Statistics & Risk Modeling 28, 389–410 (2011) / **DOI** 10.1524/strm.2011.1065
© Oldenbourg Wissenschaftsverlag, München 2011

Test on components of mixture densities

Florent Autin and Christophe Pouet

Received: April 14, 2011; Accepted: July 6, 2011

Summary: This paper deals with statistical tests on the components of mixture densities. We propose to test whether the densities of two independent samples of independent random variables $Y_1, \dots, Y_n$ and $Z_1, \dots, Z_n$ result from the same mixture of M components or not. We provide a test procedure which is proved to be asymptotically optimal according to the minimax setting. We extensively discuss the connection between the mixing weights and the performance of the testing procedure; this link had never been clearly established up to now.

1 Introduction

Since more than 20 years, the mixture model has gained a lot of attention. This is due to its ease of interpretation by viewing each component as a distinct group in the data. This model has been widely applied in several areas such as finance, economy, biology, astronomy, survey methods, etc.

Most of the theoretical results in the literature deal with the estimation of the components or of the mixing weights. There are two types of mixture models: the most popular one has fixed mixing weights and the other one has varying mixing weights.

On the one hand, many statisticians have been interested in estimating the mixing weights. For example, Hall [13], Titterington [26] and Hall and Titterington [14] have considered nonparametric estimation of the mixing weights. Two other examples about the mixing weights are the estimation of a functional of the weights by van de Geer [27] and the computation of confidence intervals by Qin [24]. On the other hand, one can think of the estimation of the mixture components. This can easily be done with varying mixture weights which were first introduced by Maiboroda [20] as far as we know. Both estimation and testing problems have been considered in this set-up. Several well-known methods have been successfully applied such as histograms in Lodakto and Maiboroda [18] and empirical distribution in Maiboroda [20, 21]. Nevertheless for the minimax approach, only the estimation problem has been considered by Pokhyl'ko [23]. He proved the optimality of wavelet thresholding methods for the estimation of components in varying mixing weights model.

AMS 2010 subject classification: Primary: 62C20, 62G10, 62G20; Secondary: 30H25, 42C40
Key words and phrases: Besov spaces, minimax theory, mixture model, nonparametric tests, wavelet decomposition

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

The mixing weights and the mixture components can also be both estimated both at the same time, this result holds in a particular setting for k-variate data introduced by Hall and Zhou [15].

More recently, the mixture model has also been studied in the testing problem framework. The usual addressed question is whether the observations come from a non-trivial mixture model or from a trivial one (i.e. with only one component). This has been done for example by Garel [10, 11] and Delmas [8] in the case of fixed mixing weights and by Maiboroda [22] in the case of varying mixing weights. Their homogeneity tests which rely respectively on the likelihood ratio test and on a Kolmogorov–Smirnov type test are proved to be consistent. All these authors only considered the mixture problem with one sample and the behaviour of the test under a sequence of simple alternatives.

Here we propose to study a testing problem with two independent samples in a mixture model with varying mixing weights in the minimax setting. Let $Y_1, \ldots, Y_n$ be a sample of independent random variables with unknown marginal densities

$$f_i(.) = \sum_{u=1}^{M} \omega_u(i)\, p_u(.), \quad 1 \le i \le n, \tag{1.1}$$

and let $Z_1, \ldots, Z_n$ be another sample of independent random variables with unknown marginal densities

$$g_i(.) = \sum_{u=1}^{M} \sigma_u(i) q_u(.), \quad 1 \le i \le n. \tag{1.2}$$

In the sequel, the mixing weights $(\omega_u(i), 1 \le u \le M, 1 \le i \le n)$ and $(\sigma_u(i),\ 1 \le u \le M, 1 \le i \le n)$ are supposed to satisfy

$$\forall (u,i) \in \{1,\ldots,M\} \times \{1,\ldots,n\},\ \min(\omega_u(i), \sigma_u(i)) \ge 0, \tag{1.3}$$

$$\forall i \in \{1,\ldots,n\},\ \sum_{u=1}^{M} \omega_u(i) = \sum_{u=1}^{M} \sigma_u(i) = 1, \tag{1.4}$$

and to be entirely known by the statistician whereas the densities p_u and q_u $(1 \le u \le M)$ are unknown.

We propose to study whether these two samples of random variables come from the same mixture of M unknown densities or not, that is to say to test the null hypothesis $\mathcal{H}_0 :\ p_u = q_u$ for any u belonging to $\{1,\ldots,M\}$ against the alternative $\mathcal{H}_1 :\ p_u \ne q_u$ for at least one u belonging to $\{1,\ldots,M\}$. A more detailed description of this testing problem is given in Section 2.2.

In Butucea and Tribouley [4] some procedures are proposed to test if two n-samples of i.i.d. variables have common probability density. Their setting is equivalent to the case $M = 1$ in our mixture problem. Here the problem appears more complex since the two samples are not based on random variables with the same marginal densities. In Section 3, our results show that there is no loss in the minimax rate compared to the simpler case studied by Butucea and Tribouley [4]. We also provide an asymptotically minimax test which is based on wavelet methods and we prove the dependence between the mixing

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

weights and the constants appearing in the definition of the minimax rate of testing. Until now this phenomenon has never been studied and is extensively discussed in this paper. Sections 4 and 5 are respectively devoted to open questions and to the proofs of our results. Finally many technical lemmas that are necessary to prove the main results are postponed in Appendix.

The varying mixing weights model is quite new and really deserves attention as it can be useful in several fields such as medicine and social science. Let us describe more precisely what can be done in social science in order to help the reader to recognize this usefulness.

Social science

Social science is a domain that can potentially take great advantage of the varying mixing weights model studied here. Let us consider an organization divided into several departments such as an enterprise. Aggregated information are only known at the department level, e.g. proportion of men and women, proportion of graduates and undergraduates, proportion of married and unmarried people, etc. The researcher is interested in a variable for these subgroups such as salary. For each person, the researcher has only recorded salary and department. The information of interest which allows to divide the sample into subgroups is unavailable at the individual level. This can happen if the researcher has forgotten to record this information when collecting the data; this also frequently happens when a new question arises during the study of the data. Another reason can be that the law forbids to record such information at the individual level. From our point of view one of the best example are National Statistics Offices which are usually allowed to supply statistics only at aggregated level. The unit can be for example an area consisting of hundreds or thousands households. The usual manner is to collect all the information needed for the study through a survey. It will lead to a lengthy questionnaire as the researcher does not exactly know in advance what is the relevant information. Therefore the mixing varying weigths approach can alleviate these drawbacks. This technique will help to concentrate on the core of the topic. The survey questionnaire will be shorter and all the questions will be geared toward collecting precise information about the subject itself. During the study, the researcher will use any available aggregated information. Any explanatory variable can be chosen at this level (e.g. income, academic level, number of cars, gender, number of children). Once a variable has been chosen, it has to be divided into M classes. These are the subpopulations in our problem. For each observation the researcher knows the statistical unit it belongs to and therefore the theoretical distribution of the explanatory variable associated with. Each line in the matrix Ω corresponds to the distribution of the explanatory variable for one observation in a specified statistical unit. Here we want to stress that this distribution is exactly known; indeed aggregated information given by National Statistics Offices are often very reliable as it is usually mandatory to answer national census. Moreover a wide range of explanatory variables is available.

Our setting can also be related to the problem of missing data. There is a wealth of works on partially missing data (see e.g. McKnight et al. [19]) but the case of entirely missing data has never been really considered. From our point of view, a varying mixing weights model is a way to cope with this lack of information at the individual level and

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

to allow the researcher to reconstruct information for each subgroup. Although we are aware of methodological problems, we want to emphasize that in this case the varying mixing weights are exactly known to the researcher; indeed, aggregated information often exists and is much easier to collect than individual information. A real-life application and numerous simulations are availaible in an extended version [2].

2 Model and background

2.1 Wavelet framework

Wavelets have often been applied in different mathematical fields such as approximation theory, signal analysis and statistics. In particular, many recent statistical works on estimation (see e.g. Autin [1], Donoho et al. [9] or Cohen et al. [6]) and on hypothesis testing (see e.g. Spokoiny [25]) use the wavelet setting to provide efficient estimators and tests. There are many explanations for the huge interest of the wavelet setting. One of them is that wavelets bases are localized both in frequency and in time, contrary to the classical Fourier basis which is only localized in frequency. As a consequence, the wavelet setting appears to be well adapted to describe local characteristics of a signal to be reconstructed.

Let ϕ and ψ be two compactly supported functions of $L_2(\mathbb{R})$ and denote for all j in $\mathbb{N}$ and all k in $\mathbb{Z}$ and all x in $\mathbb{R}$, $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$.

Suppose that for any j in $\mathbb{N}$:

- $\{\phi_{jk}, \psi_{j'k};\ j' \geq j; k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$,
- $\text{support}(\phi) \cup \text{support}(\psi) \subseteq [-L, L[$ for some $L > 0$.

Popular examples of such bases, called compactly supported orthonormal wavelet bases, are given in Daubechies [7]. The function ϕ is called the scaling function and ψ the associated wavelet.

Any function h in $L_2(\mathbb{R})$ can be represented as:

$$h(t) = \sum_{k\in\mathbb{Z}} \alpha_{jk}\phi_{jk}(t) + \sum_{j'\geq j}\sum_{k\in\mathbb{Z}} \beta_{j'k}\psi_{j'k}(t)$$

where $\forall j \in \mathbb{N}, \forall j' \geq j, \forall k \in \mathbb{Z}$:

- $\alpha_{jk} = \displaystyle\int_{I_{jk}} h(t)\phi_{jk}(t)dt \quad \text{and} \quad \beta_{j'k} = \int_{I_{j'k}} h(t)\psi_{j'k}(t)dt,$
- $I_{jk} = \left\{x \in \mathbb{R};\ -L \leq 2^j x - k < L\right\} = \left[\frac{k-L}{2^j}, \frac{k+L}{2^j}\right[.$

The wavelet framework is a good candidate as we are interested in Besov spaces. It has already been successfully applied by Pokhyl'ko [23] for the varying mixing weights model in the estimation problem.

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

2.2 Mathematical description of the testing problem

We briefly recall the setting introduced in Section 1. Let $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_n$ be two independent samples of independent random variables with unknown marginal densities respectively given by equations (1.1) and (1.2) and mixing weights $(\omega_u(i),\ 1 \leq u \leq M, 1 \leq i \leq n)$ and $(\sigma_u(i),\ 1 \leq u \leq M, 1 \leq i \leq n)$ satisfying (1.3) and (1.4). Here and what follows let $\overrightarrow{p} = (p_1, \ldots, p_M)$ and $\overrightarrow{q} = (q_1, \ldots, q_M)$.

Let $\mathcal{D}$ be the set of all probability densities with respect to the Lebesgue measure on $\mathbb{R}$. For any real number $R > 0$, we define

$$\Theta_0(R) = \left\{ \left(\overrightarrow{p}, \overrightarrow{q}\right) : \forall u \in \{1, \ldots, M\}, \quad p_u = q_u \in \mathcal{S}(R) \right\}$$

with $\mathcal{S}(R) = \mathcal{D} \cap \mathbb{L}_\infty(R) \cap \mathbb{L}_2(R)$. We consider the following null hypothesis

$$\mathcal{H}_0 : \quad \left(\overrightarrow{p}, \overrightarrow{q}\right) \in \Theta_0(R).$$

For a given $C > 0$, we define

$$\Theta_1(R, C, r_n, s) = \Big\{ \left(\overrightarrow{p}, \overrightarrow{q}\right) : \forall u \in \{1, \ldots, M\},\ p_u - q_u \in \mathcal{B}^s_{2,\infty}(R), \\ \exists u \in \{1, \ldots, M\},\ (p_u, q_u) \in \Lambda(R, C, r_n) \Big\},$$

where $\Lambda(R, C, r_n) = \left\{(p, q) \in (\mathcal{D} \cap \mathbb{L}_\infty(R))^2,\ \|p - q\|_2 \geq C r_n\right\}$, for a sequence r_n tending to 0 when n goes to infinity and $\mathcal{B}^s_{2,\infty}(R)$ is the R-ball of a functional space defined below. We consider the following alternative

$$\mathcal{H}_1 : \quad \left(\overrightarrow{p}, \overrightarrow{q}\right) \in \Theta_1(R, C, r_n, s).$$

As usual in the nonparametric setting, we focus on a large class of functions having some regularity so as to derive optimal properties. For the chosen wavelet basis, the space $\mathcal{B}^s_{2,\infty}(R)$ represents the R-ball of the so-called Besov space which consists of all functions $h \in L_2(\mathbb{R})$ whose wavelet coefficients $(\alpha_{jk},\ \beta_{j'k},\ j \in \mathbb{N},\ j' \geq j, k \in \mathbb{Z})$ satisfy:

$$\sup_{j \in \mathbb{N}} 2^{2js} \sum_{j' \geq j} \sum_{k \in \mathbb{Z}} \beta^2_{j'k} \leq R.$$

The minimax setting

In this paragraph we recall the minimax approach which is often used to evaluate the performances of testing procedures. Given the sum of the probability errors, say $\gamma \in [0,\ 1]$, we study the optimal separation rate r_n between the null hypothesis and the alternative. This rate r_n is the best possible rate separating at least one of the M pairs of density components p_u and q_u. It is usually called *the minimax rate*. Let us recall the classical definition for this rate.

Definition 2.1 *Let $0 < \gamma < 1$. We say that r_n is the minimax rate separating $\mathcal{H}_0$ and $\mathcal{H}_1$ of our testing problem at level γ if the two following statements are satisfied:*

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

1. *there exists a sequence of test procedures* Δ_n^* *and a constant* C_γ *such that for all* $C > C_\gamma$*:*

$$\limsup_{n\to\infty}\left(\sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\\ \in\Theta_0(R)}}\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta_n^*=1)+\sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\in\\ \Theta_1(R,C,r_n,s)}}\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta_n^*=0)\right)\le\gamma;$$

2. *there exists a constant* c_γ *such that for all* $C < c_\gamma$*:*

$$\liminf_{n\to\infty}\inf_{\Delta}\left(\sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\\ \in\Theta_0(R)}}\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta=1)+\sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\in\\ \Theta_1(R,C,r_n,s)}}\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta=0)\right)>\gamma,$$

where the infimum is taken over all test procedures Δ*.*

Assumption 2.2 *Let us denote by* $\Omega=(\Omega_{ui})$ *the* $(M\times n)$*-matrix with coefficients* $\Omega_{ui}=\omega_u(i)$ *and* $\Sigma=(\Sigma_{ui})$ *the* $(M\times n)$*-matrix with coefficients* $\Sigma_{ui}=\sigma_u(i)$*. We assume that the smallest eigenvalues of the* $(M\times M)$*-matrices* $\Gamma_n=\frac{\Omega\Omega^*}{n}$ *and* $\Gamma'_n=\frac{\Sigma\Sigma^*}{n}$ *are both larger than or equal to* K*, with* $0<K<1$*.*

We recall the following proposition due to Maiboroda [21].

Proposition 2.3 *Suppose that Assumption 2.2 is satisfied by the mixing weights* $(\omega_u(i),\ 1\le u\le M, 1\le i\le n)$ *and* $(\sigma_u(i), 1\le u\le M, 1\le i\le n)$ *associated with the model. Then, there exists a solution of the two problems*

$$\left[\text{find } a_l=(a_l(i))_{1\le i\le n} \text{ such that } \langle\omega_k,a_l\rangle_n:=\frac{1}{n}\sum_{i=1}^n\omega_k(i)a_l(i)=\delta_{kl}\right],$$

$$\left[\text{find } b_l=(b_l(i))_{1\le i\le n} \text{ such that } \langle\sigma_k,b_l\rangle_n:=\frac{1}{n}\sum_{i=1}^n\sigma_k(i)b_l(i)=\delta_{kl}\right],$$

where δ_{kl} *is the Kronecker delta. The solutions of interest can be viewed as the components of* $(n\times M)$*-matrices* $A=(a_l(i))_{i,l}$ *and* $B=(b_l(i))_{i,l}$ *such that:*

$$\Omega A=\Sigma B=n\ I_M,$$

where I_M *is the* $(M\times M)$*-identity matrix.*

According to Lemma 6.1 in Autin and Pouet [2], solutions satisfy

$$\text{Trace }(AA^*):=\sum_{l=1}^M\langle a_l,a_l\rangle_n=\frac{1}{n}\sum_{l=1}^M\sum_{i=1}^n a_l^2(i)\le\frac{M}{K},\tag{2.1}$$

$$\text{Trace }(BB^*):=\sum_{l=1}^M\langle b_l,b_l\rangle_n=\frac{1}{n}\sum_{l=1}^M\sum_{i=1}^n b_l^2(i)\le\frac{M}{K}.\tag{2.2}$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Remark 2.4 Note that Assumption 2.2 ensures that matrices Γ_n and Γ'_n are both invertible. As we shall see in Section 3, after *inverting* the mixing weights operators Ω and Σ, that is to say finding the components of matrices A and B, we will be able to provide a test statistic adapted to our testing problem.

Analogously to Pokhyl'ko [23], the Euclidian norms of matrices AA^* and BB^* will appear in the separation constants C_γ and c_γ. Therefore, a control on these quantities is required. A natural way to proceed is to impose a condition on the mixing weights, that is to say to control the behaviour of the smallest eigenvalues of matrices Γ_n and Γ'_n. Such kind of control is achieved through Assumption 2.2.

As it will be shown in Section 3, the performance of the test statistic will depend on the values of the smallest eigenvalues of Γ_n and Γ'_n.

3 Nonparametric test procedure

This paragraph deals with the case where the regularity s of the Besov space that appears in $\mathcal{H}_1$ is known. From now on we denote by a_l and b_l the n-vectors which are the solutions of the two optimization problems appearing in Proposition 2.3. Let us describe the asymptotically minimax decision rule.

3.1 Definition of the test procedure

For each level parameter j, we define the test procedure Δ_j comparing the test statistic

$$T_j = \frac{1}{n^2} \sum_{i_1 \neq i_2 = 1}^{n} \sum_{k \in \mathbb{Z}} \sum_{l=1}^{M} \big[a_l(i_1)\phi_{jk}(Y_{i_1}) - b_l(i_1)\phi_{jk}(Z_{i_1})\big] \cdot \big[a_l(i_2)\phi_{jk}(Y_{i_2}) - b_l(i_2)\phi_{jk}(Z_{i_2})\big]$$

with a threshold value $t_n = t\, r_n^2$ where t is a constant chosen later. We define

$$\Delta_j = \begin{cases} 1 & \text{if } T_j > t_n, \\ 0 & \text{if } T_j \leq t_n. \end{cases}$$

before studying the properties of T_j, we give some arguments to explain this choice of test statistic.

For fixed l the mixture components p_l and q_l can be decomposed in the wavelet basis as follows:

$$p_l(t) = \sum_{k \in \mathbb{Z}} \alpha_{jk}^{(l,p)} \phi_{jk}(t) + \sum_{j' \geq j} \sum_{k \in \mathbb{Z}} \beta_{j'k}^{(l,p)} \psi_{j'k}(t),$$

$$q_l(t) = \sum_{k \in \mathbb{Z}} \alpha_{jk}^{(l,q)} \phi_{jk}(t) + \sum_{j' \geq j} \sum_{k \in \mathbb{Z}} \beta_{j'k}^{(l,q)} \psi_{j'k}(t).$$

The quantities

$$\hat{\alpha}_{jk}^{(l,p)} = \frac{1}{n} \sum_{i=1}^{n} a_l(i)\phi_{jk}(Y_i) \quad \text{and} \quad \hat{\alpha}_{jk}^{(l,q)} = \frac{1}{n} \sum_{i=1}^{n} b_l(i)\phi_{jk}(Z_i) \tag{3.1}$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

are respectively the estimators of $\alpha_{jk}^{(l,p)}$ and $\alpha_{jk}^{(l,q)}$ obtained via the method of moments. Pokhyl'ko [23] has already used these estimators to construct thresholding estimators of the components in the same model.

Instead of looking at $\|p_l - q_l\|_2^2$, we look at $\sum_{k\in\mathbb{Z}} \left(\alpha_{jk}^{(l,p)} - \alpha_{jk}^{(l,q)}\right)^2$ from its empirical value:

$$\sum_{k\in\mathbb{Z}} \left(\hat{\alpha}_{jk}^{(l,p)} - \hat{\alpha}_{jk}^{(l,q)}\right)^2 = T_j + n^{-2} \sum_{l=1}^{M} \sum_{k\in\mathbb{Z}} \sum_{i=1}^{n} \left[a_l(i)\phi_{jk}(Y_i) - b_l(i)\phi_{jk}(Z_i)\right]^2.$$

Keeping in mind this decomposition, the choice of T_j as test statistic appears natural, as soon as the added term is negligible. Note that our problem can be viewed as an inverse problem which requires inversion an operator in finite dimensions. Indeed, the empirical observations of the wavelet coefficients of the mixture components p_l and q_l are not directly available. Therefore it is required to *invert*, in some sense, the mixing weights operators Ω and Σ so as to construct the estimators of these wavelet coefficients defined in (3.1). This inverse problem is potentially ill-conditioned if relevant eigenvalues are small.

3.2 Properties of the test statistic

In this section, we provide two propositions which will be crucial when evaluating the performances of our test procedure. They deal with the behaviours of its expectation and its variance.

Proposition 3.1 *Let j be any given level parameter. Then,*

$$\mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_j) = \sum_{l=1}^{M} \sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} (p_l - q_l)\phi_{jk}\right)^2 - \frac{1}{n^2} \sum_{l=1}^{M} \sum_{k\in\mathbb{Z}} \sum_{i=1}^{n} \left(\int_{\mathbb{R}} (a_l(i) f_i - b_l(i) g_i)\, \phi_{jk}\right)^2.$$

Remark 3.2 In the particular case where $\Omega = \Sigma$, the test statistic T_j is centered under the null hypothesis.

Corollary 3.3 *For any $j \in \mathbb{N}$,*

$$\left| \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_j) - \sum_{l=1}^{M} \sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} (p_l - q_l)\phi_{jk}\right)^2 \right| \leq \frac{8LMR^2}{Kn}.$$

Proposition 3.4 *There exists a constant $C_T = C_T(R, L, \|\phi\|_\infty) > 0$ such that*

$$\mathbb{V}\mathrm{ar}_{\overrightarrow{p},\overrightarrow{q}}(T_j) \leq \frac{C_T M^2}{K^2} \left(\frac{2^j}{n^2} + \frac{1}{n} \sum_{l=1}^{M} \|p_l - q_l\|_2^2 + \sqrt{\frac{2^j}{n^3}} \sum_{l=1}^{M} l \|p_l - q_l\|_2 \right).$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Remark 3.5 Under the null hypothesis the variance of the test statistic T_j is less than or equal to $C_T M^2 K^{-2} 2^j n^{-2}$.

3.3 Minimax performance of the test procedure

For any $s > 0$, let $(r_n)_{n\in\mathbb{N}}$ be the sequence such that

$$r_n = n^{-\frac{2s}{1+4s}} \quad \forall n \in \mathbb{N}^*.$$

Theorem 3.6 shows that the test procedure defined in Section 3 provides an accurate upper bound when it is correctly calibrated.

Theorem 3.6 (Upper bound) *Fix $\gamma \in]0, 1[$ and consider the test procedure $\Delta_s^* = \Delta_{j_n}$ where j_n is the smallest integer such that $2^{-j_n} \leq n^{-\frac{2}{1+4s}}$. Let t and C_γ be two positive real numbers defined as follows:*

$$t = \left(2\sqrt{\frac{C_T}{\gamma}} + 8LR^2\right)\frac{M}{K}, \qquad C_\gamma^2 = 2\left(\frac{1}{K}\sqrt{\frac{6\,C_T}{\gamma}} + R + \frac{t}{M}\right).$$

Then for all $C > C_\gamma$

$$\limsup_{n\to\infty}\left(\sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\\ \in\Theta_0(R)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta_s^* = 1) + \sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\in\\ \Theta_1(R,C,r_n,s)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta_s^* = 0)\right) \leq \gamma.$$

Although the exact value of the constant C_T is very complicated, it can be exactly calculated by following the proofs.

Now, let us focus on the lower bound associated with our nonparametric testing problem $\mathcal{H}_0$ versus $\mathcal{H}_1$. We aim at providing a constant c_γ such that we ensure that no test procedure is able to choose $\mathcal{H}_0$ or $\mathcal{H}_1$ with a sum of the probability errors less than γ $(0 < \gamma < 1)$. Obviously, the smaller the distance between c_γ and C_γ the more accurate our results. Next theorem proves that our test procedure is asymptotically minimax.

Similarly to the classical methods for providing lower bounds (see e.g. Gayraud and Pouet [12], Butucea and Tribouley [4]) we shall consider a subspace of $\Theta_1(R, C, r_n, s)$ defined for any $C_1 > 0$ as follows:

$$\begin{aligned}\tilde{\Theta}_1(R, C, C_1, r_n, s) = \Big\{ &\left(\overrightarrow{p}, \overrightarrow{q}\right) : \forall u \in \{1, \ldots, M\}, p_u - q_u \in \mathcal{B}^s_{2,\infty}(R),\\ &\forall u \in \{1, \ldots, M\}, \{x;\, p_u(x) \wedge q_u(x) \geq C_1\} \supseteq [0, 1]\\ &\exists u \in \{1, \ldots, M\}, (p_u, q_u) \in \Lambda(R, C, r_n)\Big\}.\end{aligned}$$

Theorem 3.7 (Lower bound) *Let $0 < \gamma < 1$, $s > 0$ and let $c_\gamma > 0$ satisfy the following equation*

$$c_\gamma^4 = \left(\frac{C_1^2}{L\,K^2}\ln[4(1-\gamma)^2 + 1] \wedge 2R^2\right)\frac{2^{-4s}}{4M^2}.$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Then for all $C < c_\gamma$

$$\liminf_{n\to\infty} \inf_{\Delta} \left(\sup_{\substack{(\overrightarrow{p},\overrightarrow{q}) \\ \in \Theta_0(R)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta = 1) + \sup_{\substack{(\overrightarrow{p},\overrightarrow{q}) \in \\ \Theta_1(R,C,r_n,s)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta = 0) \right) > \gamma$$

where the infimum is taken over all test procedures Δ.

From Theorems 3.6 and 3.7 we deduce the minimax rate of testing. It is the same as the one found by Butucea and Tribouley [4] when there is only one subgroup. Advances in our results are the extension to the varying mixing weights model which allows non-identically distributed random variables compared to Butucea and Tribouley [4] and the role played by the mixing weights which is clearly exposed.

Corollary 3.8 *For any* $s > 0$*, the test procedure* Δ_s^* *is asymptotically minimax and the minimax rate separating* $\mathcal{H}_0$ *and* $\mathcal{H}_1$ *is* $r_n = n^{-\frac{2s}{1+4s}}$.

3.4 Discussion about the constants c_γ and C_γ

Theorems 3.6 and 3.7 exhibit two constants C_γ and c_γ appearing respectively in the upper bound and the lower bound. We think that the connection between these constants and the parameters M and K is a novelty and deserves a discussion. Indeed, we keep in mind that

- C_γ is the minimal value for C such that our test statistic is able to detect if all the mixture components are identical in the two populations with sum of the probability errors not exceeding γ;
- c_γ is the maximal value for C such that no test statistic is able to detect if all the mixture components are identical in the two populations with sum of probability errors not exceeding γ.

As a consequence we proved that our test statistic is optimal in the minimax sense since it attains the minimax rate of convergence separating $\mathcal{H}_0$ and $\mathcal{H}_1$.

According to the definitions of c_γ and C_γ we let the reader be aware that the smaller the constant K, the larger the family of the mixing weights satisfying Assumption 2.2. Therefore it is easier to find configurations of mixing weights which increase the difficulty to detect the departure from the null hypothesis $\mathcal{H}_0$. Nearly colinear mixing weights are an example of such a configuration. Therefore, as expected, the smaller the constant K, the larger (and the worse) the constants C_γ and c_γ. It means that the null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$ have to be separated by a larger distance. This fact is important as it helps the statistician to design its experiment. Indeed a researcher looking for a subtle difference can decide to collect more information if the value of K known a priori is very small.

Computing the exact separation constant is not established in this study (since $c_\gamma < C_\gamma$) as it is a very difficult problem (see e.g. Lepski and Tsybakov [17]). Nevertheless we have clearly established that c_γ and C_γ strongly depend on the smallest eigenvalue of the matrices Γ_n and Γ'_n. This phenomenon is not a surprise when considering Operator Theory and Inverse Problems (see e.g. Brezis [3], Cavalier et al. [5]).

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

4 Open questions

As a conclusion, we have provided a statistical procedure for a testing problem on the mixture components of two populations (Y, Z). This one was proved to be optimal in the minimax sense (Theorems 3.6 and 3.7). In addition, we clearly explained that the weights of the mixture model influence the performance of the statistical rule.

It seems important to give some hints about possible extensions of this work. From the theoretical and practical points of view, it would be interesting to study the same problem without assuming that the mixing weights are exactly known to the statistician. Several explanations can be given

- the statistician can estimate the mixing weights for an observation by using covariates and an appropriate predictive model such as the logistic one,
- a Bayesian approach is chosen for the mixing weights, us information allows the statistician to roughly estimate the mixing weights.

In this case several natural questions arise

- What statistical rule should be considered?
- What kind of performance can be expected for such a rule?
- How much do random mixing weights deteriorate the performance?

Such questions are beyond the scope of this article and their answers certainly involve random matrices theory.

Finally, it would be nice to improve the choice of threshold t_n. Theorem 3.7 provides a complicated value based on asymptotic results.At least heuristics should be proposed for real data; first results in Autin and Pouet [2] seem promizing but needs improvement.

5 Proofs of main results

This section is devoted to the proofs of our results. The proofs often need technical lemmas that have been postponed in Appendix. For the sake of simplicity we sometimes omit $\overrightarrow{p}$ and $\overrightarrow{q}$ in the indices when there is no ambiguity.

5.1 Proofs of propositions and corollaries

Proof of Proposition 2.3: We refer to Maiboroda [21]. A solution of the two optimization problems is, for any indices (l, i), given by

$$a_l(i) = \frac{n}{\det(\Gamma_n)} \sum_{u=1}^{M} (-1)^{l+u} \gamma_{lu} \omega_u(i), \qquad b_l(i) = \frac{n}{\det(\Gamma'_n)} \sum_{u=1}^{M} (-1)^{l+u} \gamma'_{lu} \sigma_u(i)$$

where γ_{lu} and γ'_{lu} are respectively the minor (l, u) of the matrix Γ_n and the minor (l, u) of the matrix Γ'_n. □

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Proof of Proposition 3.1: Let us evaluate the expectation of T_j.

$$\begin{aligned} n^2\, \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_j) &= \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}\Bigg(\sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\sum_{i_1\neq i_2=1}^{n}(a_l(i_1)\phi_{jk}(Y_{i_1}) \\ &\qquad - b_l(i_1)\phi_{jk}(Z_{i_1}))(a_l(i_2)\phi_{jk}(Y_{i_2}) - b_l(i_2)\phi_{jk}(Z_{i_2}))\Bigg) \\ &= \sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\sum_{i_1\neq i_2=1}^{n}\mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}\left[a_l(i_1)\phi_{jk}(Y_{i_1}) - b_l(i_1)\phi_{jk}(Z_{i_1})\right] \\ &\qquad \cdot \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}\left[a_l(i_2)\phi_{jk}(Y_{i_2}) - b_l(i_2)\phi_{jk}(Z_{i_2})\right] \end{aligned}$$

since the random variables (Y_{i_1}, Z_{i_1}) and (Y_{i_2}, Z_{i_2}) are independent.

We have for all $1 \le i \le n$,

$$\mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}\left[a_l(i)\phi_{jk}(Y_i) - b_l(i)\phi_{jk}(Z_i)\right] = \int_{\mathbb{R}}\left(\sum_{u=1}^{M}(a_l(i)\omega_u(i)p_u - b_l(i)\sigma_u(i)q_u)\right)\phi_{jk}.$$

By introducing the diagonal term $i_1 = i_2$ in the sum, we get

$$\begin{aligned} \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_j) &= \frac{1}{n^2}\sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\left(\int_{\mathbb{R}}\phi_{jk}\left(\sum_{i=1}^{n}\sum_{u=1}^{M}a_l(i)\omega_u(i)p_u - \sum_{i=1}^{n}\sum_{u=1}^{M}b_l(i)'\sigma_u(i)q_u\right)\right)^2 \\ &\qquad - \frac{1}{n^2}\sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\left(\int_{\mathbb{R}}(a_l(i)f_i - b_l(i)g_i)\,\phi_{jk}\right)^2 \\ &= \sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\left(\int_{\mathbb{R}}(p_l - q_l)\phi_{jk}\right)^2 \\ &\qquad - \frac{1}{n^2}\sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\left(\int_{\mathbb{R}}(a_l(i)f_i - b_l(i)g_i)\,\phi_{jk}\right)^2, \end{aligned}$$

because of properties $\sum_{i=1}^{n}a_l(i)\omega_u(i) = n\delta_{lu}$ and $\sum_{i=1}^{n}b_l(i)\sigma_u(i) = n\delta_{lu}$. Thus the result for the expectation is proved. □

Proof of Corollary 3.3: According to Proposition 3.1 we only have to bound the quantity

$$D_0 := n^{-2}\sum_{l=1}^{M}\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\left(\int_{\mathbb{R}}(a_l(i)f_i - b_l(i)g_i)\,\phi_{jk}\right)^2.$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Using the Cauchy–Schwarz inequality and Lemma A.2, we have

$$\begin{aligned} n^2 D_0 &\le \sum_{l=1}^{M}\sum_{i=1}^{n}\left[\sum_{k\in\mathbb{Z}}\int_{I_{jk}}(a_l(i)f_i - b_l(i)g_i)^2\right] \\ &\le 2\sum_{i=1}^{n}\sum_{l=1}^{M}\left[\sum_{k\in\mathbb{Z}}\int_{I_{jk}}(a_l(i)f_i)^2 + \int_{I_{jk}}(b_l(i)g_i)^2\right] \\ &\le 4L\left(\sum_{i=1}^{n}\sum_{l=1}^{M}a_l^2(i)\|f_i\|_2^2 + \sum_{i=1}^{n}\sum_{l=1}^{M}b_l^2(i)\|g_i\|_2^2\right) \\ &\le \frac{8LMR^2n}{K}, \end{aligned}$$

where last inequality is due to Proposition 2.3 and the fact that for all $1 \le i \le n$ the density functions f_i and g_i belong to $\mathbb{L}_2(R)$. □

Proof of Proposition 3.4: Let us consider the variance of T_j. For all (i_1, i_2), let $h_j(i_1, i_2)$ denote the quantity

$$\begin{aligned} h_j(i_1, i_2) = \sum_{k\in\mathbb{Z}}\sum_{l=1}^{M}&\left(a_l(i_1)\phi_{jk}(Y_{i_1}) - b_l(i_1)\phi_{jk}(Z_{i_1})\right) \\ &\cdot\left(a_l(i_2)\phi_{jk}(Y_{i_2}) - b_l(i_2)\phi_{jk}(Z_{i_2})\right). \end{aligned}$$

The variance of T_j satisfies under $(\overrightarrow{p}, \overrightarrow{q})$

$$\begin{aligned} n^4\ \mathbb{V}\mathrm{ar}(T_j) &= \mathbb{V}\mathrm{ar}\left(\sum_{i_1\neq i_2=1}^{n} h_j(i_1, i_2)\right) \\ &= \sum_{i_1\neq i_2, i_3\neq i_4=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_3, i_4)\right) \\ &= \sum_{i_1\neq i_2=1}^{n}\mathbb{V}\mathrm{ar}\left(h_j(i_1, i_2)\right) + \sum_{i_1\neq i_2=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_2, i_1)\right) \\ &\quad + \sum_{i_1\neq i_2\neq i_3=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_1, i_3)\right) + \sum_{i_1\neq i_2\neq i_3=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_2, i_3)\right) \\ &\quad + \sum_{i_1\neq i_2\neq i_3=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_3, i_1)\right) + \sum_{i_1\neq i_2\neq i_3=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_3, i_2)\right) \\ &\quad + \sum_{i_1\neq i_2\neq i_3\neq i_4=1}^{n}\mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_3, i_4)\right) \\ &:= \sum_{u=1}^{7} D_u. \end{aligned}$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Independence of the random variables leads to

$$D_7 = \sum_{i_1 \neq i_2 \neq i_3 \neq i_4 = 1}^{n} \mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_3, i_4)\right) = 0.$$

Bounds for quantities D_u $(1 \leq u \leq 6)$ are still required. Since the ways to bound D_1 and D_2 (resp. D_3, D_4, D_5 and D_6) are similar, we will only bound D_1 and D_3. Such bounds are given in Lemmas A.6 and A.7. Proof of Proposition 3.4 is a direct consequence of Lemmas A.6 and A.7 by taking $C_T = 2\bar{C}_T \vee 4\tilde{C}_T$. □

5.2 Proofs of Theorems

Proof of Theorem 3.6: Let us fix $0 < \gamma < 1$ and $s > 0$. Under the null hypothesis, we directly use the well-known Bienayme–Chebyshev inequality.

$$\begin{aligned}
\mathbb{P}_{\overrightarrow{p},\overrightarrow{p}}\left(\Delta_s^* = 1\right) &= \mathbb{P}_{\overrightarrow{p},\overrightarrow{p}}\left(T_{j_n} > t_n\right) \\
&\leq \mathbb{P}_{\overrightarrow{p},\overrightarrow{p}}\left(T_{j_n} - \mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}(T_{j_n}) > t_n - \frac{8LMR^2}{Kn}\right) \\
&\leq \mathbb{V}\mathrm{ar}(T_{j_n})\left(t_n - \frac{8LMR^2}{Kn}\right)^{-2} \\
&\leq C_T M^2 2^{j_n}\left(n^2\, K^2\left(t - \frac{8LMR^2}{K}\right)^2 r_n^4\right)^{-1}.
\end{aligned}$$

The last inequality is obtained using Remark 3.5. According to the choices of level j_n and threshold t_n, we have

$$\frac{C_T\ M^2\ 2^{j_n}}{n^2\ K^2\ \left(t - \frac{8LMR^2}{K}\right)^2 r_n^4} \leq \frac{2C_T\ M^2}{K^2\left(t - \frac{8LMR^2}{K}\right)^2}.$$

Therefore it entails that

$$\mathbb{P}_{\overrightarrow{p},\overrightarrow{p}}\left(\Delta_s^* = 1\right) \leq \frac{\gamma}{2}.$$

Under the alternative, we use the expectation of the test statistic and some approximation argument. The second type error is

$$\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}\left(\Delta_s^* = 0\right) = \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}\left(-T_{j_n} + \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_{j_n}) \geq -t_n + \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_{j_n})\right).$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

The wavelet expansion in the Besov space $\mathcal{B}^s_{2,\infty}(R)$ leads to

$$\begin{aligned}
\mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_{j_n}) - t_n &= \sum_{l=1}^{M} \|p_l - q_l\|_2^2 - \sum_{l=1}^{M} \sum_{j \geq j_n} \sum_{k \in \mathbb{Z}} \left(\int_{\mathbb{R}} (p_l - q_l)\psi_{jk} \right)^2 \\
&\quad - \frac{1}{n^2} \sum_{l=1}^{M} \sum_{k \in \mathbb{Z}} \sum_{i=1}^{n} \left(\int_{\mathbb{R}} (a_l(i) f_i - b_l(i) g_i)\, \phi_{j_n k} \right)^2 - t_n \\
&\geq \sum_{l=1}^{M} \|p_l - q_l\|_2^2 - M\ R\ 2^{-2j_n s} - \frac{8LMR^2}{Kn} - t_n \\
&\geq \frac{1}{2} \sum_{l=1}^{M} \|p_l - q_l\|_2^2 - M\ R\ 2^{-2j_n s} - t_n,
\end{aligned}$$

for any n large enough.

As a consequence, applying the Bienayme–Chebychev inequality leads to

$$\begin{aligned}
&\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}} \left(-T_{j_n} + \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_{j_n}) \geq -t_n + \mathbb{E}_{\overrightarrow{p},\overrightarrow{q}}(T_{j_n}) \right) \\
&\leq \frac{C_T M^2 \left(2^{j_n} + n \sum_{l=1}^{M} \|p_l - q_l\|_2^2 + \sqrt{2^{j_n} n} \sum_{l=1}^{M} \|p_l - q_l\|_2 \right)}{n^2 K^2 \left(\frac{1}{2} \sum_{l=1}^{M} \|p_l - q_l\|_2^2 - M\ R\ 2^{-2j_n s} - t_n \right)^2}.
\end{aligned}$$

The choice of j_n and the fact that the functions are in the alternative entail the following upper bound

$$\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}} \left(\Delta_s^* = 0 \right) \leq \frac{C_T M^2 \left(2^{j_n} + n \sum_{l=1}^{M} \|p_l - q_l\|_2^2 + \sqrt{2^{j_n} n} \sum_{l=1}^{M} \|p_l - q_l\|_2 \right)}{K^2 n^2 \left(\frac{1}{2} \sum_{l=1}^{M} \|p_l - q_l\|_2^2 - M\ R\ 2^{-2j_n s} - t\ r_n^2 \right)^2}.$$

According to the choices of j_n and r_n, one gets for n large enough:

$$\begin{aligned}
\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}} \left(\Delta_s^* = 0 \right) &\leq \frac{C_T \left(2^{j_n} + n \sum_{l} \|p_l - q_l\|_2^2 + \sqrt{2^{j_n} n} \sum_{l} \|p_l - q_l\|_2 \right)}{n^2 K^2 \left(2^{-1} - RC^{-2} - tM^{-1}C^{-2} \right)^2 \left(\sum_{l=1}^{M} \|p_l - q_l\|_2^2 \right)^2} \\
&\leq 3\ C_T \left(\left(2^{-1} - RC^{-2} - tM^{-1}C^{-2} \right)^2 K^2 C^4 \right)^{-1}.
\end{aligned}$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

For all $C > C_\gamma$, we finally obtain

$$\mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}\left(\Delta_s^* = 0\right) \leq \frac{\gamma}{2}.$$

The results on the first-type and second-type errors show that if $C > C_\gamma$ the sum of the errors is less than γ. Therefore the upper bound is proved. □

Proof of Theorem 3.7: Let $\gamma \in]0,\ 1[$, $C > 0$ and $C_1 > 0$. We define a subset of $\Theta_1(R, C, r_n, s)$ by

$$\begin{aligned}\tilde{\Theta}_1(R, C, C_1, r_n, s) = \Big\{ \left(\overrightarrow{p}, \overrightarrow{q}\right) : &\forall u \in \{1, \ldots, M\},\ p_u - q_u \in \mathcal{B}^s_{2,\infty}(R),\\ &\forall u \in \{1, \ldots, M\},\ \{x;\ p_u(x) \wedge q_u(x) \geq C_1\} \supseteq [0, 1]\\ &\exists u \in \{1, \ldots, M\},\ (p_u, q_u) \in \Lambda(R, C, r_n)\Big\}.\end{aligned}$$

It is well-known that

$$\begin{aligned}&\inf_{\Delta}\left(\sup_{(\overrightarrow{p},\overrightarrow{q})\in\Theta_0(R)} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta = 1) + \sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\in\\ \Theta_1(R,C,r_n,s)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta = 0)\right)\\ &\geq \inf_{\Delta}\left(\sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\\ \in\Theta_0(R)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta = 1) + \sup_{\substack{(\overrightarrow{p},\overrightarrow{q})\in\\ \tilde{\Theta}_1(R,C,C_1,r_n,s)}} \mathbb{P}_{\overrightarrow{p},\overrightarrow{q}}(\Delta = 0)\right)\\ &\geq 1 - \frac{1}{2}\left\|\mathbb{P}_{\overrightarrow{p},\overrightarrow{p}} - \mathbb{P}_\pi\right\|,\end{aligned}$$

where $\|\cdot\|$ is the $\mathbb{L}_1$-distance and π is an a priori probability measure on the set $\Lambda(R, C, r_n)$. First we define the probability measure π and its support.

Let $\theta = (\theta_1, \ldots, \theta_M)$ denote an eigenvector associated with the smallest eigenvalue of $\Sigma\Sigma^\star$ which is Kn according to Assumption 2.2 and such that $\|\theta\|_2 = 1$.

Recall that here j_n is the same as the one defined in Theorem 3.6. Let $\mathcal{T}$ be the subset of $\mathbb{Z}$ containing every integer k satisfying the following properties

- $k \in \mathcal{T} \Longrightarrow \left[\frac{k-L}{2^{j_n}}, \frac{k+L}{2^{j_n}}\right[\subseteq [0, 1[$,
- $(k, k') \in \mathcal{T} \times \mathcal{T}$ with $k \neq k' \Longrightarrow \left[\frac{k-L}{2^{j_n}}, \frac{k+L}{2^{j_n}}\right[\cap \left[\frac{k'-L}{2^{j_n}}, \frac{k'+L}{2^{j_n}}\right[= \emptyset$.

The cardinality of $\mathcal{T}$ is clearly equal to $T = \lfloor \frac{2^{j_n-1}}{L} \rfloor$ and we denote its elements $k_1, \ldots, k_T$. Let $\zeta_k = +1$ or -1. The following parametric family of functions is considered

$$q_{l,\zeta}(z) = p_l(z) + 2^{s+1} C\sqrt{ML}\,\theta_l \sum_{k\in\mathcal{T}} \zeta_k 2^{-j_n s - \frac{j_n}{2}} \psi_{j_n k}(z).$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Remark that ζ_k does not depend on the index l. Therefore the density of Z_i is

$$g_{i,\zeta}(z) = \sum_{l=1}^{M} \sigma_l(i)\sqrt{ML}\theta_l \, 2^{s+1} C \sum_{k\in\mathcal{T}} \zeta_k 2^{-j_n s - \frac{j_n}{2}} \psi_{j_n k}(z) + \sum_{l=1}^{M} \sigma_l(i)\, p_l(z).$$

The probability measure π is such that the ζ_k's are independent Rademacher random variables with parameter $1/2$.

The function $q_{l,\zeta}$ is a density. Indeed, for n large, $q_{l,\zeta}$ is non-negative. Moreover, as $\psi_{j_n k}$ is a wavelet, we have $\int \psi_{j_n k} = 0$ and therefore $\int q_{l,\zeta} = 1$. If $C < \sqrt{R/M2^{2s+2}}$, then $q_{l,\zeta} - p_l$ belongs to the ball of the Besov space $\mathcal{B}^s_{2,\infty}(R)$. There exists l such that

$$M\theta_l^2 \geq 1 \quad \text{and} \quad \left\| p_l - q_{l,\zeta} \right\|_2^2 = TLMC^2 2^{2+2s-2j_n s - j_n} \theta_l^2 \geq C^2 \, n^{-\frac{4s}{4s+1}}.$$

Therefore the probability measure π is solely concentrated on the alternative.

It is well-known that the $\mathbb{L}_1$ distance can be bounded by the $\mathbb{L}_2$ distance. We have

$$\left\| \mathbb{P}_{\overrightarrow{p},\overrightarrow{p}} - \mathbb{P}_\pi \right\| \leq \sqrt{\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}} \left[\left(\mathbb{E}_\pi \left(\prod_{i=1}^{n} \frac{g_{i,\zeta}(Z_i)}{g_i(Z_i)} \right) \right)^2 \right] - 1}. \tag{5.1}$$

Let us introduce the following random variables

$$\tilde{Z}_{ik} = 2^{s+1} C\sqrt{ML}\, 2^{-j_n s - \frac{j_n}{2}} \frac{\psi_{j_n k}(Z_i)}{g_i(Z_i)} \sum_{l=1}^{M} \theta_l \sigma_l(i).$$

Therefore it suffices to evaluate the second-order moment of the likelihood ratio:

$$\begin{aligned} &\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}} \left[\left(\mathbb{E}_\pi \left(\prod_{i=1}^{n} \frac{g_{i,\zeta}(Z_i)}{g_i(Z_i)} \right) \right)^2 \right] \\ &= \mathbb{E}_{\overrightarrow{p},\overrightarrow{p}} \left[\left(\prod_{k\in\mathcal{T}} \int \prod_{i=1}^{n} \left(1 + \zeta_k \, \tilde{Z}_{ik} \right) \, d\pi(\zeta_1, \ldots, \zeta_T) \right)^2 \right]. \end{aligned}$$

We have

$$\begin{aligned} &\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}} \left[\left(\prod_{k\in\mathcal{T}} \int \prod_{i=1}^{n} \left(1 + \zeta_k \, \tilde{Z}_{ik} \right) \, d\pi(\zeta_1, \ldots, \zeta_T) \right)^2 \right] \\ &= \mathbb{E}_{\overrightarrow{p},\overrightarrow{p}} \left[\prod_{k\in\mathcal{T}} \frac{1}{4} \left[\prod_{i=1}^{n} \left(1 + \tilde{Z}_{ik} \right) + \prod_{i=1}^{n} \left(1 - \tilde{Z}_{ik} \right) \right]^2 \right] \\ &= \mathbb{E}_{\overrightarrow{p},\overrightarrow{p}} \left[\prod_{k\in\mathcal{T}} \frac{1}{2} \left(\prod_{i=1}^{n} \left(1 + \tilde{Z}_{ik}^2 \right) + \prod_{i=1}^{n} \left(1 - \tilde{Z}_{ik}^2 \right) \right) + \sum_{k\in\mathcal{T}} \sum_{i=1}^{n} \tilde{Z}_{ik} \tilde{h}_i(k) \right], \end{aligned}$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

where functions $\tilde{h}_i(k)$ are sums of products of $\tilde{Z}_{j\kappa}$ where the pairs (j,κ) are in the set $\{1,\dots,n\}\times\mathcal{T}\setminus\{(i,k)\}$.

As $\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}(\tilde{Z}_{ik})=0$ and $\tilde{Z}_{ik}\tilde{Z}_{ik'}=0$ for $k\neq k'$, the last term vanishes. Thus only the first term remains. As $\tilde{Z}_{ik}\tilde{Z}_{ik'}=0$ for $k\neq k'$ and the random variables $\tilde{Z}_{ik}$ and $\tilde{Z}_{i'k}$ for $i\neq i'$ are independent, we have

$$\begin{aligned}
&\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left[\prod_{k\in\mathcal{T}}\frac{1}{2}\left(\prod_{i=1}^{n}\left(1+\tilde{Z}_{ik}^2\right)+\prod_{i=1}^{n}\left(1-\tilde{Z}_{ik}^2\right)\right)\right]\\
&\leq\prod_{k\in\mathcal{T}}\left[\frac{1}{2}\left(\prod_{i=1}^{n}\left(1+\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left[\tilde{Z}_{ik}^2\right]\right)+\prod_{i=1}^{n}\left(1-\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left[\tilde{Z}_{ik}^2\right]\right)\right)\right]\\
&\leq\prod_{k\in\mathcal{T}}\cosh\left(\sum_{i=1}^{n}\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left(\tilde{Z}_{ik}^2\right)\right)\\
&\leq\exp\left(\frac{1}{2}\sum_{k\in\mathcal{T}}\left(\sum_{i=1}^{n}\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left(\tilde{Z}_{ik}^2\right)\right)^2\right).
\end{aligned}$$

Each expectation $\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left(\tilde{Z}_{ik}^2\right)$ is bounded as follows

$$\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left(\tilde{Z}_{ik}^2\right)\leq 2^{2s+2-2j_ns-j_n}C^2C_1^{-1}ML\left(\sum_{l=1}^{M}\theta_l\sigma_l(i)\right)^2.$$

Therefore this bound entails

$$\begin{aligned}
&\exp\left(\frac{1}{2}\sum_{k\in\mathcal{T}}\left(\sum_{i=1}^{n}\mathbb{E}_{\overrightarrow{p},\overrightarrow{p}}\left(\tilde{Z}_{ik}^2\right)\right)^2\right)\\
&\leq\exp\left(\frac{1}{2}\sum_{k\in\mathcal{T}}C^4 2^{4s+4-4j_ns-2j_n}L^2M^2C_1^{-2}\left(\sum_{i=1}^{n}\sum_{l,m=1}^{M}\theta_l\theta_m\sigma_l(i)\sigma_m(i)\right)^2\right)\\
&\leq\exp\left(\frac{1}{2}\sum_{k\in\mathcal{T}}2^{4s+4}C^4 2^{-4j_ns-2j_n}L^2M^2C_1^{-2}\left(\theta^\star n\Gamma_n'\theta\right)^2\right)\\
&=\exp\left(\sum_{k\in\mathcal{T}}2^{4s+3}C^4 2^{-4j_ns-2j_n}L^2M^2C_1^{-2}(Kn)^2\right)\\
&\leq\exp\left(2^{4s+2}M^2K^2\ LC^4C_1^{-2}\right).
\end{aligned}\tag{5.2}$$

Inequalities (5.1) and (5.2) lead to

$$\left\|\mathbb{P}_{\overrightarrow{p},\overrightarrow{p}}-\mathbb{P}_{\pi}\right\|\leq\sqrt{\exp\left(2^{4s+2}M^2K^2\ LC^4C_1^{-2}\right)-1}.\tag{5.3}$$

The choice of any constant C such that $C<c_\gamma$ entails that the left-hand side of (5.3) is strictly smaller than $2(1-\gamma)$. □

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

A Appendix

This section contains the technical lemmas used in the proofs of the main results. The proofs of these lemmas are given in Autin and Pouet [2].

Lemma A.1 *For all* $(j,k) \in \mathbb{Z} \times \mathbb{Z}$, *let us put*

$$I_{jk} = \Big[(k-L)2^{-j}, (k+L)2^{-j}\Big[.$$

Then for any fixed (j,k), $\mathrm{Card}\{k' \in \mathbb{Z} : \ I_{jk} \cap I_{jk'} \neq \emptyset\} \leq 4L$.

Lemma A.2 *For any function* $h \in L_1(\mathbb{R})$

$$\sum_{k\in\mathbb{Z}} \int_{I_{jk}} |h(x)|dx \leq 2L\|h\|_1.$$

Lemma A.3 *Let* W *be either* Y *or* Z. *For any* $1 \leq i \leq n$ *and any* (j,k), *we have*

$$\left|\mathbb{E}\left(\phi_{jk}(W_i)\right)\right| \leq \left(2L \sup_{1\leq l\leq M} (\|p_l\|_\infty \vee \|q_l\|_\infty)\right)^{\frac{1}{2}} 2^{-\frac{j}{2}}.$$

Lemma A.4 *Let* W *be either* Y *or* Z *and* c *be either* a *or* b. *For any* $1 \leq i \leq n$ *and any* (j,k), *the following inequalities hold*

$$\sum_{k'\in\mathbb{Z}} \left|\mathbb{E}\left(\phi_{jk}(W_i)\phi_{jk'}(W_i)\right)\right| \leq 4L \sup_{1\leq l\leq M} (\|p_l\|_\infty \vee \|q_l\|_\infty),$$

$$\sup_{1\leq l\leq M} \left|\sum_{k\in\mathbb{Z}} \int \phi_{jk}(p_l - q_l)\right| \leq 4L\,\|\phi\|_\infty\, 2^{\frac{j}{2}},$$

$$\sup_{1\leq l\leq M} |c_l(i)| \leq \sqrt{n \sum_{l=1}^{M} \langle c_l, c_l\rangle_n}.$$

Lemma A.5 *Let* p_l, q_l $p_{l'}$ *and* $q_{l'}$ *be four probability densities in* $\mathbb{L}_2(\mathbb{R})$. *Then, for any* $j \in \mathbb{N}$

$$\sum_{k\in\mathbb{Z}} \left(\int \phi_{jk}p_l - \int \phi_{jk}q_l\right)^2 \leq 2L\|p_l - q_l\|_2^2;$$

$$\sum_{k\in\mathbb{Z}} \sum_{\substack{k'\in\mathbb{Z}:\\ I_{jk}\cap I_{jk'}\neq\emptyset}} \left|\left(\int \phi_{jk}p_l - \int \phi_{jk}q_l\right)\left(\int \phi_{jk'}p_{l'} - \int \phi_{jk'}q_{l'}\right)\right|$$
$$\leq 4L^2\left(\|p_l - q_l\|_2^2 + \|p_{l'} - q_{l'}\|_2^2\right).$$

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

Lemma A.6 *There exists a constant* $\bar{C}_T = \bar{C}_T(R, L, \|\phi\|_\infty) > 0$ *such that*

$$D_1 := \sum_{i_1 \neq i_2 = 1}^{n} \underset{\vec{p}\,,\vec{q}}{\mathbb{V}\mathrm{ar}}(h_j(i_1, i_2)) \leq \bar{C}_T \frac{M^2}{K^2}\, 2^j\, n^2.$$

Lemma A.7 *There exists a constant* $\tilde{C}_T = \tilde{C}_T(R, L, \|\phi\|_\infty) > 0$ *such that for any* $j \in \mathbb{N}$*:*

$$\begin{aligned} D_3 &:= \sum_{i_1 \neq i_2 \neq i_3 = 1}^{n} \mathbb{C}\mathrm{ov}\left(h_j(i_1, i_2), h_j(i_1, i_3)\right) \\ &\leq \tilde{C}_T \frac{M^2}{K^2} \left[n^3 \sum_{l=1}^{M} \|p_l - q_l\|_2^2 + 2^{\frac{j}{2}}\, n^{\frac{5}{2}} \sum_{l=1}^{M} \|p_l - q_l\|_2 \right]. \end{aligned}$$

Acknowledgements. The authors wish to thank the referees for their useful comments and Y. Golubev for stimulating discussions.

References

[1] F. Autin. Maxiset for density estimation on $\mathbb{R}$. *Mathematical Methods Statistics*, 15:123–145, 2006.

[2] F. Autin and C. Pouet. Test on the components of mixture densities. *(Version with detailed proofs)*, arXiv:0912.0786v1, 2010.

[3] H. Brézis. *Analyse Fonctionnelle. Théorie et Applications*. Dunod, Paris, 1999.

[4] C. Butucea and K. Tribouley. Nonparametric homogeneity tests. *Journal of Statistical Planning and Inference*, 136:597–639, 2006.

[5] L. Cavalier, G. Golubev, D. Picard, and A. Tsybakov. Oracles inequalities for inverse problems. *Annals of Statistics*, 30:843–874, 2002.

[6] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Applied Computational Harmonic Analysis*, 11:167–191, 2001.

[7] I. Daubechies. *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1996.

[8] C. Delmas. On likelihood ratio tests in Gaussian mixture models. *The Indian Journal of Statistics*, 65:513–531, 2003.

[9] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet tresholding. *Annals of Statistics*, 24:508–539, 1996.

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

[10] B. Garel. Likelihood ratio test for univariate Gaussian mixture, *Journal of Statistical Planning and Inference*, 96:325–350, 2001.

[11] B. Garel. Asymptotic theory of the likelihood ratio test for the identification of a mixture, *Journal of Statistical Planning and Inference*, 131:271–296, 2005.

[12] G. Gayraud and C. Pouet. Adaptive minimax testing in the discrete regression scheme. *Probability of Theory and Related Fields*, 133:531–558, 2005.

[13] P. Hall. On the nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, Ser. B, 43:147–156, 1981.

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

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

[16] D. W. Hosmer. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three types of sample. *Biometrics*, 29:761–770, 1973.

[17] O. V. Lepski and A. B. Tsybakov. Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields*, 117:17–48, 2000.

[18] N. Lodatko and R. Maiboroda. Estimation of the density of a distribution from observations with an admixture. *Theory of Probability and Mathematical Statistics*, 73:99–108, 2007.

[19] P. E. McKnight, K. M. McKnight, A. J. Figueredo, and S. Sidani. *Missing Data: a Gentle Introduction*, Guilford Press, New York, 2007.

[20] R. E. Maiboroda. Estimates for distribution of components of mixtures with varying concentrations. *Ukrainian Journal of Mathematics*, 48:618–622, 1996.

[21] R. E. Maiboroda. An asymptotically effective estimate for a distribution from a sample with a varying mixture. *Theory of Probability and Mathematical Statistics*, 61:121–130, 2000.

[22] R. E. Maiboroda. A test for the homogeneity of mixtures with varying concentrations. *Ukrainian Journal of Mathematics*, 52:1256–1263, 2000.

[23] D. Pokhyl'ko. Wavelet estimators of a density constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics*, 70:135–145, 2005.

[24] J. Qin. Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics*, 27:1368–1384, 1999.

[25] V. G. Spokoiny. Adaptive hypothesis testing using wavelets, *Annals of Statistics*, 24:2477–2498, 1996.

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.

[26] D. M. Titterington. Minimum distance nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, Ser. B, 45:37–46, 1983.

[27] S. van de Geer. Asymptotic normality in mixture models. *ESAIM: Probability and Statistics*, 1:17–33, 1995.

Florent Autin
Université Aix-Marseille 1
C.M.I.
39, rue F. Joliot Curie
13453 Marseille Cedex 13
France
autin@cmi.univ-mrs.fr

Christophe Pouet
Ecole Centrale de Marseille
38, rue F. Joliot Curie
13451 Marseille Cedex 20
France
cpouet@centrale-marseille.fr

This article is protected by German copyright law. You may copy and distribute this article for your personal use only. Other use is only allowed with written permission by the copyright holder.