3k$. This implies that

$$\mathbb{P}(Y_{(k)} \leq d) \leq 2\exp\left[-\frac{3}{2}\frac{[nF(d)-k]^2}{3k}\right] \leq 2\exp\left[-\log\left(\frac{2}{\alpha}\right)\right] = \alpha.$$

□

Lemma 2. *If $k \geq 8\log(4/\alpha_n)$ and $\frac{k}{n} \geq n^{-\Delta}$ with $\Delta \in]0,1[$, then*

$$t_{\alpha_n,k} \leq 2\sqrt{2\Delta\log(n)}.$$

Proof.

$$\begin{aligned}\bar{\Phi}_G\left(\frac{t_{\alpha_n,k}}{2}\right) &= \frac{k}{n}\left[1 - \sqrt{\frac{2\log(4/\alpha_n)}{k}}\right] \\ &\leq \frac{1}{2}\exp\left[-\frac{1}{2}\left(\frac{t_{\alpha_n,k}}{2}\right)^2\right],\end{aligned}$$

thus

$$\exp\left[\frac{1}{2}\left(\frac{t_{\alpha_n,k}}{2}\right)^2\right] \leq \frac{1}{2}\left[1 - \sqrt{\frac{2\log(4/\alpha_n)}{k}}\right]^{-1} n^{\Delta}.$$

If $k \geq 8\log(4/\alpha_n)$, then

$$2\left[1 - \sqrt{\frac{2\log(4/\alpha_n)}{k}}\right] \geq 1$$

which leads to $t_{\alpha_n,k} \leq 2\sqrt{2\Delta\log(n)}$. □

References

Addario-Berry, L., Broutin, N., Devroye, L., and Lugosi, G. (2010). On combinatorial testing problems. *Ann. Statist.*, 38(5):3063–3092.

Azaïs, J.-M., Gassiat, É., and Mercadier, C. (2009). The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM Probab. Stat.*, 13:301–327.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):629–662.

Cai, T. T., Jin, J., and Low, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449.

Cai, T. T. and Wu, Y. (2012). Optimal detection for sparse mixtures. *ArXiv:1211.2265v1*.

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the L^2 distance between competing models. *J. Amer. Statist. Assoc.*, 99(466):488–498.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):pp. 19–29.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *Ann. Statist.*, 37(5A):2523–2542.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J. Statist. Plann. Inference*, 43(1-2):19–40.

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(4):1178–1209.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994.

Fromont, M. and Laurent, B. (2006). Adaptive goodness-of-fit tests in a density model. *Ann. of Stat.*, 34:1–45.

Garel, B. (2007). Recent asymptotic results in testing for mixtures. *Comput. Statist. Data Anal.*, 51(11):5295–5304.

Ingster, Y. (1999). Minimax detection of a signal for l^n-balls. *Mathematical Methods of Statistics*, 7(4):401–428.

Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.*, 35(5):2018–2053.

Klar, B. and Meintanis, S. G. (2005). Tests for normal mixtures based on the empirical characteristic function. *Comput. Statist. Data Anal.*, 49(1):227–242.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley series in Probability and Statistics.

Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24:2477–2498.

Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York.