Semiparametric density testing in the contamination model

Denys Pommeret, Pierre Vandekerkhove

Semiparametric density testing in the contamination model

Denys Pommeret and Pierre Vandekerkhove

Institut Mathématique de Marseille
Campus de Luminy,
13288 Marseille Cedex 9, France
e-mail: denys.pommeret@univ-amu.fr

Université Paris-Est Marne-la-Vallée
LAMA (UMR 8050), UPEMLV
F-77454, Marne-la-Vallée, France
and
UMI Georgia Tech - CNRS 2958,
Georgia Institute of Technology, USA
e-mail: pierre.vandekerkhove@u-pem.fr

Abstract: In this paper we investigate a semiparametric testing approach to answer if the parametric family allocated to the unknown density of a two-component mixture model with one known component is correct or not. Based on a semiparametric estimation of the Euclidean parameters of the model (free from the null assumption), our method compares pairwise the Fourier's type coefficients of the model estimated directly from the data with the ones obtained by plugging the estimated parameters into the mixture model. These comparisons are incorporated into a sum of square type statistic which order is controlled by a penalization rule. We prove under mild conditions that our test statistic is asymptotically $\chi^2(1)$-distributed and study its behavior, both numerically and theoretically, under different types of alternatives including contiguous nonparametric alternatives. We discuss the counterintuitive, from the practitioner point of view, lack of power of the maximum likelihood version of our test in a neighborhood of challenging non-identifiable situations. Several level and power studies are numerically conducted on models close to those considered in the literature, such as in McLachlan *et al.* (2006), to validate the suitability of our approach. We also implement our testing procedure on the Carina galaxy real dataset which low luminosity mixes with the one of its companion Milky Way. Finally we discuss possible extensions of our work to a wider class of contamination models.

MSC 2010 subject classifications: Primary 62F03, 28C20; secondary 33C45.
Keywords and phrases: Asymptotic normality, Chi-squared test, False Discovery Rate, maximum likelihood estimator, nonparametric contiguous alternative, semiparametric estimator, two-component mixture model..

1. Introduction

Let us consider n independent and identically distributed random variables $(X_1, \ldots, X_n)$ drawn from a two-component mixture model with probability den-

sity function g defined by:

$$g(x) = (1-p)f_0(x) + pf(x), \quad x \in \mathbb{R}, \tag{1.1}$$

where f_0 is a known probability density function, corresponding to a known signal, and where the unknown parameters of the model are the mixture proportion $p \in (0,1)$ and the probability density function $f \in \mathcal{F}$ (a given class of densities) associated to an unknown signal. Model (1.1) is widely used in statistics and is usually so-called the *contamination* model. This class of models is especially suitable for detection of differentially expressed genes under various conditions in microarray data analysis, see McLachlan *et al.* (2006) or Dai and Charnigo (2010). In astronomy such a model has been used to model mixtures of X-ray sources, see Melchior and Goulding (2018) and Patra and Sen (2016). Recently some applications have been also developed in selective Statistical Editing, see Di Zio and Guarnera (2013), in biology to model trees diameters, see Podlaski and Roesch (2014) or in kinetics to model plasma data, see Klingenber *et al.* (2018).

Many techniques have been proposed to estimate the Euclidean and functional parameters p and f in model (1.1). The most popular methods for known finite order mixture models, such as the moment method, see Lindsay (1989), the moment generating function based method, see Quandt and Ramsey (1978), or the maximum likelihood method, see Lindsay (1983), are largely used but suffer from the requirement of assigning a parametric form to the f density. Since then, some semiparametric approaches have been developed, such as the pioneer work by Bordes *et al.* (2006), to relax that parametric modelling. These authors only restricted, for example, their study to the class of location-shift symmetric densities in order to make model (1.1) semiparametrically identifiable. More recently, different nonparametric approaches have been also considered, such as in Nguyen and Matias (2014) where f_0 is a uniform distribution on $[0,1]$. In Ma and Yao (2015), where f_0 is only supposed to belong to a parametric family, a tail identifiability approach is used, considering symmetric distributions embedded in a nonparametric envelop. We also recommend the recent work by Al Mohammad and Boumahdaf (2018) who consider situations where the unknown component f is defined through linear constraints. In Balabdaoui and Doss (2018) a log-concave assumption is done on the family $\mathcal{F}$ to insure the identifiability of the model. In Patra and Sen (2016) the identifiability and estimation problem is considered under tail conditions with very few shape constraints assumptions.

The goal of the present paper is to answer a very natural question, explicitly raised in McLachlan *et al.* (2006, Section 6) or Patra and Sen (2016, Section 9.2), which is basically "can we test if the unknown component of the contamination model belongs to a given class of parametric densities ?", or more formally can we test

$$H_0 : f \in \mathcal{F} = \{f_\theta; \theta \in \Theta\} \qquad \text{against} \qquad H_1 : f \notin \mathcal{F}, \tag{1.2}$$

where f_θ is a probability density function parametrized by an Euclidean parameter θ belonging to a parametric space Θ. For simplicity we will restrict ourselves

to the case where f_θ is a symmetric probability density function with respect to a location parameter $\mu \in \mathbb{R}$, as described in (2.1), but discuss in Section 10 how our approach can be generalized to any class of parametric densities provided that model (1.1) can be $\sqrt{n}$-estimated semiparametrically. This problem has been considered recently by Suesse *et al.* (2017), who use a maximum likelihood estimate-based testing approach. In general the behavior of the maximum likelihood estimator is difficult to control or figure out, as illustrated in Section 7, under the alternative since the model is then misspecified. To get a consistent testing method under both H_0 and H_1, at the price of some shape restriction about H_1, we propose to use an $H_0 \cup H_1$ consistent semiparametric estimation approach in order to build a H_0-free statistic (do not forcing to fit into the parametric model). To the best of our knowledge this is the first time that an H_0-free semiparametric approach is used to test mixture models. The advantage of this new strategy will be demonstrated, both theoretically and numerically, on very counterintuitive examples in the close neighborhood of non-identifiable situations, see Fig. 1 and comments. For a general overview about semiparametric mixture models we recommend the recent surveys by Xian *et al.* (2018) or Gassiat (2018). Note that the test against a specific distribution, proposed in Bordes and Vandekerkhove (2010, Section 4.1), does not allow to test versus a complete class of probability density functions which is our goal here. To point out the interest of the statistical community about the contamination problem testing, let us mention the very recent work by Arias-Castro and Huang (2018) on the sparse variance contamination model testing and references therein.

The main idea of our test is based on the data driven smooth test procedure developed by Ledwina (1994), extending the idea of Neyman (1934), which consists in estimating the expansion coefficients of f in an orthogonal basis, first assuming $f \in \mathcal{S}$ (the set of symmetric probability density functions with respect to a location parameter $\mu \in \mathbb{R}$), and to compare thes estimates to those obtained by assuming $f \in \mathcal{F}$. This approach has been used in Doukhan *et al.* (2015), see also references therein, but the specificity of the two-component mixture model necessitates a special adaptation of the Neyman smooth test. In our case we develop a two rates procedure, one rate driven by the asymptotic normality of the test statistic and another one driven by the almost sure rate of convergence of the semiparametric estimators. As we will discuss along this paper, the approach of Suesse *et al.* (2017), restricted to model (1.1), does not allow to investigate the asymptotic behavior of the test statistic under alternative assumptions (possibly contiguous) since the asymptotic behavior of the maximum likelihood estimator cannot be controlled properly under distribution misspecification. Another aspect of our nonparametric approach is that it can easily deal with situations where f_0 is only known through a training data. This situation is illustrated in Section 9 through a real dataset collecting the radial velocity of the Carina galaxy and its companion Milky Way.

The paper is organized as follows: in Section 2 we describe our two-step test methodology; in Section 3 we state the assumptions and asymptotic results under the null hypothesis; Section 4 is dedicated to the test divergence under the alternative; Section 5 is devoted to the study of our testing procedure under

contiguous nonparametric alternatives (inspired from the parametric contiguous alternative concept); in Section 6 we discuss the choice of the reference measure when considering orthogonal bases for the unknown density decomposition; in Section 7 we conduct a power comparison between the semiparametric and maximum likelihood versions for our test, this section enlightens interestingly the fact that a maximum likelihood approach could force, in certain setups of the McLachlan *et al.* (2006, Section 6) Gaussian mixture model, to consider the number q of components defining f equal to 1 when in reality $q = 2$; Section 8 is dedicated to a simulation-based empirical and power levels study; in Section 9 we proceed with the application of our testing method to the datasets (breast cancer, colon cancer, HIV) previously studied in McLachlan *et al.* (2006) and to the Galaxy dataset studied in Patra and Sen (2016). Finally in Section 10 we discuss further leads of research connected with the contamination model testing problem.

2. Testing problem

Let us consider an independent and identically distributed sample denoted $(X_1, \ldots, X_n)$, drawn from a probability density function g defined in (1.1) with respect to a given reference measure ν. The problem addressed in this section deals with testing the unknown component f assuming the fact that f belongs to $\mathcal{S}$, the set of symmetric densities provided with the identifiability conditions in Bordes and Vandekerkhove (2010, p. 25). More precisely, denoting $\mathcal{F} = \{f_{(\mu,\theta)}; (\mu, \theta) \in \Lambda\}$ the set of densities with respect to ν, with mean μ and shape parameter θ where (μ, θ) is supposed to belong to a compact set Λ of $\mathbb{R} \times \Theta$, our goal is to test

$$H_0 : f \in \mathcal{F} \qquad \text{against} \qquad H_1 : f \in \mathcal{S} \backslash \mathcal{F}. \tag{2.1}$$

Our test procedure is based on the Ledwina (1994) approach and consists in estimating the expansion coefficients of the unknown density f in an orthogonal basis, first assuming $f \in \mathcal{S}$, and comparing in contrast these estimates to those obtained when f is supposed to belong strictly to the sub-parametric family $\mathcal{F}$. As intuitively expected, we will show how the study of the successive expansion coefficient differences helps in detecting possible departure from H_0 given the data. We will denote by $\mathcal{Q} = \{Q_k; k \in \mathbb{N}\}$, a ν-orthogonal basis satisfying $Q_0 = 1$ and such that

$$\int_{\mathbb{R}} Q_j(x) Q_k(x) \nu(dx) = q_k^2 \delta_{jk}, \tag{2.2}$$

with $\delta_{jk} = 1$ if $j = k$ and 0 otherwise, and where the normalizing factors $q_k^2 \geq 1$ will permit to control the variance of our estimators, as illustrated in Lemmas 1 and 3. We assume that $\mathcal{Q}$ is an $L^2(\mathbb{R}, \nu)$ Hilbert basis, which is satisfied if there exists $\theta > 0$ such that $\int_{\mathbb{R}} e^{\theta|x|} \nu(dx) < \infty$, and that the following integrability conditions are satisfied:

$$\int_{\mathbb{R}} f_0^2(x) \nu(dx) < \infty \qquad \text{and} \qquad \int_{\mathbb{R}} f^2(x) \nu(dx) < \infty.$$

Then, for all $x \in \mathbb{R}$, we have

$$g(x) = \sum_{k\geq 0} a_k Q_k(x) \qquad \text{with} \qquad a_k = \int_{\mathbb{R}} Q_k(x) g(x) \nu(dx)/q_k^2,$$
$$f_0(x) = \sum_{k\geq 0} b_k Q_k(x) \qquad \text{with} \qquad b_k = \int_{\mathbb{R}} Q_k(x) f_0(x) \nu(dx)/q_k^2,$$
$$f(x) = \sum_{k\geq 0} c_k Q_k(x) \qquad \text{with} \qquad c_k = \int_{\mathbb{R}} Q_k(x) f(x) \nu(dx)/q_k^2.$$

From (1.1) we have

$$a_k = (1-p)b_k + pc_k.$$

Let us denote by Z a random variable with density $f_{\mu,\theta}$ and consider

$$\alpha_k(\mu,\theta) = \mathbb{E}(Q_k(Z))/q_k^2.$$

The null hypothesis can be rewritten as $c_k = \alpha_k(\mu,\theta)$, for all $k \geq 1$, or equivalently as

$$H_0 : \ a_k = (1-p)b_k + p\alpha_k(\mu,\theta), \qquad \text{for all } k \geq 1. \tag{2.3}$$

Since the probability density function f_0 is known, the coefficients b_k are automatically known. As a consequence, for all $k \geq 1$, the coefficients a_k can be estimated empirically by:

$$a_{k,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{Q_k(X_i)}{q_k^2}, \quad n \geq 1.$$

To avoid possible compensation phenomenon under H_1 between the estimation of $\vartheta = (p,\mu)$ and the estimation of the α_k's, the estimator of (p,μ) will be obtained without assuming the null hypothesis, that is using the semiparametric estimator $\bar{\vartheta}_n = (\bar{p}_n, \bar{\mu}_n)$ introduced in Bordes *et al.* (2006) and studied more deeply in Bordes and Vandekerkhove (2010). Indeed, as numerically demonstrated in Section 7, the maximum likelihood estimator $(\widehat{p}_n, \widehat{\mu}_n, \widehat{\theta}_n)$ under the null assumption tends to provide the best H_0-fitted model when the semiparametric estimator of Bordes and Vandekerkhove (2010) is not influenced by this constraint and can provide very distant, Euclidean and functional, estimations under H_1 (when the model is misspecified under the null assumption). In the same way, considering the relation (1.1), the estimator of θ is obtained by the H_0-free semiparametric plug-in moment method satisfying

$$\mathbb{E}_\theta(X_1^p) = \frac{1}{n}\sum_{i=1}^{n} X_i^p, \tag{2.4}$$

where $\mathbb{E}_\theta(X_1^p)$ means that we express this expectation as a function of θ. The estimator of $\alpha_k(\mu,\theta)$ is obtained by using a standard plug-in approach, that is:

$$\alpha_{k,n} = \alpha_k(\bar{\mu}_n, \bar{\theta}_n).$$

To illustrate our general approach, let us detail the Gaussian case here. If $\mathcal{F}$ is equal to $\mathcal{G}$ the set of normal densities with mean μ and variance $\theta = s$, then the plug-in moment yields

$$\bar{s}_n = \frac{\bar{M}_{2,n} - (1 - \bar{p}_n)}{\bar{p}_n} - (\bar{\mu}_n)^2, \tag{2.5}$$

where $\bar{M}_{2,n} = n^{-1} \sum_{i=1}^n X_i^2$. Now coming back to generality, looking at the H_0 reformulation in (2.3) we expect that the differences

$$R_{k,n} = a_{k,n} - \bar{p}_n(\alpha_{k,n} - b_k) - b_k, \quad \text{for all } k \geq 1,$$

will allow us to detect any possible departure from the null hypothesis. For simplicity matters and without loss of generality, since the b_k's are known constants, we assume from now on them to be equal to zero. For all $k \geq 1$, we define the k-th order coefficient of our test statistic (incorporating the k-th order departure information from H_0)

$$T_{k,n} = n U_{k,n}^\top \widehat{D}_{k,n}^{-1} U_{k,n}, \tag{2.6}$$

where $U_{k,n} = (R_{1,n}, \ldots, R_{k,n})$ and where $\widehat{D}_{k,n}$ is an estimator of

$$D_{k,n} = \mathrm{diag}(\mathrm{var}(R_{1,n}), \ldots, \mathrm{var}(R_{k,n})),$$

normalizing the test statistic as in Munk *et al.* (2010). To avoid instability in the evaluation of $\widehat{D}_{k,n}^{-1}$, following Doukhan *et al.* (2015), we add a trimming term $e(n)$ to every i-th, $i = 1, \ldots, k$, diagonal element of $\widehat{D}_{k,n}$ as follows:

$$\widehat{D}_{k,n}[i] = \max(\widehat{\mathrm{var}}(R_{i,n}), e(n)), \quad 0 \leq i \leq k, \tag{2.7}$$

where $\widehat{\mathrm{var}}(R_{i,n})$ is a weakly consistent estimator of $\mathrm{var}(R_i)$ as $n \to +\infty$, and $e(n) \to 0$.

Following Ledwina (1994) and Inglot *et al.* (1997), we suggest a data driven procedure to select automatically the number of coefficients needed to answer the testing problem. We introduce the following penalized rule to pick parcimouniously (trade-off between H_0 departure detection and complexity of the procedure involved by index k) the "*best*" rank k for looking at $T_{k,n}$:

$$S_n = \min\Big\{ \underset{1 \leq k \leq d(n)}{\mathrm{argmax}} \, (s(n) T_{k,n} - \beta_k pen(n)) \Big\}, \tag{2.8}$$

where $s(n) \to 0$ is a normalizing rate, $d(n) \to +\infty$ as $n \to +\infty$, $pen(n)$ is a penalty term such that $pen(n) \to +\infty$ as $n \to +\infty$, and the β_k's are penalization factors. In practice we will consider $\beta_k = k$, $k \geq 1$, and $pen(n) = \log(n)$, $n \geq 1$. To match the asymptotic normality regime, under H_0, of the test statistic $T_{k,n}$ defined in (2.6), the normalizing factor $s(n)$ is usually taken equal to one, but in our case, due to the specificity of the semiparametric mixture estimation (possibly adapted to nonparametric contiguous alternatives), we chose:

$$s(n) = n^{\lambda - 1}, \text{ with } \lambda \in]0, 1/2[. \tag{2.9}$$

The above calibration is connected with the almost sure convergence rate of the estimators $\bar{p}_n$ and $\bar{\mu}_n$ (see Theorem 3.1 in Bordes and Vandekerkhove, 2010). Note that the selection rule in (2.8), adapted to the semiparametric framework, strongly differs from the BIC criterion used by Suesse *et al.* (2017, p. 9).

Remark 1. *It is important to notice at this point that we could have also investigated a test expressed like this:*

$$H_0 : \text{ there exists } \theta \in \Theta \text{ such that } F = F_\theta,$$

against its alternative, where F denotes the cumulative distribution function of f. To simplify, consider the Gaussian case with variance parameter $s = \theta$ and write $F_s(\cdot) = F(\sqrt{s} \times \cdot)$. Write also $F_{(0,1)}$ the standard Gaussian cumulative distribution function. In such a perspective we could have used a strategy inspired from the simple hypothesis test of Bordes and Vandekerkhove (2010, Section 4.1). Since according to Theorem 3.2 in Bordes and Vandekerkhove (2010) the semiparametric estimator $\widehat{F}_n$ of F satisfies a functional central limit theorem, one could consider s_n in (2.5) as a natural estimate of s under H_0 and evaluate the square of

$$\sqrt{n}[\widehat{F}_{n,s_n} - F_{(0,1)}] = \sqrt{n}[\widehat{F}_{n,s_n} - F_{s_n}] + \sqrt{n}[F_{s_n} - F_{(0,1)}]$$

over a set of fixed values $(x_1, \ldots, x_k)$, where $\widehat{F}_{n,s_n}(\cdot) = \widehat{F}_n(\sqrt{s_n}\times\cdot)$. By using the delta method, we can show that the second term of the above quantity is asymptotically normal, however the behavior of the first term looks much more difficult to analyze due to the random factor term s_n inside the semiparametric estimate $\widehat{F}_n$. In addition of this technical difficulty, it would also be more satisfactory to investigate a Kolmogorov type test based on $\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(\sqrt{s_n}x) - F_{(0,1)}(x)|$, embracing the whole complexity of $F_{(0,1)}$, instead of a $\chi^2(k)$-type test based on the above expression evaluated over a k-grid. Again this is a very challenging problem. In that sense our approach allows to get a sort of asymptotic framework to capture the whole complexity of f through its (asymptotically unrestricted) decomposition in a base of orthogonal functions.

3. Assumptions and asymptotic behavior under H_0

To test consistently (2.1), based on the statistic $T(n) = T_{S_n,n}$, we will suppose the following conditions:

- **(A1)** The coefficient order upper bound $d(n)$ involved in (2.8) satisfies $d(n) = O(\log(n)e(n))$, where $e(n)$ is the trimming term in (2.7).
- **(A2)** For all $k \geq 1$, $\alpha_k(\cdot,\cdot)$ is a $\mathcal{C}^1$ function and there exists nonnegative constants M_1 and M_2 such that for all $(\mu,\theta) \in \Lambda$,

$$|\alpha_k(\mu,\theta)| \leq M_1 \quad \text{and} \quad \|\dot{\alpha_k}(\mu,\theta)\| \leq M_2,$$

where $\dot{\alpha_k}$ denotes the gradient $(\partial\alpha_k/\partial\mu, \partial\alpha_k/\partial\theta)^T$ and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^2$.

(A3) There exists a nonnegative constant M_3 such that for all $(k,i) \in \mathbb{N}^* \times \mathbb{N}^*$,

$$\frac{1}{k}\sum_{i=1}^{k} \mathrm{var}\left(\frac{Q_i(X_1)}{q_i^2}\right) \leq M_3.$$

Under these three conditions, which will be checked respectively in Lemma 1 and 3 for the Gaussian and the Lebesgue reference measure, we state the following theorem.

Theorem 2. *If assumptions* **(A1-3)** *hold, then, under H_0, S_n converges in Probability towards 1 as $n \to +\infty$.*

Corollary 3. *Under* **(A1-3)**, *the test statistic $T(n)$ converges in law towards a χ^2-distribution with one degree of freedom as $n \to +\infty$.*

Remark 4. *Theorem 2 and Corollary 3 still hold if we replace in $T(n)$ the semiparametric estimators and their (asymptotic) variances by their maximum likelihood counterparts. The proofs of these two results are completely similar to the semiparametric case and rely on the asymptotic normality of the maximum likelihood estimator detailed in the supplementary material file. In this case the rate of the selection rule is the standard one, which is namely $s(n) = 1$.*

4. Asymptotic behavior under H_1

In the next proposition we study the behaviour of our test statistic under $H_1 : \ f \in \mathcal{S} \setminus \mathcal{F}$.

Proposition 1. *If $f \in \mathcal{S} \setminus \mathcal{F}$, then the test statistic $T(n)$ tends to $+\infty$ in probability with a n^λ-drift, $0 < \lambda < 1/2$, as $n \to +\infty$.*

We would like to stress out the fact that the identifiability conditions supposed when considering the class of densities $\mathcal{S}$, see definition in Section 2, are crucial in the proof of Proposition 1. As mentioned in Bordes, Delmas and Vandekerkhove (2006), there exists various non identifiability cases for model (1.1). Let us remind the following one from Bordes and Vandekerkhove (2010):

$$(1-p)\varphi(x) + pf(x-\mu) = (1-\frac{p}{2})\varphi(x) + \frac{p}{2}\varphi(x-2\mu), \quad x \in \mathbb{R}$$

where φ is an even probability density function, $p \in (0,1)$ and $f(x) = (\varphi(x-\mu) + \varphi(x+\mu))/2$. This example is very interesting since it clearly shows the danger of estimating model (1.1) when the probability density function of the unknown component has exactly the same shape as the known component. In particular if φ is a given Gaussian distribution and we want to test if the 2nd component is Gaussian, we could possibly either reject or accept H_0 with our testing procedure depending on the convergence of our semiparametric estimators. Indeed the maximum likelihood estimator would converge towards the natural underlying Gaussian model and the semiparametric method could possibly converge towards both solutions. To avoid this very well identified concern,

we recommend to check if the departures between the maximum likelihood estimator and the semiparametric one is not driven by a factor 2, *i.e* $\widehat{\mu}_n \approx 2\bar{\mu}_n$ and $\widehat{p}_n \approx \bar{p}_n/2$. To advise on this possible proximity, one could check if $\widehat{\mu}_n/2$ and $2\widehat{p}_n$ respectively belong to the 95% confidence intervals of μ and p derived from the asymptotic normality of $(\bar{p}_n, \bar{\mu}_n)$, see Bordes and Vandekerkhove (2010). Now if so, we suggest to initialize the semiparametric approach close the maximum likelihood estimator to force it to detect the possibly existing f-component in model (1.1).

5. Contiguous alternatives

5.1. Detected contiguous alternatives

We consider in this section a *vanishing* convolution-class of nonparametric contiguous alternatives. More specifically, the null hypothesis consists here in considering that the observed sample $\mathbf{X}^n = (X_1, \ldots, X_n)$ comes from

$$H_0 : \quad X_i = (1 - U_i)Y_i + U_i Z_i, \quad i = 1, \ldots, n,$$

where $(U_i)_{i\geq 1}$ and $(Y_i, Z_i)_{i\geq 1}$ are respectively independent and identically distributed sequences distributed according to a Bernoulli distribution with parameter p and $f_0 \otimes f_{\mu,\theta}$, where $f_{\mu,\theta}$ is the unknown density function with respect to the reference measure ν. For each $n \geq 1$, the contiguous alternative consists in the fact that the observed sample $\mathbf{X}^{(n)} = (X_1^n, \ldots, X_n^n)$ comes from a *row independent* triangular array:

$$H_1^{(n)} : \quad X_i^n = (1 - U_i)Y_i + U_i Z_i^n, \quad i = 1, \ldots, n,$$

where $Z_i^n = Z_i + \delta_n \varepsilon_i$, $(\varepsilon_i)_{i\geq 1}$ is an independent and identically distributed sequence of random variables, independent from the Z's and $\delta_n \to 0$ as $n \to +\infty$ (vanishing factor). We assume here that, $\forall i \geq 1$, $Z_i + \delta_n \varepsilon_i \notin \mathcal{S}$. In the Gaussian case this assumption is insured if the $\varepsilon' s$ are non Gaussian. It is also assumed that the $\mathbb{E}(e^{|\epsilon_1|} < \infty$. This type of contiguous modeling looks natural to us as, in any experimental field, measurement errors could happen, represented above by the $\delta_n \varepsilon_i$'s, and additively impact the Z true underlying phenomenon. We also remind at this point that the distribution of the Y's is theoretically known by assumption.

The whole contiguous models collection will be denoted $H_1^* = \otimes_{n=1}^{\infty} H_1^{(n)}$. To emphasize the role of index n in the triangular array, we will denote all the estimators depending on $\mathbf{X}^{(n)}$ or any function depending on $G^{(n)}$, the cumulative distribution function of the $X_i^{(n)}$'s, with the extra superscript $^{(n)}$; for example, with this new notational rule, the estimator $\bar{p}_n(\mathbf{X}^{(n)})$ of p will be denoted $\bar{p}_n^{(n)}$. Similarly we will denote by $\widehat{g}_n^{(n)}$ the kernel density estimator of $g^{(n)}$ involved in the contiguous alternative setup, see the supplementary material file, defined by

$$\widehat{g}_n^{(n)}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i^n}{h_n}\right), \quad x \in \mathbb{R}, \tag{5.1}$$

where the bandwidth h_n satisfies $h_n \to 0$, $nh_n \to +\infty$ and K is a symmetric kernel density function detailed in the supplementary file. We will denote also by $\mathbb{E}^{(n)}$ and $\mathbb{P}^{(n)}$ the expectation and probability distribution under the alternative $H_1^{(n)}$ and consider the following assumptions:

(A4) The bandwidth setup is $h_n = n^{-1/4-\gamma}$ with $\gamma \in (0, 1/12)$.
(A5) The vanishing factor satisfies $\delta_n = n^{-3/4-\xi}$, with $3\gamma < \xi < 2\gamma + 1/4$.
(A6) There exists a nonnegative constant C such that for all $k \in \mathbb{N}$,

$$|\mathbb{E}^{(n)}\left(Q_k(X_0 + \delta_n\varepsilon_1) - Q_k(X_0)\right)|/q_k^2 \quad \leq \quad C\delta_n,$$

where X_0 is H_0 distributed.

Condition **(A6)** is checked in Lemmas 2-4 for the Gaussian and the Lebesgue reference measure. It is also satisfied for any reference measure with bounded support. For simplicity, we refer to condition **(A2-3)** under H_1^* in the proposition below. This means that both conditions are satisfied for all $n \geq 1$ replacing X_1 by X_1^n. Following the proof of these conditions in Appendix under H_0 it is possible to establish explicit moment conditions on ε, adapted to the moments of Z, to insure **(A2-3)** under H_1^*. These conditions being technical and their proof being painful but straightforward we do not detail them here.

Proposition 2. *If assumptions* **(A1-6)** *hold, then, under* H_1^*, S_n *converges in Probability towards 1 and* $T(n)$ *converges in law towards a* χ^2*-distribution with one degree of freedom, as* $n \to +\infty$.

5.2. Undetected contiguous alternatives

Combining Assumptions **(A4)** and **(A5)**, we clearly have $0 < \xi < 1/3$ and then there exists $\widetilde{\xi} = 3/4+\xi \in (3/4, 13/12)$ such that $\delta_n = n^{-\widetilde{\xi}}$. The convergence rate of δ_n to zero is slow enough to distinguishe the asymptotic null hypothesis when n tends to infinity. Contrarily, we now consider two convergence rates which are too fast to recover the asymptotic null distribution of the test statistic, despite the convergence of the contiguous alternative towards the null hypothesis. These convergence rates are given under the following assumptions:

(A7) $\mathbb{E}(\varepsilon) = 0$ and there exists $0 < \xi' < 1/4$ such that $\delta_n = n^{-\xi'}$.
(A8) $\mathbb{E}(\varepsilon) \neq 0$ and there exists $0 < \xi'' < 1/8$ such that $\delta_n = n^{-\xi''}$,

where ε denotes a generic random variable involved in the above definition of the Z^n's. The rate in **(A7)** will control the mean deviation due to the perturbations ε and the rate given in **(A8)** will allow to control the variance of these perturbations when there is no mean deviation.

Proposition 3. *If assumptions* **(A7)** *or* **(A8)** *holds, then, under* H_1^*, $T(n)$ *converges in probability towards* $+\infty$. *Moreover, under* **(A7)** S_n *converges in probability towards 1, and under* **(A8)** S_n *converges in probability towards 2.*

6. Choice of the reference measure and test construction

In order to run our test, we have to select now a reference measure ν and an *ad. hoc.* orthogonal family $\mathcal{Q} = \{Q_k,\ k \in \mathbb{N}\}$. The choice of the ν depends clearly of the support of X_1. For a compact support, one can choose a uniform distribution for ν and their associated Legendre polynomials. Since our numerical studies are dedicated to the Gaussian case, we illustrate here the choice of ν corresponding to two measures on the real line: the Gaussian and the Lebesgue one. The verification of conditions **(A2–3)** for these two measures is relegated in the supplementary material file.

Gaussian reference measure. In practice, in the present paper, we chose for ν the standard normal distribution for testing the Gaussianity. This choice is adapted to any distribution having support on the real line. The set $\mathcal{Q}$ is constructed from the $f_{(0,1)}$-orthogonal Hermite polynomials defined for all $k \geq 0$ by:

$$H_k(x) = k! \sum_{m=0}^{\lfloor k/2 \rfloor} \frac{(-1)^m x^{k-2m}}{m!(k-2m)!2^m}, \quad x \in \mathbb{R}. \tag{6.1}$$

We have $\|H_k\|^2 = k!$ and, for illustration purpose, the six first polynomials are:

$$H_0 = 1,\ H_1(x) = x,\ H_2(x) = x^2 - 1,\ H_3(x) = x^3 - 3x,$$
$$H_4(x) = x^4 - 6x^2 + 3,\ H_5(x) = x^5 - 10x^3 + 15x.$$

Lemma 1. *Let H_k be defined by (6.1) and let $Q_k(x) = H_k(x)$, for all $x \in \mathbb{R}$. Assume that we want to test $H_0 : f \in \mathcal{G}$, where $\mathcal{G}$ is the set of Gaussian densities. Then conditions* **(A2–3)** *are satisfied.*

Remark 5. *Lemma 1 can be extended to non Gaussian null distribution f with known moments as discussed in Remark 1 in supplementary file.*

Lemma 2. *Let H_k be defined by (6.1) and let $Q_k(x) = H_k(x)$, for all $x \in \mathbb{R}$. Then condition* **(A6)** *is satisfied.*

Lebesgue reference measure. Another simple ν reference measure could the Lebesgue measure over $\mathbb{R}$. In that case, we would rather consider the set of orthogonal Hermite functions defined by:

$$\mathcal{H}_k(x) = h_k(x) \exp(-x^2/2), \quad x \in \mathbb{R}, \tag{6.2}$$

where $h_k(x) = 2^{k/2} H_k(\sqrt{2}x)$, with H_k defined in (6.1). In addition we have $\|\mathcal{H}_k\|^2 = k! 2^k$.

Lemma 3. *Let $\mathcal{H}_k$ be defined by (6.2) and let $Q_k(x) = \mathcal{H}_k(x)$, for all $x \in \mathbb{R}$. Then conditions* **(A2–3)** *are satisfied.*

Lemma 4. *Let $\mathcal{H}_k$ be defined by (6.2) and let $Q_k(x) = \mathcal{H}_k(x)$, for all $x \in \mathbb{R}$. Then condition* **(A6)** *is satisfied.*

Test construction. The computation of the test statistic $T(n) = T_{S_n,n}$, see expressions (2.6) and (2.8), is grounded on the computation of the $\alpha_i(\mu, s)$'s quantities. We detail here the expression of $R_{1,n}$ and $\mathrm{var}(R_{1,n})$ when the reference measure is Gaussian associated with the Hermite polynomials. To overcome the complex dependence between the estimators $a_{1,n}$, $\bar{p}_n$, $\bar{\mu}_n$ and $\bar{s}_n$, we split the sample into four independent sub-samples of size n_1, n_2, n_3, n_4, with $n_1+n_2+n_3+n_4 = n$. We use the first sample to estimate a_1, the second sample to estimate p, the third one to estimate μ, and the last one to estimate s. We get $\alpha_1(\mu, s) = \mu$ and $\alpha_{1,n} = \bar{\mu}_n$ which makes

$$\begin{aligned} R_{1,n} &= n_1^{-1}\sum_{i=1}^{n_1} X_i - \bar{p}_{n_2}\bar{\mu}_{n_3}, \quad \text{and} \\ \mathrm{var}(R_{1,n}) &= \mathrm{var}(X)/n_1 + \mathrm{var}(\bar{p}_{n_2})\mathrm{var}(\bar{\mu}_{n_3}) + \mathrm{var}(\bar{p}_{n_2})\mathbb{E}(\bar{\mu}_{n_3})^2 + \mathbb{E}(\bar{p}_{n_2})^2\mathrm{var}(\bar{\mu}_{n_3}). \end{aligned}$$

We propose a consistent estimator of $\mathrm{var}(R_{1,n})$:

$$V_{1,n} = S^2_{X,n_1} + v_{p,n_2}v_{\mu,n_3} + \bar{\mu}^2_{n_3}v_{p,n_2} + \bar{p}^2_{n_2}v_{\mu,n_3},$$

where S^2_{X,n_1} denotes the empirical variance based on $(X_1, \ldots, X_{n_1})$, and v_{p,n_2}, respectively v_{μ,n_3}, denotes the consistent estimator of $\mathrm{var}(\bar{p}_{n_2})$, respectively $\mathrm{var}(\bar{\mu}_{n_3})$), obtained from Bordes and Vandekerkhove (2010, p. 40). The computation of the test statistic first requires the choice of $d(n)$, $e(n)$ and $s(n)$. A previous study showed us that the empirical levels and powers were overall weakly sensitive to $d(n)$ for $d(n)$ large enough. From that preliminary study we decided to set $d(n)$ equal to 10. The trimming $e(n)$ is calibrated equal to $(\log(n))^{-1}$. The normalization $s(n) = n^{\alpha-1}$ is setup close enough to $n^{-1/2}$, with α equal to $2/5$, which seemed to provide good empirical levels.

Secondly, since the probability density functions considered in our set of simulation are $\mathbb{R}$-supported we use the standard Gaussian distribution for ν and its associated Hermite polynomials for $\mathcal{Q}$. All our simulations are based on 200 repetitions. Let us remind briefly that the empirical level is defined as the percentage of rejections under the null hypothesis and that the empirical power is the percentage of rejections under the alternative. Finally the asymptotic level is standardly fixed to 5%.

7. Semiparametric and maximum likelihood approaches comparison

In our testing procedure we estimate p, μ by the semiparametric estimators proposed in Bordes and Vandekerkhove (2010) instead of the maximum likelihood estimators. In the same way our estimation of θ, see expression (2.5), is H_0-free contrary to what would happen when using the maximum likelihood technique. Both approaches are asymptotically equivalent under the null hypothesis, see remark 4, and all the simulations we did shown very similar empirical levels when comparing the semiparametric and maximum likelihood approaches under null models. However, under certain types of alternatives, the maximum

likelihood approach can lead to very unexpected empirical powers. These behaviors are due to compensation phenomenon in models close, for example, to the non-identifiable one described in Section 4. To illustrate clearly this point we detail here the Gaussianity test in these cases. Write

$$g(x) = (1-p)f_{(0,1)}(x) + ph_{a,s}(x-\mu), \quad x \in \mathbb{R}, \tag{7.1}$$

where $h_{a,s}(x) = (f_{(0,s)}(x-a) + f_{(0,s)}(x+a))/2$, $a \neq 0$, $f_{(0,s)}$ being the Gaussian density, centered, with variance s. We notice that (7.1) turns to satisfy, when $\mu = a$ and $s = 1$, the following rewriting

$$g(x) = (1-\frac{p}{2})f_{(0,1)}(x) + \frac{p}{2}f_{(0,1)}(x-2\mu), \quad x \in \mathbb{R}. \tag{7.2}$$

In this case there are two different parametrizations for (7.1): one that we call the *null parametrization*, coinciding with H_0 with null parameters $p_0 = p/2, \mu_0 = 2\mu$ and $s_0 = 1$, see the right hand side of (7.2). The other one is called the *alternative parametrization*, coinciding with H_1 with $p_1 = p, \mu_1 = \mu$ and $s_1 = \mu^2+1$, see the right hand side of (7.1). By construction the maximum likelihood estimator will favor the null parameters. We study now this phenomenon through a set of simulations where the parameters are $\mu = 4$, $s = 1$ and $p = 0.4$. For comparison, we used the same initial values for the both semiparametric and maximum likelihood algorithms, namely $(p, \mu, s) = (0.3, 6, 8.5)$, which is exactly between the null parametrization $(p, \mu, s) = (0.2, 8, 1)$, and the alternative parametrization $(p, \mu, s) = (0.4, 4, 17)$. It is of interest to study now the behavior of the semiparametric and maximum likelihood testing methods when the true model deviates smoothly from the null hypothesis in two ways: i) the unknown component is a $h_{a,1}$ with $\mu \neq a$, *i.e*

$$\begin{aligned} g(x) &= (1-p)f_{(0,1)}(x) + p\underbrace{\left(\frac{1}{2}f_{(0,1)}(x-a-\mu) + \frac{1}{2}f_{(0,1)}(x+a-\mu)\right)}_{\mu\text{-symmetric mixture detected by the semiparametric method}} \\ &= \left((1-p)f_{(0,1)}(x) + \frac{p}{2}f_{(0,1)}(x+a-\mu))\right) + \frac{p}{2}f_{(0,1)}(x-a-\mu) \\ &\approx \left(1-\frac{p}{2}\right)f_{(0,1)}(x) + \frac{p}{2}\underbrace{f_{(0,1)}(x-a-\mu))}_{(a+\mu)\text{-centered Gaussian attracting the maximum likelihood method}}, \quad \text{when } \mu \to a, \end{aligned}$$

this case will be called the *mean deviation trap*, and ii) the unknown component is a $h_{a,s}$ with $\mu = a$ but $s \neq 1$, *i.e.*

$$\begin{aligned} g(x) &= (1-p)f_{(0,1)}(x) + p\underbrace{\left(\frac{1}{2}f_{(0,s)}(x-2\mu) + \frac{1}{2}f_{(0,s)}(x)\right)}_{\mu\text{-symmetric mixture detected by the semiparametric method}} \\ &= \left((1-p)f_{(0,1)}(x) + \frac{p}{2}f_{(0,s)}(x))\right) + \frac{p}{2}f_{(0,1)}(x-2\mu) \\ &\approx \left(1-\frac{p}{2}\right)f_{(0,1)}(x)) + \frac{p}{2}\underbrace{f_{(0,s)}(x-2\mu))}_{(2\mu)\text{-centered Gaussian attracting the maximum likelihood method}}, \quad \text{when } s \to 1 \end{aligned}$$

this case will be called the *variance deviation trap*.
It is very important to point out now that the above phenomenons illustrate the risk of considering only one single Gaussian component ($q = 1$) in the generic mixture model defining f in McLachlan *et al.* (2006, Section 6) when actually two Gaussian components ($q = 2$) would be necessary to accurately fit the model.

Mean deviation trap. We consider deviations from the null model obtained by considering $\mu = 3, 2, 1$ and $s = 1$. Fig. 1 shows the g probability density function under these respective alternatives. It can be observed that, if we try to visually detect a mixture of two Gaussian distributions, the probability density function of the left-side component moves clearly aside the Gaussian distribution family as μ moves largely away from $a = 4$, *i.e.* when $\mu = 1$, but we bet that many practitionners would probably vote "intuitively" for a mixture of two Gaussian distributions when $\mu = 3$ or 2. Fig. 1 in supplementary file

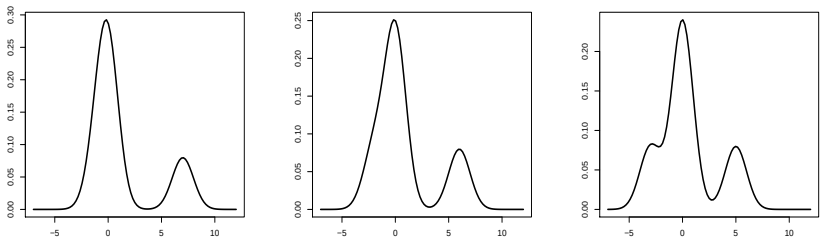

Fig 1: The probability density function g in model (7.1) when $a = 4$, $s = 1$, and $\mu = 3, 2, 1$.

illustrates the difficulty of the maximum likelihood estimator to recognize the alternative model when the mean deviation is not distant enough (here $\mu = 3$ and $a = 4$). Based on a run of 200 repetitions, it is shown that the maximum likelihood estimation is trapped at the null parametrization which namely

is $(p, \mu, s) = (0.2, 7, 1)$ when on the opposite, the semiparametric estimation detects the correct $(p, \mu, s) = (0.4, 3, 17)$ alternative parametrization. In Fig. 2 we display respectively the empirical power of our testing procedure based on the maximum likelihood and the semiparametric approach for $\mu = 3, 2, 1$, $a = 4$, $s = 1$, and for $n = 1000, 2000, 5000$. As expected the maximum likelihood approach barely detects the alternative for small values of n when its semiparametric counterpart surpasses it with up to 10 times more correct decision results. The reason of this lack of power is due to the fact that our test focuses more on the moments of the second components than those of the first one and, as seen in Fig. 1, the second components looks pretty much Gaussian even for $\mu = 1$.

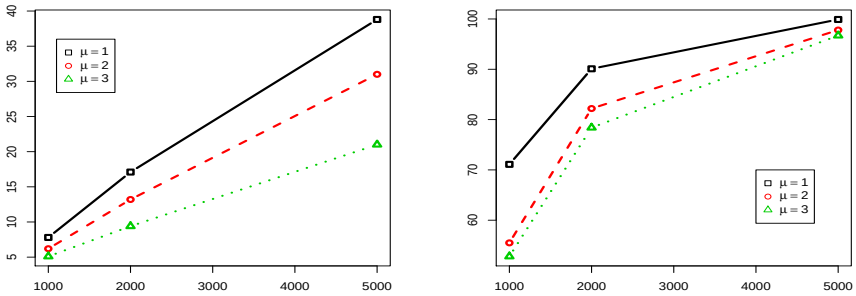

Fig 2: Empirical powers obtained with the maximum likelihood approach (left) and semiparametric approach (right) under the trap effect for $\mu = 3, 2, 1$ and $a = 4$

Variance deviation trap. We consider the variance deviations $s = 2, 3, 4$, fixing $\mu = a = 4$. Fig. 3 shows the g probability density function under these alternatives. Empirical powers are displayed in Fig. 4. We can observe that both

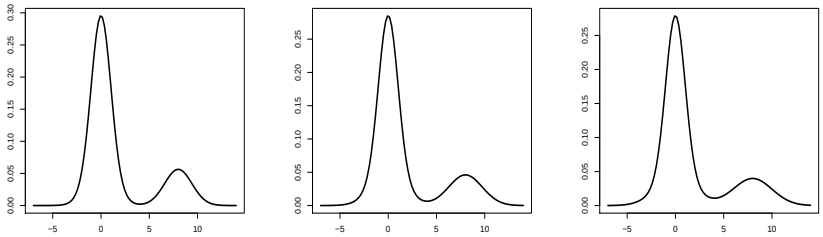

Fig 3: The probability density function g in model (7.1) with $\mu = a = 4$ and $s = 2, 3, 4$.

powers associated with the maximum likelihood and semiparametric approach increase according to the variance deviation but it is worth to notice that the detection based on the maximum likelihood approach is again very poor compared to the semiparametric approach. As a conclusion, this set of numerical exper-

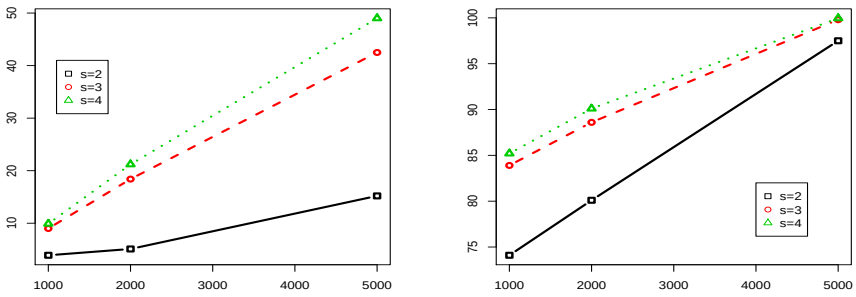

Fig 4: Empirical powers obtained with the maximum likelihood approach (left) and semiparametric approach (right) under the variance deviation trap effetct for $s = 2, 3, 4 \neq 1$, with $\mu = a = 4$.

iments shows the clear interest, in terms of testing power, of considering the semiparametric versus the maximum likelihood approach especially in a close neighborhood of non-identifiable type (1.1) Gaussian models.

8. Simulations: empirical levels and powers

8.1. Empirical levels

McLachlan *et al.* (2006) considered the two-component Gaussian version of the mixture model (1.1) through three datasets arising from the bioinformatics literature: the breast cancer data, with $n = 3226$, the colon cancer data, with $n = 2000$, and the HIV data, with $n = 7568$. The estimation of their associated parameters are respectively: $(\widehat{p}_n, \widehat{\mu}_n, \widehat{s}_n) = (0.36, 1.52, 0.99)$, $(0.58, 1.61, 2.08)$, and $(0.98, -0.15, 0.79)$. To make sure that our methodology will have reliable behaviors when applied on this collection of datasets, we investigate the empirical levels of our testing procedure across parameter values such as $n \in \{2000, 3000, 7500\}$ and $(p, \mu, s) = (1/3, 1.5, 1)$, $(0.5, 1.5, 2)$ and $(0.98, -0.15, 0.8)$ which are values in the range of the above targeted applications. For this purpose, for each value of n, p, μ and s, we compute the test statistic $T(n)$ based on the sample and compare it to the 5%-critical value of its approximated distribution under H_0 ($\chi^2(1)$ according to Corollary 3). Note that, for numerical simplicity, we initialize our parameter estimation step at the true value of the Euclidean parameter. The collection of empirical levels obtained for this set of

simulated examples is reported in Fig. 2 of the supplementary file. It appears that a significant number of observations is needed to get close to the theoretical level. This drawback can be balanced by the fact that today, as mentioned in the Introduction, genomic datasets usually contain thousands of genes which makes our methodology in practice suitable for a wide class of standard (from the sample size view point) microarray analysis problems.

8.2. Empirical powers

In this section we consider the Gaussian testing problem (1.2) with $\mathcal{F} = \mathcal{G}$ where $\mathcal{G} = \left\{f_{(\mu,s)}; (\mu, s) \in \Lambda \subset \mathbb{R} \times \mathbb{R}^{+*}\right\}$ denotes the set of Gaussian densities with mean μ and variance s, compared to Student and Laplace alternatives. First a 1-shifted Student distribution $t(3)$, having a shape far enough from the Gaussian distribution, with a shift $\mu = 1$. Second a shifted Student $t(10)$, again with a shift equal to 1, but having a shape closer to the null Gaussian distribution. Third a Laplace distribution $\mathcal{L}(1, 1)$ with mean 1 and variance 2. The last alternative is a Laplace $\mathcal{L}(1, 2)$ with mean 1 and variance 8. The empirical powers for Student and Laplace alternatives are respectively summarized in Fig. 5 and 6.

As expected, when comparing pairwise the Student alternatives, the power is greater for the $t(3)$ distribution compared to the $t(10)$ distribution. The $t(3)$ is very clearly detected by the test since the detection level is greater than 80% for all the cases and even close to 100% for $n = 7000$. Now, similarly to the mean and variance deviation trap setups investigated in Section 7, we can observe that the power is greater as p increases, which practically means that the Student component is enhanced in the model (remind that our test procedure is focused on the 2nd-component moments analysis) . We display the mixture densities corresponding to this set of alternatives in Fig. 3 of the supplementary file. For the first Student alternative, comparing $p = 1/2$ and $p = 0.98$, we can observe that a serious jump happens in terms of dissimilarity between the alternative model and the *best fitted* (same mean and variance) Gaussian null-model. For $p = 0.98$, the Student distribution strongly prevails and the test is automatically empowered. The second alternative is also detected, but with a lower power, let say between 40 % and 90%, due to the proximity of the Student $t(10)$ with the Gaussian $\mathcal{N}(0, 1)$.
In Fig. 3 of the supplementary file we can see how close the null distribution and the $t(10)$ alternative are, especially for $p = 1/3$ and $p = 1/2$, and visually evaluate how challenging these testing problems really are.

The empirical powers for Laplace alternatives are given in Fig. 6. The power is larger with the alternative $\mathcal{L}(1, 2)$ than with the alternative $\mathcal{L}(1, 1)$. Indeed the $\mathcal{L}(1, 2)$ distribution has a stronger shape departure from the Gaussian than the $\mathcal{L}(1, 1)$, and the associated mixture densities inherit these characteristics as we can see in Fig. 3 of the supplementary file. These alternatives are globally very well detected by our method and the power increases strongly when p gets closer to 1 (see Fig. 6 curve in green).

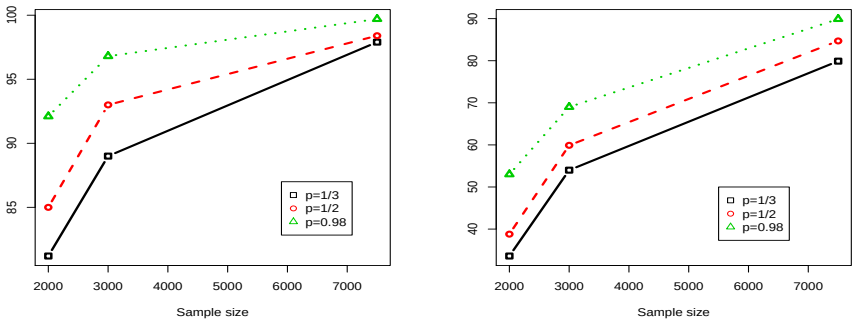

Fig 5: Empirical powers when the alternative is a shifted Student $t(3)$ (left) and a shifted Student $t(10)$ for parameter values $p = 1/3$ ($\square$), $p = 1/2$ ($\circ$) and $p = 0.98$ ($\triangle$) with sample sizes $n = 2000, 3000, 7500$.

9. Real datasets

Microarray data. We consider 3 datasets arising from the bioinformatics literature and studied in McLachlan *et al.* (2006). Fig. 7 shows the non parametric kernel estimations of their probability density functions. Each of them deals with genes expressions modeled by the two-component mixture model (1.1) in which f was arbitrarily, for simplicity matters, considered as Gaussian (without any theoretical justification). The goal of this section is to answer if the classical Gaussian assumption was a posteriori correct or not.

Breast cancer data. We consider the breast cancer data studied in Hedenfalk *et al.* (2001). It consists in $n = 3226$ gene expressions in breast cancer tissues from women with BRCA1 or BRCA2 gene mutations. The maximum likelihood parameter estimations under the Gaussian null model are $\widehat{p}_n = 0.36$, $\widehat{\mu}_n = 1.53$, $\widehat{s}_n = 0.98$. By the semiparametric method we obtain $\bar{p}_n = 0.41$, $\bar{\mu}_n = 1.35$ and $\bar{s} = 1.31$. It can be noticed here that nonparametric and maximum likelihood estimators give pretty similar results here which may corroborate the null hypothesis. Our test procedure provides a p-value equal to 0.82, with $S_n = 1$. As a consequence the normality of the second mixture component under H_0 cannot be rejected.

Colon cancer data. We consider the colon cancer data analysed in Alon *et al.* (1999). The samples comes from colon cancer tissues and normal colon tissues. It contains $n = 2000$ expressions of genes. The maximum likelihood estimations of the parameters are $\widehat{p}_n = 0.58$, $\widehat{\mu}_n = 1.61$, $\widehat{s}_n = 2.08$; The semiparametric method provides $\bar{p}_n = 0.72$, $\bar{\mu}_n = 1.28$ and $\bar{s} = 2.33$. By using our testing procedure we obtain a p-value less than 10^{-8} with $S_n = 4$. Here we clearly reject the normality under H_0. The rejection of the Gaussian mixture can be explained

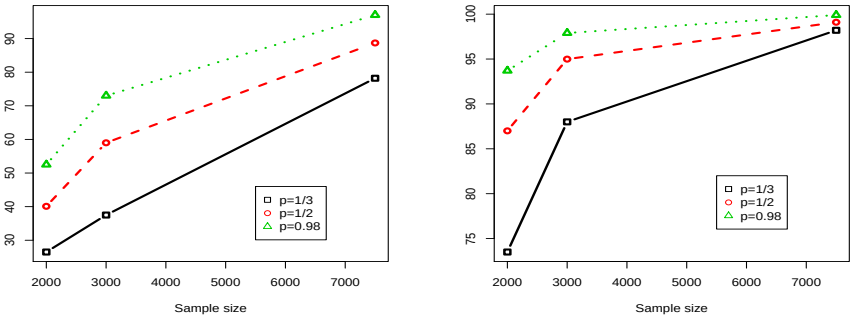

Fig 6: Empirical powers with alternative a Laplace $\mathcal{L}(1,1)$ (left) and a Laplace $S\mathcal{L}(1,2)$ (right) for parameter values $p = 1/3$ ($\square$), $p = 1/2$ ($\circ$) and $p = 0.98$ ($\triangle$) with sample sizes $n = 2000, 3000, 7500$.

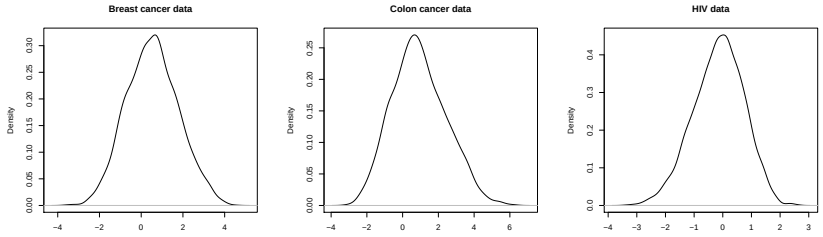

Fig 7: Respectively the kernel density estimators of the breast data, colon data and HIV data distributions.

here by the fact that the nonparametric and the maximum likelihood estimators lead to notably different values especially on p.

HIV data. We consider the HIV dataset of vant' Wout *et al.* (2003). It contains expression levels of $n = 7680$ genes in CD4-T-cell lines, after infection with the HIV-1 virus. The maximum likelihood estimations of the parameters are $\widehat{p}_n = 0.98$, $\widehat{\mu}_n = -0.15$, $\widehat{s}_n = 0.79$. The semiparametric method provides $\bar{p}_n = 0.99$, $\bar{\mu}_n = 0.20$ and $\bar{s} = 0.80$. The p-value given by our testing procedure is equal to 0.64, associated with the decision $S_n = 1$. As a consequence the normality under H_0 cannot be rejected despite the fact that the maximum likelihood and semiparametric estimations of μ are quite different but both close to 0, meaning a strong overlap of the mixed distributions (see the almost symmetry of the third probability density function in Fig. 7).

Galaxy data. We consider here the Carina dataset, see Walker *et al.* (2007), studied previously in Patra and Sen (2016). Carina is a low luminosity galaxy companion of the Milky Way. The data collects $n = 1266$ measurements of the radial velocity of stars in Carina. This is a contamination model in the sense that the measurements of stars in the Milky Way are mixed with some of Carina (overlapping). The Milky Way is largely observed, see Robin *et al.* (2003). Figure 8 shows the density f_0 of the radial velocity of Milky Way, estimated over $n' = 170,601$ observations. This density is clearly not zero-symmetric but in such a case it is enough to refer to the tail-oriented set of identifiability conditions of Proposition 3 i) in Bordes et al. (2006) to make the semiparametric estimation method still valid. Note also that the asymptotic results of Bordes and Vandekerkhove (2010) still hold if the cumulative distribution function F_0 is replaced by a smooth empirical estimate $\widetilde{F}_{0,n'}$ based on a $n' = \varphi(n)$ sized training data provided with $n/n' \to 0$ as $n \to +\infty$. Unfortunately the study of the maximum likelihood estimate, see Section 5 of the supplementary file, cannot be generalized straightforwardly since the non-parametric estimation of the Kullback distance, obtained by replacing f_0 by a kernel density estimate $\widehat{f}_{0,n}$ in the log-likelihood, is known to be very a delicate problem, see Berrett *et al.* (2018) and references therein. Though, the fact that the unknown component of g under H_0 is supposed to have a parametric form should definitely help to control some technical tail issues specific to the Kullback estimation. We obtained for p and μ, respectively the proportion and the mean of the Carina radial velocity, the following estimations:

$$\bar{p}_n = 0.361 \quad \text{and} \quad \bar{\mu}_n = 222.60.$$

In their study, Patra and Sen (2016) obtained very similar values: $\widetilde{p} = 0.323$ and $\widetilde{\mu} = 222.9$. However, the estimation of the variance s appears to be highly sensitive to the estimation of p. Using the plug-in estimator given by (2.5) we get $\bar{s}_n = 453.93$. Note that the estimation given in Patra and Sen (2016) was $\widetilde{s}_n = 56.4$ which looks far from the expected value given the data. To illustrate this remark, we compare in Fig. 8 the kernel density estimate of the observed data with the probability density of model (1.1), obtained by replacing (p, μ, s) by our estimates $(\bar{p}_n, \bar{\mu}_n, \bar{s}_n)$ and the Patra and Sen (2016)'s estimates $(\widetilde{p}_n, \widetilde{\mu}_n, \widetilde{s}_n)$. We can observe that our estimation provides an excellent fitting when the variance estimated by Patra and Sen (2016) appears to be way too small. Our test procedure yields a p-value equal to 0.75 with a test statistic $T_{S_n,n} = T_1 = 0.097$. As a consquence, there is no evidence here to reject the normality of the Carina radial velocity.

10. Discussion and perspectives

In this paper we proposed an H_0-free testing procedure to deal with the delicate problem of the contamination model parametrization. In our numerical study we focused our attention on the Gaussianity testing problem however it is very important to remind that our asymptotic results can be generalized to

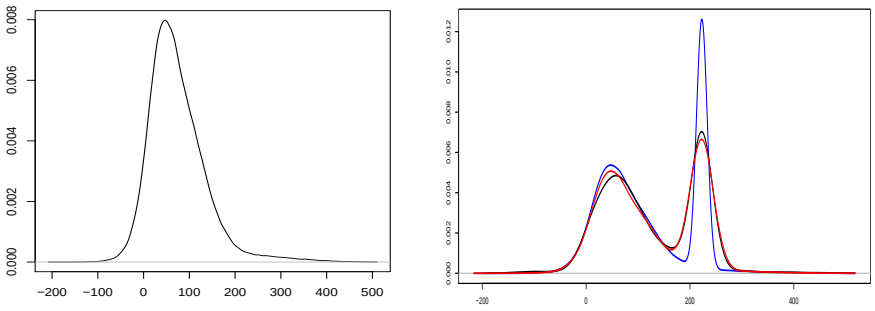

Fig 8: Left side: estimated density of the Milky Way Radial velocities. Right side: in black, the plot of the Carina dataset nonparametric density estimate. In red, resp. in blue, the plot of the model (1.1) probability density function under $f = f_{(\mu,s)}$ and obtained by plugging $(\bar{p}_n, \bar{\mu}_n, \bar{s}_n)$, resp. $(\widetilde{p}_n, \widetilde{\mu}_n, \widetilde{s}_n)$, into (p, μ, s).

any suitable distribution (possibly non-symmetric). Indeed, if the unknown distribution of model (1.1) is embedded in a nonparametric envelop $\mathcal{S}$ provided with identifiability constraints and if there exists a corresponding semiparametric $\sqrt{n}$-consistent method, then the asymptotic results in Sections 3-4 extends straightforwardly. For this latter case, we recommend the recent work by Al Mohammad and Boumahdaf (2018) who consider in model (1.1) an unknown component defined through linear constraints. In their paper, the authors derive an original consistent and asymptotically normally distributed semiparametric estimation method with asymptotic closed form variance expressions. Indeed, when considering null assumptions different from the Gaussian case, basically only the shape parameter estimation, usually deduced from moment equations, and the choice of the orthogonal basis described in Section 2 could possibly change, depending on the support of the tested distribution. Wavelet functions and Laguerre polynomials could respectively be used for probability density functions on the whole, respectively positive, real line, when Legendre, or cosine bases could be used for densities with compact support. Also, with a slight adaptation of our work, we could definitely test the unknown component of the contamination model considered in the recent work by Ma and Yao (2015) where the first component density is only supposed to belong to a parametric family (the first component is not entirely known anymore). For each case, the use of the maximum likelihood or semiparametric approach could be again discussed. On the other hand, as it has been demonstrated in Section 7, see Figs. 2 and 4, the semiparametric testing approach shows better power performances than the maximum likelihood version especially in the neighborhood of the mean and variance deviation trap situations (up to 10 times more efficient for small sample sizes). We also proposed in Section 5 a vanishing convolution-class of nonparametric contiguous alternatives and studied theoretically their detectability

under certain convergence rate conditions. In a futur work it would be very interesting to address the contiguous detection problem associated with the mean and variance deviation trap setups. This would namely consist in looking at the asymptotic behavior of our test when replacing respectively the parameters μ and s in the mean and variance deviation trap setups by sequences μ_n and s_n converging respectively towards a and 1 as n goes to infinity. The major technical difficulty here is that we are not able to establish yet optimal bounds of convergence for the semiparametric Euclidean estimator associated with a triangular array driven by the above asymptotic parametrization, see Remark 4 in the supplementary material file. Future work is also to consider a K-sample extension, $K \geq 2$, in the spirit of Wylupeck (2010), Ghattas *et al.* (2011), or more recently Doukhan *et al.* (2015). More precisely, we could test the equality of K unknown components through K observed mixture models.

Acknowlegement. The authors acknowledge the Office for Science and Technology of the Embassy of France in the United States, especially its antenna in Atlanta, for its valuable support to this work.

11. Appendix: proofs of the main results

Theorem 2. Let us prove that $\mathbb{P}(S_n \geq 2)$ vanishes as $n \to +\infty$. By definition of S_n in (2.8) and $\widehat{D}_{k,n}[\cdot]$ in (2.7) we have for all $\lambda \in]0, 1/2[$:

$$\begin{aligned}
&\mathbb{P}(S_n \geq 2)\\
&= \mathbb{P}\left(\text{there exists } k \in \{2,\dots,d(n)\} :\ n^\lambda U_{k,n}^\top \widehat{D}_{k,n}^{-1} U_{k,n} - k\log(n) \geq n^\lambda U_{1,n}^\top \widehat{D}_{1,n}^{-1} U_{1,n} - \log(n)\right)\\
&\leq \mathbb{P}\left(\text{there exists } k \in \{2,\dots,d(n)\} :\ n^\lambda U_{k,n}^\top \widehat{D}_{k,n}^{-1} U_{k,n} \geq (k-1)\log(n)\right)\\
&\leq \mathbb{P}\left(\text{there exists } k \in \{2,\dots,d(n)\} :\ \sum_{j=2}^{k} n^\lambda (R_{j,n})^2 \geq (k-1)\log(n)e(n)\right)\\
&\leq \mathbb{P}\left(\text{there exists } (j,k) \text{ with } 2 \leq j \leq k \leq d(n) :\ n^\lambda (R_{j,n})^2 \geq \log(n)e(n)\right)\\
&\leq \mathbb{P}\left(\sum_{j=2}^{d(n)} n^\lambda (R_{j,n})^2 \geq \log(n)e(n)\right). \qquad (11.1)
\end{aligned}$$

It is important for us to keep the summation term up to $d(n)$ in the left hand side of the above inequality-type event in order to straightforwardly use the almost sure rate of convergence of the semiparametric Euclidean parameters, see (11.5)–(11.6). We decompose $R_{k,n}$ as follows:

$$R_{k,n} = (a_{k,n} - \mathbb{E}(a_{k,n})) - (\bar{p}_n \alpha_{k,n} - p_0 \alpha_k(\mu_0, \theta_0)), \quad 1 \leq k \leq d(n). \qquad (11.2)$$

By using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, for all $(a,b) \in \mathbb{R}^2$, we get

$$\mathbb{P}\left(\sum_{k=2}^{d(n)} n^{\lambda}(R_{k,n})^2 \geq \log(n)e(n)\right) \leq \mathbb{P}\left(\sum_{k=2}^{d(n)}(a_{k,n}-\mathbb{E}(a_{k,n}))^2 \geq \frac{\log(n)e(n)}{4n^{\lambda}}\right) + \mathbb{P}\left(\sum_{k=2}^{d(n)}(\bar{p}_n\alpha_{k,n}-p_0\alpha_k(\mu_0,\theta_0))^2 \geq \frac{\log(n)e(n)}{4n^{\lambda}}\right). \tag{11.3}$$

We study now all the above quantities separately. By the Markov inequality, we first have

$$\begin{aligned}\mathbb{P}\left(\sum_{k=2}^{d(n)}(a_{k,n}-\mathbb{E}(a_{k,n}))^2 \geq \frac{\log(n)e(n)}{4n^{\lambda}}\right) &\leq \frac{4n^{\lambda}}{\log(n)e(n)}\sum_{k=2}^{d(n)}\mathbb{E}\left((a_{k,n}-\mathbb{E}(a_{k,n}))^2\right)\\ &= \frac{4n^{\lambda}}{\log(n)e(n)}\sum_{k=2}^{d(n)}\frac{1}{n}\mathrm{var}\left(\frac{Q_k(X_1)}{q_k^2}\right)\\ &\leq \frac{4d(n)}{n^{1-\lambda}\log(n)e(n)}M_3, \end{aligned}\tag{11.4}$$

where the right hand side term goes to zero as $n \to +\infty$ since $d(n)/\log(n)e(n) = O(1)$ according to **(A1)** and (2.9).
Secondly, by decomposing $\bar{p}_n\alpha_{k,n} - p_0\alpha_k(\mu_0,\theta_0) = (\bar{p}_n - p_0)\alpha_{k,n} + p_0(\alpha_{k,n} - \alpha_k(\mu_0,\theta_0))$, we obtain the following majorization

$$\mathbb{P}\left(\sum_{k=2}^{d(n)}(\bar{p}_n\alpha_{k,n}-p_0\alpha_k(\mu_0,\theta_0))^2 \geq \frac{\log(n)e(n)}{4n^{\lambda}}\right) \leq \mathbb{P}\left(\sum_{k=2}^{d(n)}(\alpha_{k,n})^2(\bar{p}_n-p_0)^2 \geq \frac{\log(n)e(n)}{8n^{\lambda}}\right) + \mathbb{P}\left(\sum_{k=2}^{d(n)}p_0^2(\alpha_{k,n}-\alpha_k(\theta_0,\mu_0))^2 \geq \frac{\log(n)e(n)}{8n^{\lambda}}\right).$$

Since the $\alpha_{k,n}$'s are bounded by M_1 according to **(A2)**, we have

$$\mathbb{P}\left(\sum_{k=2}^{d(n)}\alpha_{k,n}^2(\bar{p}_n-p_0)^2 \geq \frac{\log(n)e(n)}{8n^{\lambda}}\right) \leq \mathbb{P}\left((\bar{p}_n-p_0)^2 \geq \frac{\log(n)e(n)}{8n^{\lambda}M_1 d(n)}\right), \tag{11.5}$$

where the last right hand side term goes to zero as $n \to +\infty$ since $\lambda \in (0,1/2)$ and $|\bar{p}_n - p_0|^2 = o_{a.s.}(n^{-1/2+\alpha})$ for all $\alpha > 0$, by Bordes and Vandekerkhove (2010). By denoting $\rho_0 = (\mu_0,\theta_0)$ and $\bar{\rho}_n = (\bar{\mu}_n,\bar{\theta}_n)$, we also have $\|\bar{\rho}_n - \rho_0\|^2 = o_{a.s.}(n^{-1/2+\alpha})$, for all $\alpha > 0$. Since the $\dot{\alpha}_{k,n}$'s are bounded by $M2$ according to **(A2)**, using the *mean value* theorem we obtain:

$$\mathbb{P}\left(\sum_{k=2}^{d(n)}(\alpha_{k,n}-\alpha_k(\mu_0,\theta_0))^2 \geq \frac{\log(n)e(n)}{8n^{\lambda}}\right) \leq \mathbb{P}\left(\|\bar{\rho}_n-\rho_0\|^2 \geq \frac{\log(n)e(n)}{8n^{\lambda}M_2^2 d(n)}\right), \tag{11.6}$$

which last term goes to zero as $n \to +\infty$. Hence from (11.1) and the controls in probability (11.3–11.6), we obtain that $\mathbb{P}(S_n \geq 2) \to 0$ as $n \to +\infty$. □

Corollary 3. From Theorem 2, $T_{S_n,n}$ has the same limiting distribution as $T_{1,n} = nR_{1,n}^2/V_{1,n}$. Since the estimators $\bar{\theta}_n$ and $\bar{\mu}_n$ are independent and asymptotically Normally distributed towards the true values θ_0 and μ_0 we get, by using the delta method, the following convergence in distribution:

$$\sqrt{n}\alpha_1(\bar{\mu}_n, \bar{\theta}_n) \longrightarrow \mathcal{N}(\alpha_1(\mu_0, \theta_0), D(\mu_0, \theta_0)VD(\mu_0, \theta_0), \quad \text{as } n \to +\infty,$$

where $D(\cdot, \cdot)$ is the gradient $\dot{\alpha}_1(\cdot, \cdot)$, and where V is the asymptotic variance of $(\sqrt{n}\bar{\mu}_n, \sqrt{n}\bar{\theta}_n)$. Combining this convergence in law with the following convergence in probability:

$$V_{1,n} \longrightarrow \text{var}(R_{1,n}) \quad \text{and} \quad \bar{p}_n \longrightarrow p_0, \quad \text{as } n \to +\infty,$$

along with the independence and the asymptotic normality of the first estimated coefficient $a_{1,n} = \sum_{i=1}^{n} Q_1(X_i)/nq_1^2$, we get, by using the Slutsky's Theorem, the following limiting distribution:

$$\sqrt{n}\frac{R_{1,n}}{\sqrt{V_{1,n}}} = \sqrt{\frac{n}{V_{1,n}}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Q_1(X_i)}{q_1^2} - \bar{p}_n\bar{\mu}_n\right) \longrightarrow \mathcal{N}(0,1), \quad \text{as } n \to +\infty,$$

which concludes the proof. □

Proposition 1. The advantage of considering the semiparametric approach in Bordes and Vandekerkhove (2010) versus the maximum likelihood method is that under H_1 we keep the following consistency results in probability:

$$\bar{\vartheta}_n = (\bar{p}_n, \bar{\mu}_n) \longrightarrow (p_0, \mu_0), \ \bar{\theta}_n \longrightarrow \theta_0, \ R_i \longrightarrow r_i = \mathbb{E}(Q_i(X)/q_i^2) - p_0\alpha_i(\mu_0, \theta_0),$$

as $n \to +\infty$, for $i \geq 1$, along with their associated asymptotic normality. As a consequence, by using the Slutsky's Theorem, the terms $\sqrt{n}(R_{i,n} - r_i)/\sqrt{\widehat{D}_{k,n}[i]}$, $1 \leq i \leq k$, are asymptotically normally distributed since $\widehat{D}_{k,n}[i]$ is a weakly consistent estimator of $\text{var}(R_i)$. Now, Clearly by (1.1) (with $b_i = 0$), $\mathbb{E}(Q_i(X)) = p_0\mathbb{E}(Q_i(Y))$, where Y is a f-distributed random variable. Then we have the following equivalence

$$r_i = 0, \quad \text{for all } \ i \geq 1 \iff \mathbb{E}(Q_i(Y)/q_i^2) = \alpha_i(\mu_0, \theta_0), \quad \text{for all } \ i \geq 1.$$

This condition implies that the expansion of the Ys' density matches with the expansion of the unknown density f with mean μ_0 and parameter θ_0, which is in contradiction with the semiparametric identifiability of model/setup H_1, see Bordes *et al.* (2006). Thus we can state that there exists an index j such that

$r_j \neq 0$. For simplicity matters let us consider $j_0 = \min\{j \geq 1 :\ r_j \neq 0\}$. Since from (2.6), for every $k \geq 1$ fixed, we can decompose $T_{k,n}$ as follows:

$$\begin{aligned} s(n)T_{k,n} &= n^{\lambda} U_{k,n}^T \widehat{D}_{k,n}^{-1} U_{k,n} \\ &= n^{\lambda-1} \sum_{\ell=1}^{k} \left(\sqrt{n} \left[\frac{R_{\ell,n} - r_\ell}{\sqrt{\widehat{D}_{k,n}[\ell]}} \right] \right)^2 + 2n^{\lambda-1/2} \sum_{\ell=1}^{k} \sqrt{n} \left[\frac{R_{\ell,n} - r_\ell}{\sqrt{\widehat{D}_{k,n}[\ell]}} \right] r_\ell \\ &\quad + n^{\lambda} \sum_{\ell=1}^{k} r_\ell^2, \end{aligned}$$

it comes that for all $k < j_0$, $T_{k,n} = O_p(n^{\lambda-1})$ since the r_ℓ's are all equal to zero for $1 \leq \ell \leq k$, when instead for the index j_0 we have $T_{j_0,n} \geq n^{\lambda} r_{j_0}^2 + O_p(n^{\lambda-1/2})$. It comes that for all $k < j_0$ we have

$$\mathbb{P}\left(s(n)T_{k,n} - \beta_k pen(n) < s(n)T_{j_0,n} - \beta_{j_0} pen(n)\right) \longrightarrow 1, \quad \text{as} \quad n \to +\infty.$$

This obviously shows, according to S_ns' definition (2.8), that $S_n \geq j_0$ with probability one as $n \to +\infty$. Now, since $T_{k,n}$ is a k-increasing sequence for every given $n \geq 1$, we have that $T_{S_n,n} \geq T_{j_0,n} \geq n^{\lambda} r_{j_0}^2 + O_p(n^{\lambda-1/2})$ which proves the wanted result. Note that the right hand side of the previous inequality shows clearly a drift of our test statistic in $O_p(n^{\lambda})$, $0 < \lambda < 1/2$, under the alternative H_1. □

Proposition 2. Similarly to the proof of Theorem 2, we have

$$\mathbb{P}\left(S_n^{(n)} \geq 2\right) \leq \mathbb{P}\left(\sum_{k=2}^{d(n)} n^{\lambda} \left(R_{k,n}^{(n)}\right)^2 \geq \log(n) e(n) \right). \tag{11.7}$$

To prove that the right hand side term of the above probability goes to zero as $n \to +\infty$, we decompose $R_{k,n}^{(n)}$ as follows:

$$R_{k,n}^{(n)} = \left(a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)})\right) - \left(\bar{p}_n^{(n)} \alpha_{k,n}^{(n)} - p_0 \alpha_k(\mu_0, \theta_0)\right) + \psi_{k,n}, \tag{11.8}$$

with $\alpha_{k,n}^{(n)} = \alpha_k(\bar{\mu}_n^{(n)}, \bar{\theta}_n^{(n)})$, and

$$\psi_{k,n} = p_0 \mathbb{E}^{(n)} \left(Q_k(X_0 + \delta_n \varepsilon_1) - Q_k(X_0)\right) / q_k^2, \tag{11.9}$$

which denotes the expectation of the k-th difference between the $H_1^{(n)}$ and H_0-distribution type supported by the second component in the mixture model (1.1), X_0 being H_0 distributed. By **(A6)** there exists $c > 0$ such that

$$\psi_{k,n}^2 \quad \leq c\delta_n^2. \tag{11.10}$$

We then have

$$
\begin{aligned}
&\mathbb{P}(S_n^{(n)} \geq 2) \\
&= \mathbb{P}\left(n^\lambda \sum_{k=2}^{d(n)} \left((a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)})) - (\bar{p}_n^{(n)}\alpha_{k,n}^{(n)} - p_0\alpha_k(\mu_0,\theta_0)) + \psi_{k,n} \right)^2 \geq \log(n)e(n) \right) \\
&\leq \mathbb{P}\left(n^\lambda \sum_{k=2}^{d(n)} \left(((a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)})) - (\bar{p}_n^{(n)}\alpha_{k,n}^{(n)} - p_0\alpha_k(\mu_0,\theta_0)))^2 + \psi_{k,n}^2 \right) \geq \log(n)e(n)/2 \right) \\
&\leq \mathbb{P}\left(n^\lambda \sum_{k=2}^{d(n)} \left((a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)})) - (\bar{p}_n^{(n)}\alpha_{k,n}^{(n)} - p_0\alpha_k(\mu_0,\theta_0)) \right)^2 \geq \log(n)e(n)/2 - cn^\lambda d(n)\delta_n^2 \right) \\
&\leq \mathbb{P}\left(\sum_{k=2}^{d(n)} \left((a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)}))^2 + (\bar{p}_n^{(n)}\alpha_{k,n}^{(n)} - p_0\alpha_k(\mu_0,\theta_0))^2 \right) \geq \log(n)e(n)/(4n^\lambda) - cd(n)\delta_n^2/2 \right) \\
&\leq \mathbb{P}\left(\sum_{k=2}^{d(n)} \left(a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)}) \right)^2 \geq C(k,n)/(8n^\lambda) \right) \\
&+ \mathbb{P}\left(\sum_{k=2}^{d(n)} \left(\bar{p}_n^{(n)}\alpha_{k,n}^{(n)} - p_0\alpha_k(\mu_0,\theta_0) \right)^2 \geq C(k,n)/(8n^\lambda) \right)
\end{aligned}
$$

where $C(k,n) = \log(n)e(n) - 2cd(n)n^\lambda\delta_n^2$. By **(A1)** we have $d(n) = O(\log(n)e(n))$, and $n^\lambda\delta_n^2 \to 0$ as $n \to +\infty$ due to **(A5)** (key point of the proof). It follows that

$$C(k,n) \quad = \quad \log(n)e(n) + o(\log(n)e(n)). \tag{11.11}$$

We study the two above probabilities separately. First we have, according to the Markov inequality and Condition **(A3)**, that

$$
\begin{aligned}
\mathbb{P}\left(\sum_{k=2}^{d(n)} \left(a_{k,n}^{(n)} - \mathbb{E}^{(n)}(a_{k,n}^{(n)}) \right)^2 \geq \frac{C(k,n)}{8n^\lambda} \right) &\leq \frac{8n^\lambda}{C(k,n)} \sum_{k=2}^{d(n)} \frac{1}{n} \mathrm{var}\left(\frac{Q_k(X_1^n)}{q_k^2} \right) \\
&\leq \frac{8d(n)}{n^{1-\lambda}C(k,n)} M_3,
\end{aligned}
$$

where the last right hand side term goes to zero as $n \to +\infty$ according to **(A1)**. Secondly we have

$$
\begin{aligned}
\mathbb{P}\left(\sum_{k=2}^{d(n)} \left(\bar{p}_n^{(n)}\alpha_{k,n}^{(n)} - p_0\alpha_k(\mu_0,\theta_0) \right)^2 \geq \frac{C(k,n)}{8n^\lambda} \right) &\leq \mathbb{P}\left(\sum_{k=2}^{d(n)} \left(\alpha_{k,n}^{(n)} \right)^2 \left(\bar{p}_n^{(n)} - p_0 \right)^2 \geq \frac{C(k,n)}{16n^\lambda} \right) \\
&+ \mathbb{P}\left(p_0^2 \sum_{k=2}^{d(n)} \left(\alpha_{k,n}^{(n)} - \alpha_k(\mu_0,\theta_0) \right)^2 \geq \frac{C(k,n)}{16n^\lambda} \right).
\end{aligned}
$$

By **(A2)** the α_k's are bounded by M_1 which leads to

$$\mathbb{P}\left(\sum_{k=2}^{d(n)}\left(\alpha_{k,n}^{(n)}\right)^2\left(\bar{p}_n^{(n)}-p_0\right)^2\geq\frac{C(k,n)}{16n^\lambda}\right)\leq\mathbb{P}\left(\sum_{k=2}^{d(n)}M_1^2\left(\bar{p}_n^{(n)}-p_0\right)^2\geq\frac{C(k,n)}{16n^\lambda}\right)$$
$$\leq\mathbb{P}\left(\left(\bar{p}_n^{(n)}-p_0\right)^2\geq\frac{C(k,n)}{16n^\lambda d(n)M_1^2}\right).$$

We next prove that the last right hand side term goes to zero as $n\to+\infty$. Combining **(A4)-(A5)** with (ii) of Theorem 2 (in Supplementary file) we have for all $\alpha>0$ and $0<\delta<1/2$,

$$\begin{aligned}\|\bar{\vartheta}_n^{(n)}-\vartheta_0\| &= O_{a.s.}\left(\left(n^{-1/2+\alpha}+n^{-1/4+(2\delta-\xi)}\right)^{1/2-\delta}\right)\\ &= O_{a.s.}\left(\left(n^{-1/2+\alpha}+n^{-1/2+(2\delta-\xi+1/4)}\right)^{1/2-\delta}\right),\end{aligned}$$

with $0<2\delta-\xi+1/4<1/4$. It follows that

$$\|\bar{\vartheta}_n^{(n)}-\vartheta_0\| = O_{a.s.}\left(\left(n^{-1/2+\alpha}\right)^{1/2-\delta}\right),$$

for all $0<\alpha<2\delta-\xi+1/4$ and $0<\delta<1/2$, and finally

$$\|\bar{\vartheta}_n^{(n)}-\vartheta_0\| = O_{a.s.}\left(n^{-1/4+\beta}\right), \tag{11.12}$$

for all $\beta>0$ small enough. Since $\lambda\in]0,1/2[$ we obtain $|\bar{p}_n^{(n)}-p_0|^2=o_{a.s.}(n^{-\lambda})$ and the assertion follows from 11.11 and **(A1)**. Writing $\rho_0=(\mu_0,\theta_0)$ and $\bar{\rho}_n^{(n)}=(\bar{\mu}_n^{(n)},\bar{\theta}_n^{(n)})$, similarly **(A5)-(A6)** give $\|\bar{\rho}_n^{(n)}-\rho_0\|^2=o_{a.s.}(n^{-\lambda})$. Since the $\dot{\alpha}_k$'s are bounded by M_2 according to **(A2)**, using the *mean value* Theorem, we obtain:

$$\mathbb{P}\left(p_0^2\sum_{k=2}^{d(n)}\left(\alpha_{k,n}^{(n)}-\alpha_k(\mu_0,\theta_0,)\right)^2\geq\frac{C(k,n)}{16n^\lambda}\right)\leq\mathbb{P}\left(\|\bar{\rho}_n^{(n)}-\rho_0\|^2\geq\frac{C(k,n)}{16n^\lambda d(n)M_2^2}\right),$$

which last term goes to zero as $n\to+\infty$ according to **(A1)**. Hence from (11.7), we obtain that $\mathbb{P}(S_n\geq 2)\to 0$ as $n\to+\infty$. Therefore, using the proofs of Corollary 3 we get the limiting distribution of the test statistic $T(n)$ under H_1^*. □

Proposition 3. Let us compute the close forms of the quantities $\psi_{1,n}$ and $\psi_{2,n}$ defined in (11.9). It first comes

$$\begin{aligned}\psi_{1,n}&=p_0\mathbb{E}^{(n)}(Q_1(X_0+\delta_n\varepsilon_1)-Q_1(X_0))\\ &=p_0\mathbb{E}^{(n)}(a_{1,1}(X_0+\delta_n\varepsilon_1)+a_{1,0}-a_{1,1}(X_0)-a_{1,0})\\ &=p_0\delta_n\mathbb{E}^{(n)}(\varepsilon_1),\end{aligned}$$

and we have

$$\begin{aligned} R_{1,n}^{(n)} &= \left(a_{1,n}^{(n)} - \mathbb{E}^{(n)}(a_{1,n}^{(n)})\right) - \left(\bar{p}^{(n)}\alpha_{1,n}^{(n)} - p_0\alpha(\mu_0,\theta_0)\right) + \psi_{1,n} \\ &= A - B + \psi_{1,n}. \end{aligned}$$

Combining Markov inequality and **(A3)** we obtain

$$\mathbb{P}^{(n)}(|a_{1,n}^{(n)} - \mathbb{E}^{(n)}(a_{1,n}^{(n)})| \geq 1/n) \quad \leq \quad \mathrm{var}\left(\frac{Q_1(X_1^n)}{q_1^2}\right) \quad < M3,$$

ensuring that $A = O_{a.s.}(1/n)$. Moreover

$$B = \alpha_{k,n}^{(n)}\left(\bar{p}^{(n)} - p_0\right) + p_0\left(\alpha_{k,n}^{(n)} - \alpha(\mu_0,\theta_0)\right) = B_1 + B_2.$$

From **(A2)** we have $|\alpha_{k,n}^{(n)}| \leq M1$ and from 11.12 we have $(\bar{p}^{(n)} - p_0) = o_{a.s.}(n^{-\lambda/2})$ which prove that $B_1 = o_{a.s.}(n^{-\lambda/2})$. In the same way, using **(A2)** we can show that $B_2 = o_{a.s.}(n^{-\lambda/2})$.

By **(A7)** it follows that almost surely $n^\lambda (R_{1,n}^{(n)})^2 \approx n^{\lambda - 2\xi'} \to +\infty$, as $n \to +\infty$. By construction we have $T_{1,n}^{(n)} \geq n(R_{1,n}^{(n)})^2/(\widehat{\mathrm{var}}(R_{1,n}^{(n)}) + e(n))$ which leads to the almost sure convergence

$$s(n)T_{1,n}^{(n)} - \log(n) \longrightarrow +\infty, \quad \text{as } n \to +\infty.$$

Under **(A8)** we obtain immediately that $\psi_{1,n} = 0$ and $R_{1,n} = o_{a.s.}(n^{-\lambda/2})$. Since $T_{1,n}^{(n)} \leq n(R_{1,n}^{(n)})^2/e(n)$, it follows that almost surely

$$s(n)T_{1,n}^{(n)} - \log(n) \longrightarrow -\infty, \quad \text{as } n \to +\infty.$$

We also have

$$\begin{aligned} \psi_{2,n} &= p_0\mathbb{E}^{(n)}(Q_2(X_0 + \delta_n\varepsilon_1) - Q_2(X_0)) \\ &= p_0(\mathbb{E}^{(n)}(a_{2,2}(X_0 + \delta_n\varepsilon_1)^2 + a_{2,1}(X_0 + \delta_n\varepsilon_1) + a_{2,0} \\ &\quad - a_{2,2}(X_0)^2 - a_{2,1}(X_0) - a_{2,0})) \\ &= 2p_0 a_{2,2}\delta_n^2\mathbb{E}(\varepsilon_1^2). \end{aligned}$$

From the above expressions and by definition of $R_{2,n}^{(n)}$ in (11.8) we can mimic the previous arguments to show that almost surely $R_{2,n}^{(n)} \approx \delta_n^2$ and that

$$\begin{aligned} &s(n)T_{2,n}^{(n)} - 2\log(n) \\ &= s(n)\left(n(R_{1,n}^{(n)})^2\widehat{D}_{1,n}^{-1} + n(R_{k,n}^{(n)})^2\widehat{D}_{2,n}^{-1}\right) - 2\log(n) \\ &\geq n^\lambda\left((R_{1,n}^{(n)})^2/(e(n) + \widehat{\mathrm{var}}(R_{1,n}^{(n)})) + (R_{2,n}^{(n)})^2/(e(n) + \widehat{\mathrm{var}}(R_{2,n}^{(n)}))\right) - 2\log(n), \end{aligned}$$

where the last right hand side term goes to infinity as $n \to +\infty$ which gives us the wanted result. □

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

[2] Al Mohamad, D. and Boumahdaf, A. (2018) Semiparametric two-component mixture models when one component is defined through linear constraints. *IEEE Trans. Information Theory*, **64**, 795–830.

[3] Arias-Castro, E. and Huang, R. (2018) The sparse variance contamination model. *Preprint.* `arXiv:807.10785v1.`

[4] Balabdaoui, F. and Doss, C.R. (2018) Inference for a two-component mixture of symmetric distributions under log-concavity. *Bernoulli,* **24**, 1053–1071.

[5] Di Zio, M. and Guarnera, U. (2013) A Contamination Model for Selective Editing. *J. official Statist.*, **29**, 539–555.

[6] Berrett, T.B., Samworth, R.J, and Yuan, M. (2019) Efficient multivariate entropy estimation via k k-nearest neighbour distances, *Ann. Statist.* **47**, 288–318.

[7] Bordes, L., Delmas, C. and Vandekerkhove, P. (2006) Semiparametric estimation of a two-component mixture model when a component is known. *Scand. J. Statist.*, **33**, 733–752.

[8] Bordes, L. and Vandekerkhove, P. (2010) Semiparametric two-component mixture model when a component is known: an asymptotically normal estimator. *Math. Meth. Statist.*, **19**, 22–41.

[9] Dai, H. and Charnigo, R. (2010) Contaminated normal modeling with application to microarray data analysis. *Can. J. Statist.* **38**, 315–332.

[10] Doukhan, P., Pommeret, D. and Reboul, L. (2015) Data driven smooth test of comparison for dependent sequences. *J. Multivar. Analys.*, **139**, 147–165.

[11] Gassiat, E. (2018) Mixtures of Nonparametric Components and Hidden Markov Models. Handbook of Mixture Analysis (ed. G. Celeux, S. Fruhwirth-Schnatter, C. Robert, Chap. 12) *To appear.*

[12] Ghattas, B., Pommeret, D., Reboul, L. and Yao, A. F. (2011) Data driven smooth test for paired populations. *J. Stat. Plan. Inference* 141: 262–275.

[13] Hedenfalk, I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.

[14] Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1997) Data driven smooth tests for composite hypotheses. *Ann. Statist.*, **25**, 1222–1250.

[15] Klingenberg, C., Pirner, M. and Puppo, G. (2017) A consistent kinetic model for a two-component mixture with an application to plasma. *Kinet. Relat. Models*, **10**, 445–465.

[16] Ledwina, T. (1994) Data-driven version of Neyman's smooth test of Fit.*J. Amer. Statist. Assoc.* **89**, 1000–1005.

[17] Lindsay, B. G. (1983) The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86–94.

[18] Lindsay, B. G. (1989) Moment matrices: applications in mixtures. *Ann.*

Statist., **17**, 722–740.

[19] Ma, Y. and Yao, W. (2015) Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electr. J. Statist.*, **9**, 444–474.

[20] McLachlan, G. J., Bean, R.W. and Ben-Tovim Jones, L. (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

[21] Nguyen, V. H. and Matias, C. (2014) On Efficient Estimators of the Proportion of True Null Hypotheses in a Multiple Testing Setup. *Scan. J. Statist.*, **41**, 1167–1194.

[22] Melchior, P. and Goulding, A. D. (2018) Filling the gaps: Gaussian mixture models from noisy, truncated or incomplete samples. *Astronomy and Computing,* **25**, 183–194.

[23] Munk, A., Stockis, J. P., Valeinis, J. and Giese, G. (2010) Neyman smooth goodness-of-fit tests for the marginal distribution of dependent data. *Ann. Instit. Statist. Math.,* **63**, 939–959.

[24] Neyman, J. (1937) Smooth Test for Goodness of Fit, *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

[25] Patra, R. K. and Sen, B. (2016) Estimation of a Two-component Mixture Model with Applications to Multiple Testing. *J. Roy. Statist. Soc., Series B*, **78**, 869–893.

[26] Podlaski, R. and Roesch, F.A. (2014) Modelling diameter distributions of two-cohort forest stands with various proportions of dominant species: A two-component mixture model approach. *Math. Biosci.*, **249**, 60–74.

[27] Quandt, R. E. and Ramsey, J. B. (1978) Estimating mixtures of normal distributions and switching regressions (with comments). *J. Am. Statist. Ass.*, **73**, 730–752.

[28] Robin, A. C., Reyl, C., Derrire, S. and Picaud, S. (2003) A synthetic view on structure and evolution of the Milky Way. *Astron. Astrophys.*, **409**, 523–540.

[29] Suesse, T., Rayner, J. C. W. and Thas, O. (2017) Assessing the fit of finite mixture distributions. *Aust. N. Z. J. Stat.*, **59**, 463–483.

[30] Szegö, G. (1939) *Orthogonal Polynomials.* Amer. Math. Soc., Colloquium Publications Volume XXIII.

[31] Walker, M. G., Mateo, M., Olszewski, E. W., Sen, B. and Woodroofe, M. (2009) Clean kinematic samples in drarf spheroidals: an algorithm for evaluating membership and estimating distribution parameters when contamination is present. *The Astronomical Journal,* **137**, 3109–3138.

[32] van't Wout, A.B. *et al.* (2003) Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-cell lines. *J. Virol.*, **77**, 1392–1402.

[33] Wylupek, G. (2010) Data driven K-sample tests. *Technometrics,* **52**, 107–123.

[34] Xiang, S., Yao, W. and Yang, G. (2018) An overview of Semiparametric Extensions of finite Mixture Models. *Preprint.* `arXiv:1811.05575v1`.