Semiparametric two-sample mixture components comparison test

Xavier Milhaud, Denys Pommeret, Yahia Salhi, Pierre Vandekerkhove

Semiparametric two-sample mixture components comparison test

Xavier Milhaud[(1)] Denys Pommeret[(1,2)] Yahia Salhi[(1)] and Pierre Vandekerkhove[(3)]

[(1)]Univ Lyon, UCBL, ISFA LSAF EA2429, F-69007, Lyon, France
[(2)]Aix-Marseille University, Campus de Luminy, 13288 Marseille cedex 9, France
[(3)]Université Gustave Eiffel, LAMA (UMR 8050), 77420 Champs-sur-Marne, France

February 25, 2020

Abstract

We consider in this paper two-component mixture distributions having one known component. This is the case when a gold standard reference component is well known, and when a population contains such a component plus another one with different features. When two populations are drawn from such models, we propose a penalized χ^2-type testing procedure able to compare pairwise the unknown components, *i.e.* to test the equality of their residual features densities. An intensive numerical study is carried out from a large range of simulation setups to illustrate the asymptotic properties of our test. Moreover the testing procedure is applied on two real cases: i) mortality datasets, where results show that the test remains robust even in challenging situations where the unknown component only represents a small percentage of the global population, ii) galaxy velocities datasets, where stars luminosity mixed with the Milky Way are compared.

Keywords: finite mixture model; semiparametric estimator; Chi-squared test; mortality.

1 Introduction

Let us consider the two-component mixture model with probability density function (pdf) h defined by

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

where g is a known pdf and where the unknown parameters are the mixture proportion $p \in]0,1[$ and the pdf f. This model has been widely investigated in the last decades, see for instance Bordes and Vandekherkove [3], Matias and Nguyen [23], Cai and Jin [5] or Celisse and Robin [6] among others. Numerous applications of model (1) can be found in topics such as: i) genetics regarding the analysis of gene expressions from microarray experiments such as in Broët *et al.* [4]; ii) the false discovery rate problem (used to assess and control multiple error rates such as in Efron and Tibshirani [9]), see McLachlan *et al.* [19]; iii) astronomy, in which this model arises when observing variables such as metallicity and radial velocity of stars such as in Walker *et al.* [29]; iv) biology to model trees diameters, see Podlaski and Roesch [25]; v) kinetics to model plasma data, see Klingenberg *et al.* [13].

In this paper, the data of interest is made of two i.i.d. samples $X = (X_1, \ldots, X_{n_1})$ and $Y = (Y_1, \ldots, Y_{n_2})$ with respective probability density functions:

$$\begin{cases} h_1(x) = (1-p_1)g_1(x) + p_1 f_1(x), & x \in \mathbb{R}, \\ h_2(x) = (1-p_2)g_2(x) + p_2 f_2(x), & x \in \mathbb{R}, \end{cases} \tag{2}$$

where p_1, p_2 are the unknown mixture proportions and f_1, f_2 are the unknown component densities with respect to a given reference measure ν. Note that we can also consider discrete measures, as Poisson or Binomial, extending the notion of probability density function to mass probability function when ν refer to the counting measure. All our results will be still valid in such setups. Given the above model, our goal is now to answer the following statistical problem:

$$H_0: \quad f_1 \text{ is equal to } f_2 \qquad \text{against} \qquad H_1: \quad f_1 \text{ is different from } f_2 \ , \tag{3}$$

without assigning any specific parametric family to the f_i's.

This problem is a natural extension of a recent work by Pommeret and Vandekherkove [26]. Our contribution here is twofold: we extend their results to the two-samples case, and the proposed method enables to consider problem (3) without specifying any type of distribution (Pommeret and Vandekherkove [26] focused their study on the continuous case with symmetric densities). Basically our test procedure consists in expanding the two unknown densities in an orthogonal polynomial basis, and then in comparing, with an *ad. hoc.* method, their coefficients up to a parsimonious rank selected according to a data-driven technique detailed latter on in the paper.

In our case, the first practical interest of this paper relates to the demography analysis and corresponding applications in actuarial science. Mortality is shown to vary across individuals due to many factors; including age, sex, education, health and marital status among others. In insurance, the probability distribution of the age-at-death random variable is of paramount importance. Indeed, this is the basis for the premium calculations as well as the solvency capital requirement assessment, see Barrieu *et al.* [1]. In such a context, we are generally facing some heterogeneity which arises from the adverse selection effect (caused by the insurance guarantees themselves). Death insurance contracts are thus sold to people presenting heterogeneous mortality risks: some individuals have a mortality profile similar to the (average) national population; whereas others have a different risk profile due to anti-selection. Insurers are therefore concerned with the mitigation of those risks, especially when the latter population is composed of less healthy lives. Of course, this adverse selection effect may have various impacts on different portfolios, since the pace of change and level of mortality are highly heterogeneous from one insurance contract to another. Ordinary life table analyses implicitly assume that the population is homogeneous, which clearly looks unrealistic in a practical setting. Keyfitz and Littman [15] illustrated the potential bias due to heterogeneity using a life table model where the age-at-death density follows a discrete mixture of homogeneous subgroups, as in (1). They showed that, in general, ignoring heterogeneity leads to incorrect assessments of life expectancies. From an insurance risk management viewpoint, the heterogeneity brings about two main challenges. First, to get an efficient probabilistic representation of this unobserved heterogeneity is not trivial: given that the first component adjusts to the national population, the statistical estimation of the second component f in (1) can sometimes be tricky. This is the so-called basis risk, typically arising as a result of the anti-selection effect such as in Salhi and Loisel [28]. Secondly, we are interested in comparing the heterogeneity across different guarantees (portfolios). Since insurance risk management rests on the pooling of a large number of ideally uncorrelated risks, the main component of the mortality risk is generally well understood and managed. As a matter of fact, most of the individuals exhibit a mortality pattern similar to the national population. However, unusual mortality

profiles pose an issue, as mitigation is made infeasible due to low levels of representation. Therefore, it is of great interest to develop a hypothesis testing that aims at comparing the heterogeneous component of two different populations, allowing the insurer to benefit from the pooling of those two populations.

Eventually, our method can be used in many other areas than actuarial science. As illustration we also consider kinematic datasets from two Milky Way dwarf spheroidal (dSph) satellites: Carina and Sextans, see for instance Walter *et al.* [29]. More precisely, we consider the heliocentric velocities (HV) of stars in these satellites, that is the velocities defined with respect to the solar system. These measurements are mixed with the HV of stars in the Milky Way. The Milky Way is largely observed, see Robin *et al.* [27], and can be assumed perfectly known. One problem is to compare the HV distributions of both satellites Carina and Sextans through such mixture models with a common Milky Way known component. We therefore encounter a problem of two sample comparison of mixing components.

The remainder of the paper is organized as follows. We introduce the testing problem and describe our methodology in Section 2. In Section 3, we state the assumptions and asymptotic results under the null hypothesis, along with the test divergence under the alternative. Section 4 provides details about the adequate polynomial decomposition depending on the nature of the distributions support. In Section 5, we implement a simulation-based study to evaluate the empirical level and power of the test. Finally, Section 6 is devoted to applications on real datasets: first in mortality (with insurance contracts embedding death guarantees); second in kinematic (with galaxies heliocentric velocities comparisons). A discussion closes the paper when proofs are relegated in Appendices.

2 Testing problem

Our test procedure is based on the expansion coefficients comparison of the two probability density functions h_1 and h_2, defined in (2), in an orthonormal polynomial basis. Such an approach was originally introduced by Neyman [22] and extended in a data-driven context by Ledwina [16]. Our test procedure will permit to detect asymptotically any departure between two expansion coefficients, screened pairwise, along the indices.

Remark 1. *For technical reasons, we assume in the sequel that*

$$n_1/(n_1+n_2) \to a \in]0,1[\text{ as } n_1, n_2 \to \infty. \tag{4}$$

This condition is not restrictive and is obviously fulfilled when using the technique on real datasets, which corresponds to finite sample applications.

Let us denote by $\mathcal{Q} = \{Q_k; k \in \mathbb{N}\}$, an ν-orthonormal basis satisfying $Q_0 = 1$ and such that

$$\int_{\mathbb{R}} Q_j(x) Q_k(x) \nu(dx) = \delta_{jk},$$

with $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.

We assume that the following integrability conditions are satisfied

$$\int_{\mathbb{R}} h_1^2(x)\nu(dx) < \infty \quad \text{and} \quad \int_{\mathbb{R}} h_2^2(x)\nu(dx) < \infty.$$

Then, for all $x \in \text{supp}(\nu)$, we have for $i = 1, 2$

$$h_i(x) = \sum_{k\geq 0} h_{i,k} Q_k(x) \quad \text{with} \quad h_{i,k} = \int_{\mathbb{R}} Q_k(x) h_i(x) \nu(dx),$$
$$g_i(x) = \sum_{k\geq 0} g_{i,k} Q_k(x) \quad \text{with} \quad g_{i,k} = \int_{\mathbb{R}} Q_k(x) g_i(x) \nu(dx),$$
$$f_i(x) = \sum_{k\geq 0} f_{i,k} Q_k(x) \quad \text{with} \quad f_{i,k}s = \int_{\mathbb{R}} Q_k(x) f_i(x) \nu(dx),$$

and, from (2), we deduce that

$$h_{i,k} = (1 - p_i) g_{i,k} + p_i f_{i,k}.$$

Note that there is no restriction on the support of f_1 and f_2, excepted it must be known. The null hypothesis can be rewritten as $f_{1,k} = f_{2,k}$, for all $k \geq 1$. Or equivalently

$$H_0 : \; p_2(h_{1,k} - (1 - p_1) g_{1,k}) = p_1(h_{2,k} - (1 - p_2) g_{2,k}), \qquad k \geq 1. \tag{5}$$

Since the pdfs g_1 and g_2 are known, the coefficients $g_{i,k}$, $i = 1, 2$, are automatically known. For all $k \geq 1$, the coefficients $h_{i,k}$ can be estimated empirically by:

$$\widehat{h}_{1,k} = \frac{1}{n_1} \sum_{i=1}^{n_1} Q_k(X_i) \quad \text{and} \quad \widehat{h}_{2,k} = \frac{1}{n_2} \sum_{i=1}^{n_2} Q_k(Y_i). \tag{6}$$

The estimation of the proportions p_i, $i = 1, 2$, involved in model (2) will depend on some technical assumptions. In fact, we can distinguish the following cases.

- **Semiparametric conditions**: the following assumptions allow to semiparametrically identify model (2) and to estimate the parameters p_1 and p_2 of crucial importance in our testing method, see expression (10).

  (A1) The regularity and identifiability conditions required in Bordes and Vandekerkhove [3] and Bordes *et al.* [2] are satisfied. Regarding specifically the identifiability conditions we will suppose either:

  a) The densities g_i and f_i (i=1,2) are respectively supposed to be odd and symmetric about a location parameter μ_i, *i.e.* there exists $\mu_i \in \mathbb{R}$ such that for all $x \in \mathbb{R}$ $f_i(x + \mu_i) = f_{i,s}(x) = f_{i,s}(-x)$, with 2nd order moments supposed to satisfy

  $$m(g_i) \neq m(f_{i,s}) + \mu_i \frac{2 \pm k}{3k}, \qquad \text{for } k \in \mathbb{N}^*, \text{ and } i = 1, 2, \tag{7}$$

  where $m(f)$ generically denotes the 2nd order moment according to the f density.

  b) The densities g_i and f_i (i=1,2) are respectively supposed to be strictly positive over $\mathbb{R}$ and symmetric about a location parameter μ_i, both having first order moments and satisfying the following tail conditions:

  $$\text{for all } \beta \in \mathbb{R} : \quad \lim_{x\to+\infty} \frac{f_{i,s}(x - \beta)}{g_i(x)} = 0, \quad \text{or} \quad \lim_{x\to-\infty} \frac{f_{i,s}(x - \beta)}{g_i(x)} = 0, \quad i = 1, 2.$$

The central role of the above conditions in the semiparametric literature are detailed in the recent and very well documented survey by Xiang et al. [31].

(A2) The regularity and identifiability conditions required in Patra and Sen [24] are satisfied.

Note that in Patra and Sen [24] the estimation of the proportion p in model (1) is based on the fact that the correct p should be defined as

$$p_0 = \inf \{p \in (0,1] : \ [H-(1-p)G]/p \text{ is a cdf}\}.$$

where we recognize the inversion formula $H = (1-p)G + pF \Leftrightarrow F = [H-(1-p)G]/p$ under $p \in (0,1]$ when G is known and H can be estimated from the data. This second semiparametric method is more flexible than Bordes and Vandekerkhove [3] since it can be basically used on any sort of distribution support, *i.e.* continuous, discrete or a combination of both, as described in Patra and Sen [24, Lemmas 2-4]. However, they do not obtain a central limit theorem for their estimators and therefore a bootstrap technique is required to calibrate the distribution of the test statistics.

- **Parametric condition**:

  (A3) For $i = 1, 2$, there exists known functions ℓ_i and intervals I_i such that the quantities $F_i^\ell = \int_{I_i} \ell_i(x) f_i(x) \nu(dx)$ are known.

Under parametric condition **(A3)** parameters p_i can be estimated by

$$\bar{p}_i = \frac{\widehat{H}_i^\ell - G_i^\ell}{F_i^\ell - G_i^\ell}, \tag{8}$$

where $G_i^\ell = \int_{I_i} \ell_i(x) g_i(x) \nu(dx) \neq F_i$ and $\widehat{H}_i$ is an estimator of $H_i^\ell = \int_{I_i} \ell_i(x) h_i(x) \nu(dx)$, typically:

$$\widehat{H}_i^\ell = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell_i(X_j) \mathbb{I}_{X_j \in I_i}. \tag{9}$$

Remark 2. *Assumption* **(A3)** *can be reduced to the knowledge of a restricted mean when* $\ell_i(x) = x$*, or a value of a distribution function when* $\ell(x) = 1$*. In the case where* I_i *is not in the support of* g_i*, we could estimate* I_i *from the sample. The sole knowledge of* G_i *instead of* g_i *is also possible. Other various parametric assumptions are studied in the literature (see for instance Celisse and Robin [6] and the references given there)*

Henceforth, we will denote by $\widehat{p}$ the estimator of p, whatever the considered assumption among **(A1-3)**, *i.e.* $\widehat{p}$ will be equal to $\widetilde{p}$ under **(A1-2)**, and to $\bar{p}$ under **(A3)**. Finally, to answer the H_0 testing problem (3), we consider the following double-sourced differences

$$\widehat{R}_k := \widehat{p}_2(\widehat{h}_{1,k} - (1-\widehat{p}_1)g_{1,k}) - \widehat{p}_1(\widehat{h}_{2,k} - (1-\widehat{p}_2)g_{2,k}), \quad k \geq 1, \tag{10}$$

allowing to detect any possible departure from the null hypothesis.

For all $k \geq 1$, we define $\widehat{U}_k = (\widehat{R}_1, \ldots, \widehat{R}_k)$, and

$$\widehat{T}_k = \frac{n_1 n_2}{n_1 + n_2} \widehat{U}_k^\top \widehat{D}_k^{-1} \widehat{U}_k, \tag{11}$$

where $\widehat{D}_k = \mathrm{diag}(\widehat{d}_1, \ldots, \widehat{d}_k)$ is a consistent estimator of $\mathrm{diag}(\mathbb{V}(\widehat{R}_1), \ldots, \mathbb{V}(\widehat{R}_k))$.

To avoid instability in the evaluation of $\widehat{D}_k^{-1}$, we add a trimming term $e(n_1, n_2)$ satisfying $e(n_1, n_2) \to 0$ as n_1, n_2 tend to infinity, and we finally consider

$$\widehat{d}_j = \max\left(\widehat{w}_j, e(n_1, n_2)\right), \quad 1 \leq j \leq k, \tag{12}$$

where $\widehat{w}_j$ is an estimator of $\mathbb{V}(\widehat{R}_j)$.

Following Ledwina [16] and Kallenberg and Ledwina [14], we suggest a data-driven procedure to select automatically the number of coefficients needed to answer the testing problem. Formally, we introduce the following penalized rule to select the rank k of the statistic $\widehat{T}_k$:

$$S(n_1, n_2) = \min\left\{ \underset{1 \leq k \leq d(n_1, n_2)}{\mathrm{argmax}} \left(s(n_1, n_2)\widehat{T}_k - \beta_k \mathrm{pen}(n_1, n_2) \right) \right\}, \tag{13}$$

where $d(n_1, n_2) \to +\infty$ as $n_1, n_2 \to +\infty$, $pen(n_1, n_2)$ is a penalty term such that $pen(n_1, n_2) \to +\infty$ as $n_1, n_2 \to +\infty$, the β_k's are penalization factors, and $s(n_1, n_2)$ is a normalization factor which depends on the convergence rate of the estimators of p_1 and p_2. In practice, we will consider $\beta_k = k$, $k \geq 1$, and $pen(n_1, n_2) = \log(n_1 n_2/(n_1 + n_2))$, $n_1, n_2 \geq 1$. The rate $s(n_1, n_2)$ will depend on the assumptions made about the densities f_1 and f_2 and is related to the rate of convergence of the parameter estimators.

- Under Assumptions **(A1-2)**, we fix $s(n_1, n_2) = \left(\dfrac{n_1 n_2}{n_1 + n_2} \right)^{s-1}$, with $0 < s < 1/2$.
- Under Assumption **(A3)**, we fix $s(n_1, n_2) = 1$.

Finally the associated data-driven test statistic is $T(n_1, n_2) = \widehat{T}_{S(n_1, n_2)}$.

3 Additional assumptions and main results

To test consistently (3), based on the statistic $T(n_1, n_2)$, we will suppose the following conditions:

(A4) The coefficient order upper bound $d(n_1, n_2)$ involved in (13) satisfies

$$d(n_1, n_2) = o(\log(n_1 n_2/(n_1 + n_2)) e(n_1, n_2)).$$

(A5) There exist nonnegative constants M_1, M_2 such that for all $k \geq 1$, under H_0

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{E}(Q_j^2(X)) < M_1, \qquad \text{and} \qquad \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}(Q_j^2(Y)) < M_2.$$

The next two theorems state respectively the asymptotic behavior of the selected rank $S(n_1, n_2)$ defined in (13) under nonparametric, resp. parametric, conditions.

Theorem 3. *If one of these three assumptions is satisfied:* **(A1)**, *or* **(A2)**, *or* **(A3)**, *and if* **(A4)** *and* **(A5)** *hold, then, under* H_0, $S(n_1, n_2)$ *converges in probability towards 1 as* $n_1, n_2 \to +\infty$.

From Theorem 3, $T(n_1, n_2)$ and $\widehat{T}_1$ have the same limiting distribution in both parametric or nonparametric cases. Under **(A1)** or under **(A3)**, the estimators $\widehat{p}_1$ and $\widehat{p}_2$ are asymptotically Gaussian under the null and then we deduce the limit distribution of the test statistic from the previous theorems.

Proposition 1. *Assume that the estimators $\widehat{p}_1$ and $\widehat{p}_2$ are asymptotically Gaussian with convergence rate $\sqrt{n_1}$ and $\sqrt{n_2}$, respectively. Then T_1 converges in law towards a χ^2-distribution with one degree of freedom as $n_1, n_2 \to +\infty$.*

Corollary 4. *Assume that* **(A1),(A4-5)** *hold, or* **(A3-5)** *hold, then, under H_0, $T(n_1, n_2)$ converges in law towards a χ^2-distribution with one degree of freedom as $n_1, n_2 \to +\infty$.*

Under **(A2)** we do not have such an asymptotic results. Nevertheless, based on our numerical studies we conjecture that we are very close, in practice, to a normal asymptotic behavior for $\widehat{p}_1$ and $\widehat{p}_2$. We give a numerical illustration in Fig.7 Appendix B where the variance of these estimators is obtained by a bootstrap procedure. We used Patra and Sen [24] estimators instead of those proposed in [3] in our simulation study when the known density was not symmetric. In case of symmetry, both estimation methods yield very similar results, going into the direction of the above conjecture.

We consider now the collection of H_1-type alternatives defined as follows: there exists $q \in \mathbb{N}^*$ such that

$$H_1(q): \quad f_{1,j} = f_{2,j},\ j = 1, \ldots, q-1, \quad \text{and} \quad f_{1,q} \neq f_{2,q},$$

which describes a departure between the q-th order coefficients of h_1 and h_2. Writing

$$\delta(k) := p_2(h_{1,k} - (1-p_1)g_{1,k}) - p_1(h_{2,k} - (1-p_2)g_{2,k}), \qquad k \geq 1, \tag{14}$$

the alternative hypothesis $H_1(q)$ tells that $\delta(q)$ is the first non null coefficient along this series. We can now state the following proposition that describes the asymptotic drift of the test statistics under $H_1(q)$.

Proposition 2. *Assume that* **(A1),(A4-5)** *hold, or* **(A3-5)** *hold, then, under $H_1(q)$, we have $S(n_1, n_2) \to s \geq q$ and $T(n_1, n_2) \to +\infty$, when $n_1, n_2 \to \infty$, that is, for all $\epsilon > 0$, $\mathbb{P}(T(n_1, n_2) < \epsilon) \to 0$.*

4 Choice of the reference measure and test construction

4.1 Choice of the adequate reference measure

We propose in this section to advice on the most relevant reference measure ν to be used for the computation of coefficients $h_{i,k}$, $g_{i,k}$ and $f_{i,k}$ given in Section 2.

i) Real line support: the Gaussian measure. When the support of both unknown mixture components is the real line, we can chose for ν the standard normal distribution. The set $\{Q_k, k \in \mathbb{N}\} = \{H_k,\ k \in \mathbb{N}\}$ is constructed from the orthogonal Hermite polynomials, defined for all $x \in \mathbb{R}$ by:

$$H_0 = 1,\ H_1(x) = x,\ H_{k+1}(x) = 2xH_k(x) + 2nH_{k-1}(x), \quad k \geq 1. \tag{15}$$

ii) Real line support: the Lebesgue measure. When the support is the real line, another choice for ν is the Lebesgue measure on $\mathbb{R}$. In that case we can choose for $\{Q_k, k \in \mathbb{N}\} = \{\mathcal{H}_k,\ k \in \mathbb{N}\}$ the set of orthogonal Hermite functions, defined for all $x \in \mathbb{R}$ by:

$$\mathcal{H}_k(x) = H_k(x)\sqrt{f_{\mathcal{N}(0,1)}(x)}, \quad k \geq 0.$$

where H_k is the k-th Hermite polynomial defined in (15).

iii) Positive real line support: the Gamma measure. When the support of both unknown mixture components is the positive real line, we can chose for ν a gamma distribution $\Gamma(1,\alpha)$, with $\alpha > -1$. The set $\{Q_k, k \in \mathbb{N}\}$ is then constructed from the orthogonal Laguerre polynomials defined for all $x \in \mathbb{R}$ by:

$$\begin{aligned} \mathcal{L}_0^\alpha(x) = 1, \quad \mathcal{L}_1^\alpha(x) = -x + \alpha + 1, \\ -x\mathcal{L}_k^\alpha(x) = (k+1)\mathcal{L}_{k+1}^\alpha(x) - (2k+\alpha+1)\mathcal{L}_k^\alpha(x) + (k+\alpha)\mathcal{L}_{k-1}^\alpha(x), \quad k \geq 1. \end{aligned}$$

iv) Discrete support: the Poisson measure. If the common support is the set of integers then the choice of ν can be the Poisson distribution with mean $\alpha > 0$ and with associated orthogonal Charlier polynomials defined by:

$$\mathcal{C}_0^\alpha = 1, \quad \mathcal{C}_1^\alpha(x) = (\alpha - x)/\alpha, \qquad x\mathcal{C}_n^\alpha(x) = -\alpha\mathcal{C}_{n+1}^\alpha(x) + (n+\alpha)\mathcal{C}_n^\alpha(x) - n\mathcal{C}_{n-1}^\alpha(x), \quad k \geq 1.$$

v) Bounded support. If the supports are a bounded interval (a,b), $a < b$, we can use a uniform measure for ν and its associated Legendre polynomials. For instance, when $(a,b) = (-1,1)$ these polynomials are defined for all $x \in \mathbb{R}$ by:

$$L_0 = 1, \quad L_1(x) = x, \quad (k+1)L_{k+1}(x) = (2k+1)xL_k(x) - kL_{k-1}(x), \quad k \geq 1.$$

vi) Wavelets. Another approach is to consider an orthogonal basis of wavelets, say $\{\phi_i, \psi_{i,j}; i,j \in \mathbb{Z}\}$, see Daubechies [7]. Note here that the measure ν is the Lebesgue one and we can change $\phi_i, \psi_{i,j}$ into $\phi_i/h, \psi_{i,j}/h$ to keep our assumptions **(A4)**. Then the density expansions would take the following generic form:

$$f = \sum_{i\in\mathbb{Z}} \langle f, \phi_i\rangle\overline{\phi_i} + \sum_{i\in\mathbb{N}, j\in\mathbb{Z}} \langle f, \psi_{ij}\rangle\overline{\psi_{ij}},$$

with a double sum, heavier to implement in practice.

4.2 Construction of the test statistic

The computation of the test statistic $T(n_1,n_2) = \widehat{T}_{S(n_1,n_2)}$, requires the estimation of the variances $\mathbb{V}(\widehat{R}_1), \ldots, \mathbb{V}(\widehat{R}_k)$. To overcome the complex dependence between the estimators of p_1, p_2 and the estimators of the coefficients associated to h_1 and h_2, we split each sample into two independent sub-samples of size n_1', n_1'' for X and n_2', n_2'' for Y, with $n_1' + n_1'' = n_1$ and $n_2' + n_2'' = n_2$. Then we use the first sub-samples to estimate the proportions p_1 and p_2, and the second sub-samples to estimate the coefficients of h_1 and h_2.

Under **(A1)** or **(A3)**, Corollary 4 gives the asymptotic distribution of the test statistic. Under **(A2)**, there is no result on the asymptotic distribution of the estimators proposed in Patra and Sen [24]. We conjecture that a central limit theorem occurs for $\widehat{p}$. To verify empirically this assumption, we use a bootstrap procedure. By doing so, we estimate the asymptotic variance of the test statistic, and we consider numerically the distribution of our test statistic for different cases detailed in Appendix B.

The computation of the test statistic first requires the choice of $d(n_1,n_2)$, $e(n_1,n_2)$ and $s(n_1,n_2)$. A previous study (see Pommeret and Vandekherkove [26]) showed us that the empirical levels and powers were overall weakly sensitive to $d(n_1,n_2)$ for $d(n_1,n_2)$ large enough. From that preliminary study we decided to set $d(n_1,n_2)$ equal to 10. The trimming $e(n_1,n_2)$ is calibrated equal to $(\log(\max(n_1,n_2)))^{-1}$. The power of the normalization $s(n_1,n_2) = (n_1n_2/(n_1+n_2))^{s-1}$ is setup close enough to $(-1/2)$, with s equal to $2/5$, which seemed to provide good empirical levels.

5 Monte-Carlo simulations

Recall that X and Y respectively follow mixture densities h_1 and h_2, given as in (2), where p_1, p_2 are the unknown mixing proportions and f_1, f_2 are the unknown component densities with respect to some reference measure ν. Hereafter, simulations are performed to evaluate the empirical level of the test. This level corresponds to the probability of rejecting H_0 when H_0 is true ($f_1 = f_2$). In practice, this level is expected to asymptotically reach 5%, since one compares our test statistic to the 95-percentile of the χ^2-distribution (Corollary 4). We also assess the power of the test, i.e. the probability to reject H_0 given that H_0 is false, which gives an idea of the ability of the test to detect departures from the null hypothesis.

Usually, statisticians initially check for low levels of the test in many frameworks before analyzing its power. To have a deeper understanding of the strengths and weaknesses of the test, various simulation schemes are considered including finite mixture models with different component distributions and weights. Also, to check whether the test quality remains acceptable in diverse settings, we make the parameters of component distributions vary. Basically, we introduce two opposite situations. In the former, the two component densities are in close proximity; whereas component densities are far apart in the latter. In the sequel, for each case of the simulation study, the test is performed one hundred times to evaluate the empirical level (or power). We also fix $n = n_1 = n_2$ for conciseness when presenting the results (the case where $n_1 \neq n_2$ naturally arises with real datasets, see Section 6).

5.1 Test empirical levels

Firstly, our objective is to check whether the results significantly differ when changing the component weights and the component distributions of the mixture models. About the weights, we focus on values ranging from 10% to 70%. Such weights are typical in most of applications, where the unknown mixture component can be prevalent or not (see Section 6). One of the challenges is that the test remains effective when the unknown component proportions are low, meaning that few observations would be assigned to them in practice.

To begin with, consider two-component mixtures of Gaussian distributions. Recall that, under the null hypothesis, $f = f_1 = f_2$. The test is conducted in the following cases: a) $g_1 = g_2$ with g_1 far from (ff) f; b) $g_1 = g_2$ with g_1 close to (ct) f; c) $g_1 \neq g_2$ with g_1 ff g_2 and g_1, g_2 ff f; d) $g_1 \neq g_2$ with g_1 ct f and g_2 ff f (the case where $g_1 \neq g_2$ with g_1 ct g_2 is similar to b)). As an illustration, Fig. 1 shows the densities obtained for each of the aforementioned situations (see Tab. 7 in Appendix C.1 for associated parameters), with $n = 4,000$ observations. Notice that it is sometimes not obvious that h_i follows a mixture distribution (due to component proximity, see cases b) and d)), which is all the more interesting when testing H_0.

Table 1 summarizes the empirical level of the test depending on the situation. Globally, the test results are rather satisfactory, since empirical levels roughly equal 5% whatever the context. Worst cases correspond to situations where at least one mixture component has low weight. This is not surprising, since it is tricky to get accurate estimates of the mixture weights in this situation, thus impacting the quality of the test because very few observations relates to the component density to test. Despite being sensitive to the component weights, our test procedure does not seem to be highly sensitive to the number of observations. As expected, empirical levels tend to asymptotically decrease.

Let us now study the level of the test, based on the Patra and Sen [24] estimator of the weights, when working with other supports and other distributions. In this view, Tab. 2 shows the results for distributions on various supports; namely $\mathbb{R}^+$, $\mathbb{N}$ and $[0, 1]$. The idea is to check whether using different polynomials when decomposing the mixture densities affect the quality of our procedure. For one given

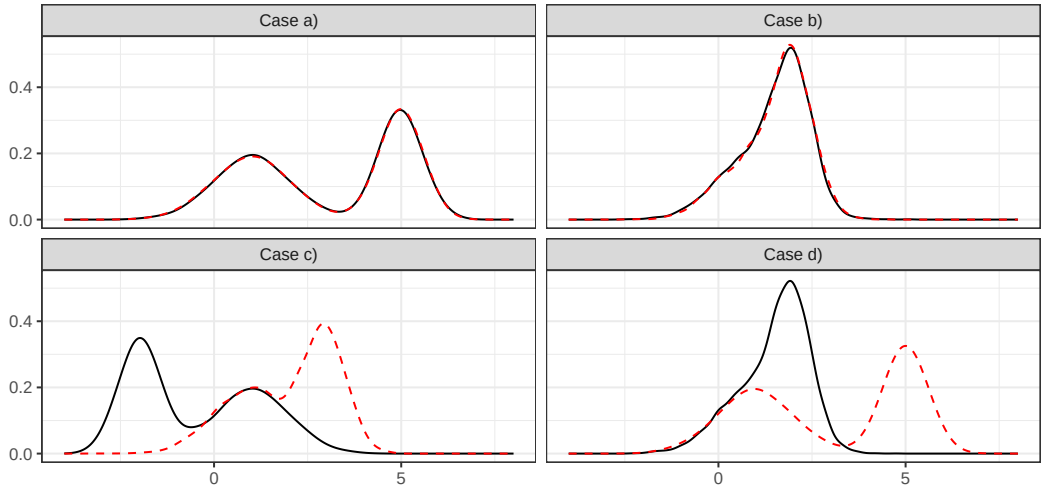

Figure 1: Under H_0: densities of X (solid) and Y (dashed) in various settings: a) $g_1 = g_2$ with g_1 far from (ff) f; b) $g_1 = g_2$ with g_1 close to (ct) f; c) $g_1 \neq g_2$ with g_1 ff g_2 and g_1, g_2 ff f; d) $g_1 \neq g_2$ with g_1 ct f and g_2 ff f.

Table 1: Empirical level (in %) of the test with two-component Gaussian mixtures in various settings. Mixture parameters are listed in Tab. 7, see Appendix C.1.

			Case a)			Case b)			Case c)			Case d)		
			p_2			p_2			p_2			p_2		
			0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7
$n = 1,000$	p_1	0.1	4	8	2	2	6	4	5	7	7	7	5	3
		0.25	3	7	6	5	3	6	2	7	4	6	3	9
		0.7	4	3	5	5	3	9	5	5	6	13	6	6
$n = 4,000$	p_1	0.1	11	6	6	6	6	8	1	4	5	7	5	7
		0.25	7	5	5	4	7	7	1	8	4	1	6	4
		0.7	5	3	4	3	6	3	7	6	5	8	6	5
$n = 10,000$	p_1	0.1	5	6	9	4	7	7	8	3	5	9	3	5
		0.25	3	4	4	6	3	3	4	8	7	4	3	9
		0.7	4	3	4	8	3	5	5	7	1	7	6	4

support, Tab. 2 stores the worst case (in terms of empirical level) associated to the aforementioned situations a), b), c) and d). This leads to an interesting conclusion: the worst case stands for situation d) whatever the support. As in case b), the mixture weight p_1 (see the distribution of X, plain curve in case d), in Fig. 1) may be tricky to estimate. However, contrary to case b), there is no compensation effect here since p_2 is very likely to be well estimated. This asymmetric behavior, when estimating the mixture parameters related to X and Y, deteriorates the quality of the test. On the other hand, this effect tends to vanish when increasing the sample size, which is rather reassuring. Provided that there are enough observations, the type of distribution involved in the mixture densities does not seem to affect the test quality (note that other distributions than those presented in this paper were also tested, with similar results).

Nevertheless, the bad results sometimes observed when the sample size is too small, are likely to

come from the choice of the distributions themselves (see the associated mixture parameters in Tab. 8 of Appendix C.1). Indeed, it seems that high variances of the components distributions cause troubles when estimating the weights p_1 and p_2. Again, this issue diminishes when increasing the sample size, thanks to the asymptotic properties of the estimators used in our procedure.

Table 2: Empirical level (in %) of the test corresponding to case d), identified as the situation providing the worst results whatever the support. Parameters are given in Tab. 8 of Appendix C.1.

Support:			$\mathbb{R}^{+\star}$			$\mathbb{N}$			$[0,1]$		
			p_2			p_2			p_2		
			0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7
$n = 1,000$	p_1	0.1	12	10	15	6	8	9	6	10	9
		0.25	13	19	10	8	9	9	9	8	4
		0.7	22	26	21	7	7	8	5	4	4
$n = 4,000$	p_1	0.1	8	3	8	9	12	3	4	2	5
		0.25	3	2	1	12	8	3	4	2	7
		0.7	18	12	11	5	11	6	4	2	5
$n = 10,000$	p_1	0.1	5	8	6	6	3	6	5	4	3
		0.25	3	6	5	4	3	6	4	2	7
		0.7	7	3	6	2	1	4	7	5	7

5.2 Test empirical powers

We now evaluate the ability of our test to detect departures from the null hypothesis. As a starting point, consider the situations where f_1 and f_2 belong to same distribution family, but have different moments. In our study, the difference can originate from the expectation (case e)) or the variance. When the variances of f_1 and f_2 differ, we are interested in two separate cases: either the difference is big (case f)) or small (case g)). Lastly, we analyse the behaviour of the test when f_1 and f_2 belong to distribution families, with same two first order moments (case h)).

Similarly to Section 5.1, Tab. 3 provides the empirical power of the test related to Gaussian mixtures

Table 3: Empirical powers of the test in two-component Gaussian mixtures in various settings. Mixture parameters are listed in Tab. 9 of Appendix C.2).

			Case e) $\mathbb{E}[f_1] \neq \mathbb{E}[f_2]$			Case f) $\mathbb{V}(f_1) \neq \mathbb{V}(f_2)$			Case g) $\mathbb{V}(f_1) \simeq \mathbb{V}(f_2)$			Case h) $\mathcal{N}(\mu,\sigma)$ vs $\mathcal{L}(\theta,\nu)$		
			p_2			p_2			p_2			p_2		
			0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7
$n = 1,000$	p_1	0.1	51	62	70	41	59	72	13	6	12	4	2	8
		0.25	69	93	100	62	97	100	9	18	34	2	6	6
		0.7	83	100	100	83	100	100	14	41	97	3	5	5
$n = 4,000$	p_1	0.1	100	99	100	96	100	100	17	21	39	1	7	2
		0.25	100	100	100	100	100	100	28	69	93	2	6	3
		0.7	100	100	100	100	100	100	36	96	100	1	3	5
$n = 10,000$	p_1	0.1	100	100	100	100	100	100	31	58	68	3	5	5
		0.25	100	100	100	100	100	100	63	98	100	8	5	4
		0.7	100	100	100	100	100	100	88	100	100	8	10	4

Table 4: Empirical power of the test ($n = 1,000$); depending on the support, the weights and the order of the moment differentiating f_1 from f_2. Parameters are stored in Tab. 10, Appendix C.2.

			Case e)			Case f)			Case g)			Case h)		
			p_2			p_2			p_2			p_2		
			0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7	0.1	0.25	0.7
$\mathbb{R}^+$	p_1	0.1	58	75	85	29	28	33	9	7	5	8	7	9
		0.25	82	100	100	60	92	98	3	6	7	11	5	10
		0.7	84	100	100	61	100	100	8	14	17	9	4	6
$\mathbb{N}$	p_1	0.1	22	26	27	14	22	43	5	11	7	14	5	5
		0.25	30	73	94	18	76	96	6	6	19	5	10	6
		0.7	43	97	100	16	91	100	12	11	78	12	4	14
$[0, 1]$	p_1	0.1	16	22	30	19	45	43	6	5	10	10	7	8
		0.25	26	72	95	45	100	100	16	33	62	9	20	20
		0.7	37	90	100	38	100	100	7	60	100	12	34	83

in the four aforementioned cases. Then, Tab. 4 summarizes the results when making the support and distribution change. As it can be noticed, the power of the test is very strongly influenced by the number of observations, and is much more sensitive to the sample size than when considering empirical levels performances. Indeed, detecting some differences between f_1 and f_2 sometimes requires a lot of data and is more conservative. Concretely, as soon as the difference lies in the skewness, the kurtosis, or higher order moments of the distributions, it becomes very hard to get high powers (except when the size of the data becomes huge). As an example, our trials show that at least $25,000$ observations are needed to reach acceptable powers (70%) in case h) of Tab. 3 (with $p_1 = p_2 = 0.1$ and other parameters listed in Tab. 9 of Appendix C.2). Of course, for the same reason, the weights also play a key role. Indeed, they somewhat represent the exposure of the unknown component: the bigger they are, the higher the power of the test is. Table 4 shows that the same conclusions apply when changing the support and the distributions of the mixture components. Differences are very likely to come from the choice of the parameters and the types of the component distributions (see Tab. 10 in Appendix C.2), but results are in line with our expectations. Moreover, Tab. 4 gives an additional information about some kind of lower bound for the powers in the cases studied. Indeed, the number n of observations was set to $1,000$, which is clearly the worst case in our simulations.

Apart from these statements, those simulations also enable to verify empirically Proposition 2. Under the alternative hypothesis, the selected order of the test statistic should be greater or equal than the moment order differentiating them. Among the 100 times the test was performed (with $n = 10,000$), more than 80% of the tests have selected the right order following the penalization rule (13); i.e. $k = 1$ in case e), $k = 2$ in case f) or g), and $k \geq 3$ in case h). Let us mention that more than 90% of the tests selected the first rank ($k = 1$) in the decomposition when testing under the null hypothesis. This is good news that confirms our theoretic results.

6 Real datasets applications

6.1 Mortality Dataset

In order to understand the heterogeneity of insured populations, we propose an application of the hypothesis testing developed so far on real-world datasets of insured populations gathering information on

the age-at-death random variable. These datasets come from studies conducted by the *French Institute of Actuaries* and cover the period 2007-2011.

Here, we consider three populations of female individuals holding death guarantees. The portfolios characteristics are reported in the left panel of Tab. 5 and the pdfs of the age-at-death for each population are depicted in Fig. 2. In Tab. 5, we report the life expectancy, i.e. $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, estimated over the populations. It is stable over the three considered populations and would suggest a comparable mortality profiles for the three of them. However, when looking at the pdfs in Fig. 2 the comparison between these three populations is not straightforward.

Table 5: Left panel: The characteristics of the age-at-death random variable over the three portfolios as well as the estimated proportion p of the unknown component (bottom panel). Right panel: The test statistics (upper triangle) and the test p-value (lower triangle) for the insured population age-at-death unknown mixture component.

	Size (n)	Life expectancy	Proportion $\widehat{p}$	P1	P2	P3
P1	1,251	75.42	0.4603	—	23.28	0.717
P2	7,356	74.91	0.7003	1.4e-06	—	18.48
P3	3,456	75.56	0.6281	0.397	1.7e-05	—

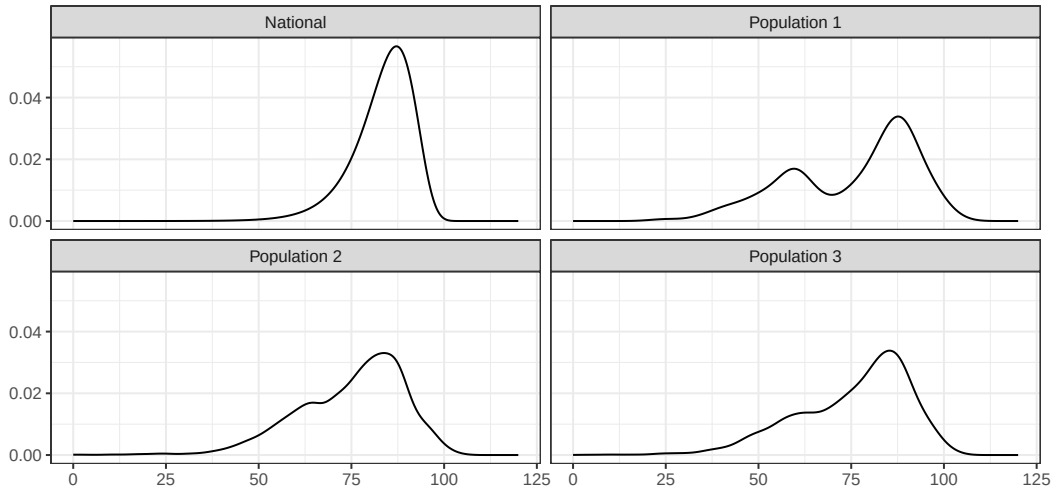

Figure 2: The probability density estimations of the age-at-death for the French (female) national populations together with the considered populations (portfolios).

Before proceeding to the hypothesis testing, we need to specify the known densities in model (2). In fact, each population should encompass individuals whose mortality profile is described by the (mother) national one. Formally, we assume that the densities g_1 and g_2 are identical and calibrated using the national French mortality observations. Although, various parametric forms are usually used in demographic application, we choose the well-celebrated Gompertz law [20], i.e. $g(x) = b\exp(ax)\exp(-b/a(\exp(ax) - 1))$, with shape a and rate b. We calibrate these parameters on the basis of the mortality records of the French female population over the aforementioned period using data of the Human Mortality Database

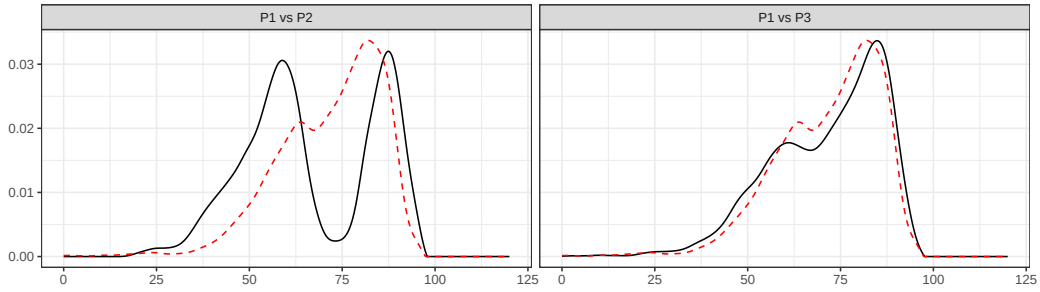

Figure 3: Decontaminated density estimations of the age-at-death. The left panel depicts the pdf of the population P1 (dashed) against P2 (solid) and the right panel compares the pdf of P1 (dashed) to P3 (solid).

(HMD)[1]. This gives the following estimated shape and rate $a = 0.125$ and $b = 2.182 \times 10^{-6}$. In Fig. 2, we depicted the Gompertz law for the national population (top left panel). Hence, using this density, we reported the estimated proportion $\widehat{p}$ in model (1) for each population using the estimator by Patra and Sen [24]. In the right panel of Tab. 5, we summarize the outputs of the hypothesis testing for pairs of portfolios. In the upper triangle we reported the test statistics in (11) as well as the p-values in the lower triangle. The heterogeneous part of the portfolios is tested on the pairs and the p-values of the test does accept the equality for populations 1 and 3, suggesting a similar behaviour of the heterogeneous component, see also the decontaminated (unknown) densities f_i in Fig. 3. Here, we use a natural estimator of the density functions

$$\widehat{f}_i(x) = \frac{\widehat{h}_i(x) - (1 - \widehat{p}_i)g_i(x)}{\widehat{p}_i}, \quad i = 1, 2, \tag{16}$$

where $\widehat{h}_i$ denotes a classical kernel density estimator of h_i. In this figure, we can see that the remaining subgroup in each portfolio is also a mixture of two (or more) components, which is coherent with the demographic literature [20]. In order to understand the outputs, we should look at the main determinants that can drive the heterogeneity of such populations. In fact, a significant factor that can explain this is the level of underwriting, which depends on the insured sums. For populations 1 and 3, a medical exam at underwriting took place as the insured sums are relatively high. We should note that this cannot be the sole explanation for the comparable behaviour of the two populations as the test suggested. Indeed, these portfolios come from two different insurers, and the underwriting strategy and level is obviously not the same from one to another. From a risk management point of view, a re-insurer can benefit from a mitigation effect by grouping these two populations in order to reduce the uncertainty, together with the cost of capital.

6.2 Galaxies Dataset

We consider here two datasets from the SIMBAD Astronomical Database (Observatoire Astronomique de Strasbourg). They are made of stars heliocentric velocities evolution measurements coming from two dwarf spheroidal (dSph) galaxies: Carina and Sextans. These dSph galaxies are low luminosity galaxies that are companions of the Milky Way. In Lokas [18] it is noted that these two dwarfs are highly dark matter dominated and similar in this respect to the so called Draco dwarf.

[1]The dataset was downloaded from http://www.mortality.org on September 2019.

These models are contamination models in the sense that both stars measurements from Carina and Sextans are mixed with stars from the Milky Way in the stellar landscape. Since the Milky Way is very largely observed, see Robin *et al.* [27], it is commonly accepted that its heliocentric velocity (HV) can be viewed as a random variable with known probability density function $g = g_1 = g_2$ as in (2). Then we can assume that our two datasets are drawn from (2), where f_1 and f_2 stand respectively for the unknown density of Carina and Sextans galaxies. Figure 4 shows the probability densities estimations of the heliocentric velocity associated to Milky Way and its companions Carina and Sextans. It is based on $n = 170,601$ observations without contamination for Milky Way obtained by Magellan telescope [30], $n_1 = 2,400$ contaminated observations from Carina and $n_2 = 1,488$ contaminated observations from Sextans. It is similar to the Fig. 1 shown in [30] where Gaussian assumptions on the densities were done. Such assumptions are not necessary with our method, and only the knowledge of the moments of the Milky Way HV allows us to implement our test.

We aim here to test if both Carina and Sextans heliocentric velocities have the same distribution (H_0). Using the semiparametric estimation procedure in Bordes and Vandekherkove [3] we obtain $\widehat{p}_1 = 0.4446$ and $\widehat{p}_2 = 0.5693$, which means that 44.46% of the Carina HV and 56.93% of the Sextans HV are captured through these datasets. In addition the corresponding location estimators for Carina and Sextans are respectively: $\widehat{\mu}_1 = 224.5$ and $\widehat{\mu}_2 = 226.4$. Our testing procedure selects the first rank, that is $S(n_1, n_2) = 1$, and provide a test statistic value $\widehat{T}_1 = 0.02567$ with a p-value equal to 0.87. The meaning of this is that there is no reason to reject the null hypothesis, or more practically, that we can reasonably decide that the Carina and Sextans HV distributions are similar. Note that this conclusion can be visually validated by looking at the decontaminated Carina and Sextans HV densities (see Fig. 3), where only very slight bumps on the left tails do not fit exactly. These tiny differences are possibly artefacts due to the inversion formula (16) applied on the approximate $\widehat{p}_i$ and kernel density estimates $\widehat{h}_i$, $i = 1, 2$.

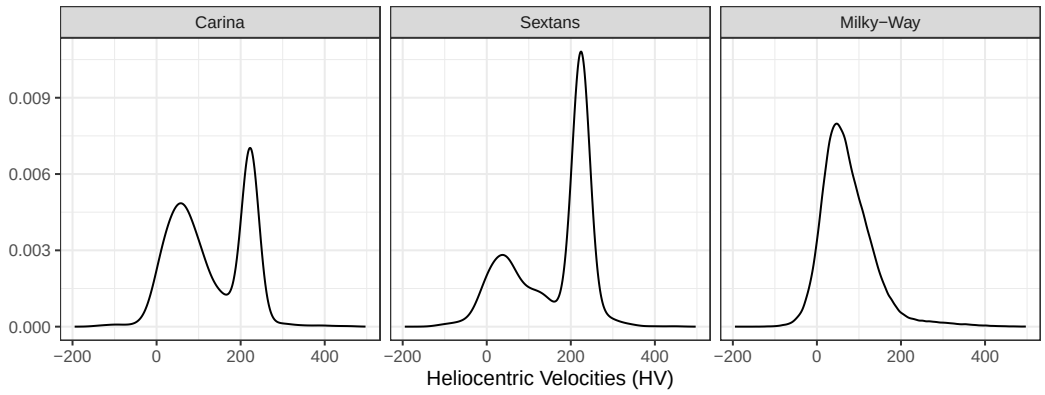

Figure 4: The probability density estimations of the heliocentric velocities of the Carina (contaminated), Sextans (contaminated) and Milky Way galaxies.

Conclusion

In this work we both theoretically and numerically addressed the two-sample comparison testing of the unknown component in the contamination model (2). We implemented our methodology, with satisfactory results, on a large range of situations, as summarized in Tables 1- 4, including Gaussian distributions but also more challenging distributions supported on $\mathbb{R}^+$, $\mathbb{N}$ or $[0, 1]$ which are considered as very non-standard

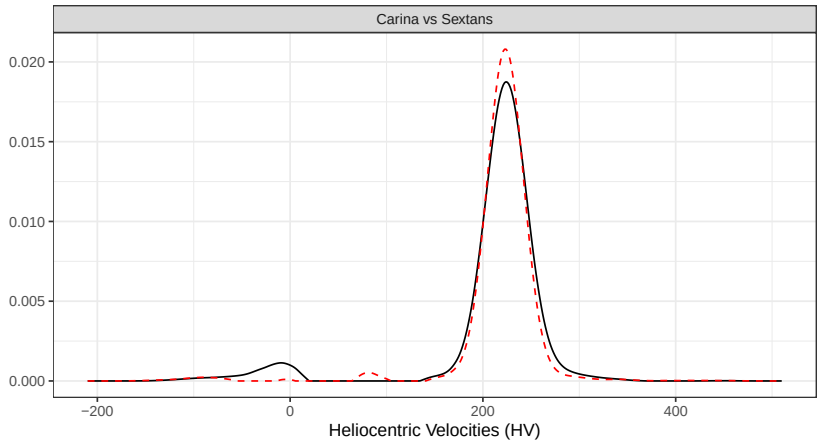

Figure 5: Decontaminated density estimations of the heliocentric velocities of the Carina (dashed) and Sextans (solid) galaxies using Bordes and Vandekerkhove [3] with $(\widehat{p}_1, \widehat{p}_2) = (0.44, 0.56)$.

in the semiparametric mixture models literature. We then used our testing procedure on two real-life problems: i) age-at-death distribution testing in actuarial science and ii) heliocentric velocity comparison for Carina and Sextan galaxies. These two real datasets applications successfully demonstrate the utility and interpretability of our testing procedure validated by the features comparison of the decontaminated densities, see Fig. 5.

We think that this work could be extended in many interesting ways. First we could consider the case where the two samples are paired, with $n_1 = n_2$, as in Ghattas *et al.* [10]. We could probably adapt our testing procedure to obtain results similar to those in Proposition 2 when Corollary 4 could be extended to the paired case by considering the Central Limit Theorem applied on $n^{-1}\sum_{j=1}^{n}(Q_1(X_j) - Q_1(Y_j))$. This would particularly be interesting for time-varying models consideration. Another interesting problem would be the K-sample version of this procedure, which would enable us to deal with time series applications. Coming back to mortality, this would allow us to consider a longevity model integrating future improvements of mortality rates, for instance the prospective Lee and Carter model [17]. Concerning galaxy data, this would allow us to compare the heliocentric velocities from more than two galaxies simultaneously. Moreover, extensions to censored and truncated data would also be interesting, especially for insurance applications where it is frequent to retrieve incomplete observations. In the same spirit, we could also extend our applications in various fields to compare the heterogeneity of populations when a gold standard is established. All these works are also intended to be included in a package in progress.

References

[1] Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C. and Salhi, Y. (2010). Understanding, modelling and managing longevity risk: key issues and main challenges. *Scand. Actuar. J.*, **3**, 203–231.

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

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

[4] Broët, P., Lewin, A., Richardson, S., Dalmasso, C. and Magdelenat, H. (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, **20**, 2562–2571.

[5] Cai, T.T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Stat.*, **38**, 100–145.

[6] Celisse, A. and Robin, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Stat. Plan. Infer.*, **140**, 3132–3147.

[7] Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM: CBMS-NSF Regional Conf. Ser. Appl. Math.

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

[9] Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates. *Genet. Epidemiol.*, **23**, 70–86.

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

[11] Inglot, T. and Ledwina, T. (2006). Data-driven score tests for homoscedastic linear regression model: asymptotic results. *Probab. Math. Stat.* , **26**, 41–61.

[12] Janic-Wróblewska, A. and Ledwina, T. (2000). Data driven rank test for two-sample problem. *Scand. J. Stat.* **27**, 281–297.

[13] 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.

[14] Kallenberg, W.C. and Ledwina, T. (1995). Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann. Stat.*, **23**(5), 1594-1608.

[15] Keyfitz, N. and Littman, G. (1979). Mortality in a heterogeneous population. *Popul. Stud.*, **33**, 333-342.

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

[17] Lee, R.D. and Carter L.R. (1992). Modeling and Forecasting U.S. mortality. *JASA*, **87**, 659–671.

[18] Lokas, E.L. (2009). The mass and velocity anisotropy of the Carina, Fornax, Sculptor and Sextans dwarf spheroidal galaxies. *Mon. Not. R. Astron. Soc.*, 394, 102–106.

[19] 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.

[20] Manton, K.G., Stallard, E. and Vaupel, J. W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *JASA*, **81**, 635-644.

[21] Munk, A., Stockis, J. P., Valeinis, J. and Giese, G. (2011). Neyman smooth goodness-of-fit tests for the marginal distribution of dependent data. *Ann. I. Stat. Math.*, **63**, 939-959.

[22] Neyman, J. (1937). Smooth test for goodness of Fit. *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

[23] Nguyen, V.H. and Matias, C. (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Stat.*, **41**, 1167-1194.

[24] 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.

[25] 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.

[26] Pommeret, D. and Vandekerkhove, P. (2019). Semiparametric density testing in the contamination model. *Electron. J. Stat.*, **13**, 4743-4793.

[27] 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.

[28] Salhi, Y. and Loisel, S. (2017). Basis risk modelling: a cointegration-based approach. *Statistics*, **51**, 205-221.

[29] Walker, M., Mateo, M., Olszewski, E., Sen, B. and Woodroofe M. (2009). Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present. *Astron. J.*, **137**, 3109–3138.

[30] Walker, M.G., Mateo, M., Olszewski, E.W., Gnedin, O.Y., Wang, X., Sen, B. and Woodroofe, M. (2007). Velocity dispersion profiles of seven dwarf spheroidal galaxies. *Astrophys. J.*, **667**, L53–L56.

[31] Xiang, S., Yao, W. and Yang, G. (2018) An overview of Semiparametric Extensions of finite Mixture Models. *Statist. Sci.*, **34**, 391–404.

A Proofs

Proof of Theorem 3. For simplicity matters, let us write $\widetilde{n} = n_1 n_2/(n_1+n_2)$. We need to prove that $\mathbb{P}(S(n_1,n_2) \geq 2)$ vanishes as $n_1, n_2 \to +\infty$. By definition of $S(n_1,n_2)$, using the positivity of $\widehat{T}_1$, we have

$$\begin{aligned}
\mathbb{P}(S(n_1,n_2) \geq 2) &= \mathbb{P}\left(\max_{2\leq k\leq d(n_1,n_2)} \{\widetilde{n}^{s-1}\widehat{T}_k - k\log\widetilde{n}\} \geq \widetilde{n}^{s-1}\widehat{T}_1 - \log\widetilde{n} \right), \\
&= \mathbb{P}\left(\exists k,\ 2\leq k\leq d(n_1,n_2):\ \widetilde{n}^{s-1}\widehat{T}_k - k\log\widetilde{n} \geq \widetilde{n}^{s-1}\widehat{T}_1 - \log\widetilde{n} \right), \\
&\leq \mathbb{P}\left(\exists k,\ 2\leq k\leq d(n_1,n_2):\ \sum_{j=2}^{k} \widetilde{n}^{s}(R_j)^2/\widehat{D}[j] \geq (k-1)\log\widetilde{n} \right), \\
&\leq \mathbb{P}\left(\exists k,\ 2\leq k\leq d(n_1,n_2):\ \widetilde{n}^{s}(R_k)^2 \geq e(n_1,n_2)\log\widetilde{n} \right), \\
&\leq \mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^{s}(R_k)^2 \geq e(n_1,n_2)\log\widetilde{n} \right).
\end{aligned}$$

We decompose now R_k as follows:

$$\begin{aligned} R_k &= \widehat{h}_{1,k}\widehat{p}_2 - \widehat{h}_{2,k}\widehat{p}_1, \\ &= (\widehat{h}_{1,k} - p_1\alpha_{1,k})\widehat{p}_2 - (\widehat{h}_{2,k} - p_2\alpha_{2,k})\widehat{p}_1 + \alpha_{1,k}p_1(\widehat{p}_2 - p_2) + \alpha_{2,k}p_2(\widehat{p}_1 - p_1), \end{aligned}$$

where $\alpha_{i,k} = \int_{\mathbb{R}} Q_k(z)h_i(z)\nu(dz)$. Combining two times the inequality $(a+b)^2 \leq 2(a^2+b^2)$, for all $(a,b) \in \mathbb{R}^2$, with $\mathbb{P}(X^2+Y^2 \geq z) \leq \mathbb{P}(X^2 \geq z/2) + \mathbb{P}(Y^2 \geq z/2)$, for all random variable X and Y, we deduce that

$$\begin{aligned} \mathbb{P}(S(n_1,n_2) \geq 2) \leq\ & \mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s(\widehat{h}_{1,k} - p_1\alpha_{1,k})^2\widehat{p}_2^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right) \\ &+ \mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s(\widehat{h}_{2,k} - p_2\alpha_{2,k})^2\widehat{p}_1^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right) \\ &+ \mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s\alpha_{1,k}^2p_1^2(\widehat{p}_2 - p_2)^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right) \\ &+ \mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s\alpha_{2,k}^2p_2^2(\widehat{p}_1 - p_1)^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right). \end{aligned}$$

We study these four quantities separately. First, by Markov inequality, we obtain

$$\begin{aligned} \mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s(\widehat{h}_{1,k} - p_1\alpha_{1,k})^2\widehat{p}_2^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right) &\leq \frac{16\widetilde{n}^s}{e(n_1,n_2)\log\widetilde{n}} \sum_{k=2}^{d(n_1,n_2)}\sum_{j=1}^{n_1} \frac{\mathbb{V}(Q_k(X_j))}{n_1}, \\ &\leq \frac{16M_1d(n_1,n_2){n_1}^{s-1}}{e(n_1,n_2)\log\widetilde{n}}\left(\frac{n_2}{n_1+n_2}\right)^s, \end{aligned}$$

which tends to zero as n_1, n_2 tend to infinity. Similarly,

$$\mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s(\widehat{h}_{2,k} - p_2\alpha_{2,k})^2\widehat{p}_1^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right) \to 0,$$

as n_1, n_2 tend to infinity. We now consider the two last quantities. Since

$$\begin{aligned} p_i^2\alpha_{i,k}^2 &\leq p_i^2\int_{\mathbb{R}} Q_k(z)^2h_i(z)\nu(dz), \\ &= p_i\left(\int_{\mathbb{R}} Q_k(z)^2f_i(z)\nu(dz) - (1-p_i)\int_{\mathbb{R}} Q_k(z)^2g_i(z)\nu(dz)\right), \\ &\leq \int_{\mathbb{R}} Q_k(z)^2f_i(z)\nu(dz \leq M_1, \end{aligned}$$

we have

$$\mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s \alpha_{1,k}^2 p_1^2 (\widehat{p}_2 - p_2)^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right)$$
$$\leq \mathbb{P}\left((\widehat{p}_2 - p_2)^2 \geq \frac{e(n_1,n_2)\log\widetilde{n}}{16 M_1 d(n_1,n_2)\widetilde{n_2}^s}\left(\frac{n_1}{n_1+n_2}\right)^{-s}\right),$$

which tends to zero since $(\widehat{p}_2 - p_2)^2 = o_{a.s.}(n_2^{1/2+\alpha})$ for all $\alpha > 0$ (see Bordes and Vandekerkhove [3] or Patra and Sen [24]). The same conclusion holds for the last quantity, that is,

$$\mathbb{P}\left(\sum_{k=2}^{d(n_1,n_2)} \widetilde{n}^s \alpha_{2,k}^2 p_2^2 (\widehat{p}_1 - p_1)^2 \geq e(n_1,n_2)\log\widetilde{n}/16\right) \to 0,$$

and we finally get $\mathbb{P}(S(n_1,n_2) \geq 2) \to 0$ as $n_1, n_2 \to +\infty$. □

Proof of Proposition 1. We give the proof under Assumption **(A1)**. The proof under **(A3)** is very similar. Recall that $n_1/n_2 \to a$ as n tends to infinity. From Theorem 3, $T_{S(n_1,n_2)}$ has the same limiting distribution as $\widehat{T}_1$. Combining the independence of $\widehat{p}_1, \widehat{p}_2, \widehat{h}_{1,1}$ and $\widehat{h}_{2,1}$ with their asymptotic normality we have the following convergence in law:

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\begin{pmatrix} \widehat{p}_1 - p_1 \\ \widehat{p}_2 - p_2 \\ \widehat{h}_{1,1} - h_{1,1} \\ \widehat{h}_{2,1} - h_{2,1} \end{pmatrix} \quad \to \quad \mathcal{N}(0, D),$$

with $D = diag((1-a)\ v_1, a\ v_2, (1-a)\ v_3, a\ v_4)$, where v_1 and v_2 are the asymptotic variances of $\widehat{p}_1$ and $\widehat{p}_2$, $v_3 = \mathbb{V}(Q_1(X))$ and $v_4 = \mathbb{V}(Q_1(Y))$. Write $L(x,y,z,w) = y(z - (1-x)g_1 - x(w - (1-y)g_2)$. Under the null, (5) implies that $L(p_1,p_2,h_{1,1},h_{2,1}) = 0$ and then we have

$$\widehat{R}_1 \quad = \quad L(\widehat{p}_1,\widehat{p}_2,\widehat{h}_{1,1},\widehat{h}_{2,1}) = L(\widehat{p}_1,\widehat{p}_2,\widehat{h}_{1,1},\widehat{h}_{2,1}) - L(p_1,p_2,h_{1,1},h_{2,1}).$$

From the delta method we get the following convergence in law

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\Big(L(\widehat{p}_1,\widehat{p}_2,\widehat{h}_{1,1},\widehat{h}_{2,1}) - L(p_1,p_2,h_{1,1},h_{2,1})\Big) \quad \to \quad \mathcal{N}(0, w_1),$$

where $w_1 = V'DV$, with V the gradient vector of L. We obtain

$$\begin{aligned} V' &= (p_2 g_1 - (h_{2,1} - (1-p_2)g_2), -p_1 g_2 - (h_{1,1} - (1-p_1)g_1), p_2, p_1), \\ w_1 &= (p_2 g_1 - (h_{2,1} - (1-p_2)g_2))^2 (1-a)\ v_1 + (p_1 g_2 - (h_{1,1} - (1-p_1)g_1))^2 a\ v_2 + p_2^2 (1-a)\ v_3 + p_1^2 a\ v_4, \end{aligned}$$

that we can estimate, replacing $p_1, p_2, v_1, v_2, v_3, v_4, h_{1,1}, h_{2,1}$ by their consistent estimators. Finally we get

$$\widehat{T}_1 = \frac{n_1 n_2}{n_1+n_2}\widehat{R}_1^2/\widehat{d}_1 \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where $\widehat{d}_1$ is given by (12). □

PROOF OF COROLLARY 4. The proof follows from the asymptotic normality of both estimators $\widehat{p}_1$ and $\widehat{p}_2$ under **(A1)** and **(A3)**. □

PROOF OF PROPOSITION 2. We first prove that under $H_1(q)$ we have $\mathbb{P}(S(n_1,n_2)<q)\to 0$ when n_1 and n_2 tend to infinity. We write $\widetilde{n}=n_1n_2/(n_1+n_2)$ and we distinguish the following two cases.
i) Parametric case. We have for all $k<q$:

$$\begin{aligned}
\mathbb{P}\left(S(n_1,n_2)=k\right) &\leq \mathbb{P}\left(\widehat{T}_k-k\log(\widetilde{n})\geq \widehat{T}_q-q\log(\widetilde{n})\right),\\
&= \mathbb{P}\left(\widehat{T}_q-\widehat{T}_k\leq (q-k)\log(\widetilde{n})\right),\\
&= \mathbb{P}\left(\widetilde{n}\sum_{j=k+1}^{q}\widehat{R}_j^2/\widehat{d}_j\leq (q-k)\log(\widetilde{n})\right),\\
&\leq \mathbb{P}\left(\widetilde{n}\widehat{R}_q^2/\widehat{d}_q\leq (q-k)\log(\widetilde{n})\right),\\
&= \mathbb{P}\left(\frac{\sqrt{\widetilde{n}}|\widehat{R}_q|}{\sqrt{\widehat{d}_q\log(\widetilde{n})}}\leq\sqrt{(q-k)}\right).
\end{aligned}$$

We can decompose

$$\begin{aligned}
\frac{\sqrt{\widetilde{n}}\widehat{R}_q}{\sqrt{\widehat{d}_q\log(\widetilde{n})}} &= \frac{1}{\sqrt{\log(\widetilde{n})}}\times\frac{\sqrt{\widetilde{n}}}{\sqrt{\widehat{d}_q}}\left(\widehat{R}_q-\delta(q)\right)+\frac{\sqrt{\widetilde{n}}}{\widehat{d}_q\log(\widetilde{n})}\delta(q),\\
&= A+B,
\end{aligned}$$

where $\delta(q)$ is defined in (14) and $\widehat{d}_q$ is a consistent estimator given by (12). Mimicking the proof of Proposition 2 we can show that $\sqrt{\widetilde{n}}\left(\widehat{R}_q-\delta(q)\right)$ is asymptotically Gaussian and then A converges to a Dirac at point zero. Moreover B converges to $+\infty$ since $\delta(q)>0$. Then for all $k<q$, we have

$$\mathbb{P}\left(S(n_1,n_2)=k\right)\to 0,$$

along with $T_q>\widetilde{n}\widehat{R}_q^2/\widehat{d}_q\to+\infty$ as n_1,n_2 tend to infinity.
ii) Nonparametric case. Mimicking the parametric case we obtain for all $k<q$:

$$\mathbb{P}\left(S(n_1,n_2)=k\right)\leq\mathbb{P}\left(\frac{\sqrt{\widetilde{n}^s}|\widehat{R}_q|}{\sqrt{\widehat{d}_q\log(\widetilde{n})}}\leq\sqrt{(q-k)}\right),$$

with $0<s<1/2$. Again, we have the following decomposition

$$\begin{aligned}
\frac{\sqrt{\widetilde{n}^s}\widehat{R}_q}{\sqrt{\widehat{d}_q\log(\widetilde{n})}} &= \frac{1}{\sqrt{\widetilde{n}^{1-s}\log(\widetilde{n})}}\times\frac{\sqrt{\widetilde{n}}}{\sqrt{\widehat{d}_q}}\left(\widehat{R}_q-\delta(q)\right)+\frac{\sqrt{\widetilde{n}^s}}{\widehat{d}_q\log(\widetilde{n})}\delta(q).\\
&= A'+B'.
\end{aligned}$$

As previously, the random variable A' converges to a Dirac at point zero and the random variable B' converges to $+\infty$ since $\delta(q)>0$. Then for all $k<q$, we have

$$\mathbb{P}\left(S(n_1,n_2)=k\right)\quad\to\quad 0.$$

Finally $T_q>\widetilde{n}\widehat{R}_q^2/\widehat{d}_q\to+\infty$ as n_1,n_2 tend to infinity. □

B Behaviour of the test statistics using Patra and Sen estimator

In addition to the fact that the estimator by Patra and Sen [24] is based on non-restrictive assumptions, we highlight here the interest to use it instead of the classical parametric approach (i.e. the EM algorithm) when fitting non-standard finite mixture models (like in our simulation study). Indeed, although this estimator tends to slightly underestimate the true parameter in practice, it is consistent in all the settings studied in this paper. Moreover, this estimator is adapted to study real datasets, where the true distribution of some components is unknown and thus requires a semiparametric approach.

Under H_0, consider that X and Y have the same mixture density:

$$h_1(x) = h_2(x) = (1-p)g(x) + pf(x) = 0.75\, g(x) + 0.25\, f(x), \quad x \in \mathbb{R},$$

where $g \sim \mathcal{N}(0,1)$, and $f(x) = 0.5\, k(x) + 0.5\, l(x)$ with $k \sim \mathcal{N}(2, 0.5)$ and $l \sim \mathcal{N}(6, 0.5)$.
On the contrary, under the alternative hypothesis H_1, consider the following framework:

$$\begin{cases} h_1(x) = 0.75\, g(x) + 0.25\, f(x), & x \in \mathbb{R}, \\ h_2(x) = 0.75\, g(x) + 0.25\, f_2(x), & x \in \mathbb{R}, \end{cases}$$

where $f_2 \sim \mathcal{N}(6, 0.5)$, and the other component densities were defined previously.
Such mixture densities are depicted in Fig. 6.

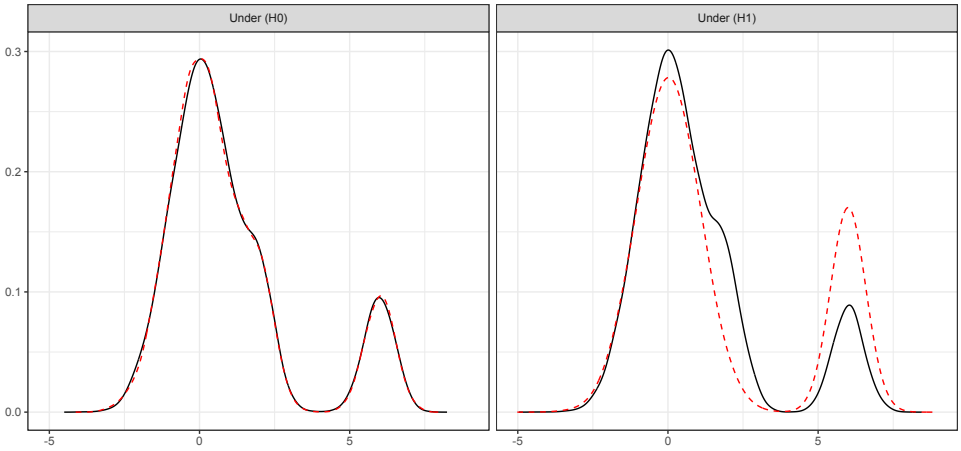

Figure 6: Mixture densities of X (solid) and Y (dashed), under H_0 (left panel) and under H_1 (right panel).

Performing 100 times the testing procedure with $n = 20,000$ observations, we compare the estimates of the weight p given by both the EM algorithm and the Patra and Sen estimator. Table 6 sums up the results obtained, and gives meaningful conclusions. In particular, under H_0, the empirical level of the test does not seem to be affected by the bad estimation of p given by the EM algorithm. On average, $\widehat{p}^{EM} \simeq 13.5\%$, which means that this three-component mixture was understood as a two-component one. More precisely, the second component ($\mathcal{N}(2, 0.5)$) has been "merged" with the first one ($\mathcal{N}(0,1)$). This issue occurs when estimating the parameters related to the two densities of X and Y, and thus some compensation operates when making the difference that defines the test statistics (10). However, in the case of the alternative hypothesis, such a compensation does not apply. Indeed, under H_1, the EM algorithm is expected to provide bad estimates of p for the distribution of X, but good results for the

estimation of the parameters related to the distribution of Y. The test has therefore a very low power when using the EM estimators in this case, since the third component density of X is understood as the second component of the mixture distribution, which is to be compared with the second component density of Y. Of course f and f_2 strongly differ in reality, which is easily detected by our test procedure based in Patra and Sen estimator, but is not seen when using the parametric estimation procedure.

Table 6: Empirical level and power of the test depending on the estimator used for the weight p.

	EM algorithm		Patra and Sen	
	Under H_0	Under H_1	Under H_0	Under H_1
Empirical level (H_0) / power (H_1)	8	4	6	100
Mean of $\widehat{p}_1$	0.1356	0.1255	0.249	0.2463
Median of $\widehat{p}_1$	0.1255	0.1252	0.249	0.2465
Mean of $\widehat{p}_2$	0.1394	0.2502	0.245	0.2513
Median of $\widehat{p}_2$	0.1248	0.2508	0.244	0.251
Mostly selected $S(n_1, n_2)$	1	1	1	10

Finally, some of the theoretic results given in this paper rely on the asymptotic normality of the estimator of p_i ($i = 1, 2$). Using Bordes and Vandekerkhove, this is guaranteed. However, there is no such result when using the estimator by Patra and Sen. We thus decided to plot in Fig. 7 the behaviour of the centered and scaled version of this estimator, based on $2,000$ bootstrap samples (with $n = 1,000$ observations, still within the same framework as above). Despite the fact that the asymptotic normality does not hold for this estimator, its behaviour is close to a Gaussian one, which explains why our test procedure remains powerful using this estimator in practice.

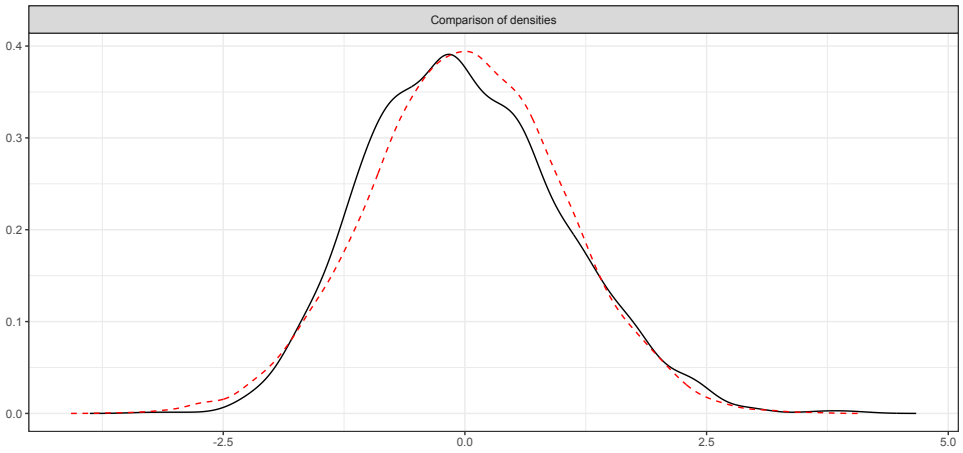

Figure 7: Density of $\widehat{p}_1$ when using Patra and Sen estimator, based on $2,000$ bootstrap samples.

C Additional tables for Monte-Carlo experiments

C.1 Empirical levels

Table 7: Parameters corresponding to Fig. 1.

	Case a)	Case b)	Case c)	Case d)
f	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$
g_1	$\mathcal{N}(5,0.5)$	$\mathcal{N}(2,0.5)$	$\mathcal{N}(-2,0.5)$	$\mathcal{N}(2,0.5)$
g_2	$\mathcal{N}(5,0.5)$	$\mathcal{N}(2,0.5)$	$\mathcal{N}(3,0.5)$	$\mathcal{N}(5,0.5)$
p_1	50%	50%	50%	50%
p_2	50%	50%	50%	50%

Table 8: Parameters corresponding to Tab. 2, in case d). Notations used for the distributions: $\mathcal{G}$ = Gamma, $\mathcal{E}$ = Exponential, $\mathcal{P}$ = Poisson, $\mathcal{BN}$ = Negative Binomial, and $\mathcal{U}$ = Uniform.

	Support $\mathbb{R}^+$			Support $\mathbb{N}$			Support $[0,1]$		
	distribution	$\mathbb{E}$	$\mathbb{V}ar$	distribution	$\mathbb{E}$	$\mathbb{V}ar$	distribution	$\mathbb{E}$	$\mathbb{V}ar$
f	$\mathcal{G}(16,4)$	4	1	$\mathcal{BN}(1,10)$	1	1.1	$\mathcal{B}eta(1.2,5)$	0.2	0.02
g_1	$\mathcal{E}(1/4)$	4	16	$\mathcal{P}(1)$	1	1	$\mathcal{U}(0,0.4)$	0.2	0.013
g_2	$\mathcal{E}(2)$	0.5	0.25	$\mathcal{P}(4)$	4	4	$\mathcal{U}(0.05,1)$	0.55	0.075

C.2 Empirical powers

Table 9: Parameters corresponding to Tab. 3, for cases e), f), g) and h).

	e) $\neq$ means	f) very $\neq$ variances	g) $\neq$ variances	h) $\neq$ distributions
f_1	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	$\mathcal{N}(1,\sqrt{2})$
g_1	$\mathcal{N}(5,0.5)$	$\mathcal{N}(5,0.5)$	$\mathcal{N}(5,0.5)$	$\mathcal{N}(5,0.5)$
f_2	$\mathcal{N}(2,1)$	$\mathcal{N}(1,3)$	$\mathcal{N}(1,\sqrt{2})$	$\mathcal{L}aplace(1,1)$
g_2	$\mathcal{N}(5,0.5)$	$\mathcal{N}(5,0.5)$	$\mathcal{N}(5,0.5)$	$\mathcal{N}(5,0.5)$

Table 10: Parameters corresponding to Tab. 4. Notations: $\mathcal{G}$ = Gamma, $\mathcal{E}$ = Exponential, $\mathcal{P}$ = Poisson, $\mathcal{BN}$ = Negative Binomial, and $\mathcal{U}$ = Uniform, $\mathcal{L}og\mathcal{N}$ = Logit Normal, $\mathcal{G}o$ = Gompertz.

	Support $\mathbb{R}^+$		Support $\mathbb{N}$		Support $[0,1]$	
	e)	f)	e)	f)	e)	f)
f_1	$\mathcal{G}(16,4)$	$\mathcal{G}(8,2)$	$\mathcal{BN}(1,10)$	$\mathcal{BN}(2,10)$	$\mathcal{B}eta(0.8,5)$	$\mathcal{B}eta(12,50)$
g_1	$\mathcal{E}(1/1.1)$	$\mathcal{E}(1/1.1)$	$\mathcal{P}(5)$	$\mathcal{P}(5)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$
f_2	$\mathcal{G}(16,5)$	$\mathcal{N}(32,8)$	$\mathcal{BN}(2,10)$	$\mathcal{BN}(2,0.5)$	$\mathcal{B}eta(1.2,5)$	$\mathcal{B}eta(1.2,5)$
g_2	$\mathcal{E}(1/1.1)$	$\mathcal{E}(1/1.1)$	$\mathcal{P}(5)$	$\mathcal{P}(5)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$
	g)	h)	g)	h)	g)	h)
f_1	$\mathcal{G}(8,2)$	$\mathcal{G}(1.47,0.56)$	$\mathcal{BN}(2,10)$	$\mathcal{BN}(3,100)$	$\mathcal{B}eta(1.2,5)$	$\mathcal{B}eta(5,2)$
g_1	$\mathcal{E}(1/1.1)$	$\mathcal{E}(1/1.1)$	$\mathcal{P}(5)$	$\mathcal{P}(3)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$
f_2	$\mathcal{G}(10,2.5)$	$\mathcal{G}o(0.1,0.3)$	$\mathcal{BN}(2,2)$	$\mathcal{B}(50,0.06)$	$\mathcal{B}eta(2.4,10)$	$\mathcal{L}og\mathcal{N}(0.9,0.8)$
g_2	$\mathcal{E}(1/1.1)$	$\mathcal{E}(1/1.1)$	$\mathcal{P}(5)$	$\mathcal{P}(3)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0,1)$