Two-sample contamination model test

Xavier Milhaud, Yahia Salhi, Pierre Vandekerkhove, Denys Pommeret

Submitted to Bernoulli

Two-sample contamination model test

XAVIER MILHAUD[1,a], DENYS POMMERET[2,b], YAHIA SALHI[3,c] and PIERRE VANDEKERKHOVE[4,d]

[1]*Aix-Marseille University, Centrale Marseille, 13013 Marseille, France,* [a]*xavier.milhaud@univ-amu.fr*
[2]*Aix-Marseille University, Campus de Luminy, 13288 Marseille cedex 9, France,* [b]*denys.pommeret@univ-amu.fr*
[3]*Univ Lyon, UCBL, ISFA LSAF EA2429, F-69007, Lyon, France,* [c]*yahia.salhi@univ-lyon1.fr*
[4]*Université Gustave Eiffel, LAMA (UMR 8050), 77420 Champs-sur-Marne, France,*
[d]*pierre.vandekerkhove@univ-eiffel.fr*

In this paper, we consider two-component mixture models having one single known component. This type of model is of particular interest when a known random phenomenon is contaminated by an unknown random effect. We propose in this setup to test the equality in distribution of the unknown random sources involved in two separate samples generated from such a model. For this purpose, we introduce the so-called IBM (Inversion-Best Matching) approach resulting in a tuning-free relaxed semiparametric Cramér-von Mises type two-sample test requiring minimal assumptions about the unknown distributions. The accomplishment of our work lies in the fact that we establish, under some natural and interpretable mutual-identifiability conditions specific to the two-sample case, a functional central limit theorem about the proportion parameters along with the unknown cumulative distribution functions of the model. An intensive numerical study is carried out from a large range of simulation setups to illustrate the asymptotic properties of our test. Finally, our testing procedure, implemented in the `admix` `R package`, is applied to a real-life situation through pairwise post COVID-19 mortality excess profile testing across a panel of European countries.

MSC2020 subject classifications: Primary 62G05; 62G20; secondary 62E10
Keywords: Cramér-von Mises; finite mixture model; mortality excess; semiparametric estimation

1. Introduction

The aim of this paper lies in a better understanding of random phenomena resulting from a contamination of a known random source by an unknown/unexpected random effect using advanced statistical models that enable to deal with unobserved heterogeneity. Despite the wide generality of the results we will show later on, we propose to present, as an introductory example, the starting motivation of this work in connection with the COVID-19 pandemic. For more than two years now, the COVID-19 pandemic has affected most populations from all around the world, in very different and unexpected ways. As observed in Kontis et al. (2020), on top of the direct infectious impact of the pandemic on populations, the indirect effects, acting through social, economic, environmental and healthcare pathways are also very substantial. Indirect effects, which can be negative or positive, include for instance denied or delayed disease prevention and medical procedures for acute and chronic conditions; loss of jobs and income; disruption of social networks; increase of self-harm and crime; changes in quantity and quality of food and use of tobacco and other drugs, other injuries and air quality resulting from modified social contacts, mobility and transportation. As it can be seen in Figure 1 (left panel), the all cause-of-death records in European countries over the first 25 weeks strikingly shot up in 2020 (black line), as compared to the historical corresponding mortality records exhibited over year 2019 (greyed curve). On top of this, the mortality profile itself has been largely distorted by the sanitary crisis, as illustrated on the right-hand side of Figure 1. Basically, elderly people have been more severely impacted than other age classes. Such a phenomenon can be interpreted as a contamination of some well-known/regular random behaviour, described by the regular mortality distribution, and modelled by using a relevant mixture

model. Indeed let us consider an event A indicating whether the COVID-19 pandemic has impacted (directly or indirectly) some given individual, and X the random variable indicating the age category for a deceased person picked at random during the 25 first weeks of 2020. In the context of Figure 1, X is labelled from 1 to 4 (#1=[15-64], #2=[65,74], #3=[75,84], #4=[85,+[). Considering that during the early times of the COVID-19 pandemic, because of the sudden nature of the crisis and the lack of preparation of the countries, all the populations were uniformly (in age) exposed to the virus, we propose to define the cumulative distribution function of X based on the Bayes principle:

$$\begin{aligned} L(k) = P(X \leq k) &= P(\bar{A})P(X \leq k \mid \bar{A}) + P(A)P(X \leq k \mid A) \\ &= (1-p)G(k) + pF(k), \qquad k = 1, \ldots, 4, \end{aligned} \tag{1}$$

where the probability $p = P(A)$ of being impacted by COVID-19 is unknown, the probability of death occurring up to class k given the fact that the person died from a regular cause, *i.e.* $G(k) = P(X \leq k \mid \bar{A})$, is known and represents the regular per age mortality distribution, whereas the probability of death happening before class k given the fact that the person died from consequences of the COVID-19 pandemic, *i.e.* $F(k) = P(X \leq k \mid A)$ is unknown. Given this modelling of the mortality distribution distortion effect due to the pandemic, we propose in this paper to test if some of the European countries included in our panel reacted similarly, up to an impact parameter p, in terms of mortality over ages in the early stage of the pandemic, or equivalently if some countries had similar F components involved in their contamination model (1).

To answer the above problem we propose to include it in a more general framework, which is the univariate semiparametric two-component mixture model with Cumulative Distribution Function (cdf)

$$L(x) = (1-p)G(x) + pF(x), \quad x \in \mathbb{R}, \tag{2}$$

where G is a known cdf, and where the two unknown quantities are the mixture proportion $p \in]0,1[$ and the cdf F, which is not assumed to belong to any parametric family. This model, sometimes called

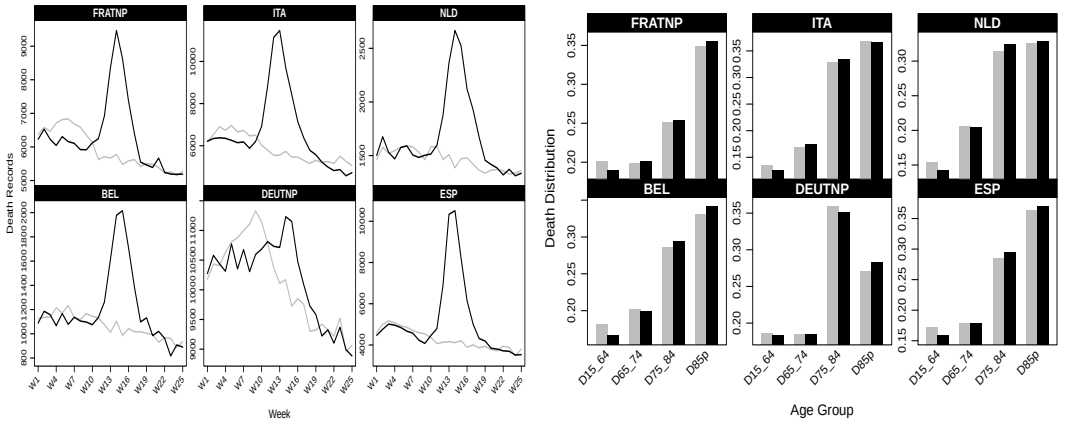

Figure 1. Mortality across six European countries: France, Italy, Netherlands, Belgium, Germany and Spain. Left panel: total death records over the first 25 weeks of 2019 (grey) and 2020 (black curve). Right panel: distribution of the proportion of deaths per age group among all deaths for the first 25 weeks of years 2019 (grey) and 2020 (black).

contamination or *admixture* model, has been widely investigated in the last decades; see for instance Bordes and Vandekerkhove (2010), Nguyen and Matias (2014), Cai and Jin (2010) or Celisse and Robin (2010) among others. On top of the COVID-19 application described earlier, numerous applications of this model can be found in topics such as: i) genetics regarding the analysis of gene expressions from microarray experiments as done in Broët et al. (2004); ii) the false discovery rate problem (used to assess and control multiple error rates such as in Efron and Tibshirani (2002)), see McLachlan, Bean and Jones (2006); iii) astronomy, in which this model arises when observing variables such as metallicity and radial velocity of stars as in Walker et al. (2009); iv) biology to model trees diameters, see Podlaski and Roesch (2014); v) kinetics to model plasma data, see Klingenberg, Pirner and Puppo (2017), vi) genomics to represent populations formed by admixture with known ancestral founding populations as in Chakraborty and Weiss (1988) or Loh et al. (2013), among many other fields of applications. We recommend also the excellent survey on semiparametric mixture models by Xiang, Yao and Yang (2019) to have a panoramic view on this last generation of mixture models.

Consider now that we observe two independent datasets drawn from two separate distributions $L_i(x) = (1-p)G_i(x) + pF_i(x)$, $x \in \mathbb{R}$ and $i = 1, 2$. We propose in this paper to address the following general pairwise testing problem:

$$H_0: \ F_1 \text{ is equal to } F_2 \qquad \text{against} \qquad H_1: \ F_1 \text{ is different from } F_2, \tag{3}$$

without assigning any specific parametric family to these distributions.

In the one-sample case very few works have been proposed to test parametrically the unknown cdf F in (2). In Suesse, Rayner and Thas (2017) a maximum likelihood approach is proposed but without guarantee of convergence of the test under alternatives. In Pommeret and Vandekerkhove (2019) a consistent test is proposed but only under symmetry assumptions reducing the number of possible applications. In the two-sample case, Milhaud et al. (2022) used a semiparametric penalized χ^2-type test based on a $\sqrt{n}$-consistent estimation of the parameters p_1 and p_2, where n stands for the sample size, which again requires symmetry assumptions according to Bordes and Vandekerkhove (2010). In the p-value contamination setup (support over $[0, 1]$), Nguyen and Matias (2014) establish asymptotic efficiency results in two different cases: whether the unknown component F is constant on a set with non-null Lebesgue measure or not. In the first case, the authors exhibit estimators converging at parametric rate and conjecture that no estimator is asymptotically efficient. In the second case the authors prove that the quadratic risk of any estimator does not converge at parametric rate. We could have investigated, under the first case, a testing method based on the available $\sqrt{n}$-convergence result but the attached conditions were too restrictive for the range of applications targeted in our paper (no objective reasons to assume that the contamination phenomenon does not overlap the regular component on a certain interval). On the other hand when considering the very general Patra and Sen (2016, Theorem 3-4) setup, the $\sqrt{n}$-consistency is unfortunately not theoretically achieved which seriously compromises the chances of building a test statistic, grounded on their estimators, with an asymptotically identified distribution. As mentioned earlier, in many applications, including the one we investigate here, symmetry assumptions could also happen to be too restrictive and unrealistic given the nature of the data. In fact, when considering the COVID-19 case, it is easy to figure out that, since older people are much more impacted than the rest of the population, the distribution associated with the F_i's will assign less mass to the youngest age classes than to the older ones (see Figure 7). Therefore, the testing strategy we propose here is very different from the one proposed in Milhaud et al. (2022), which only applies in the symmetric case or from the other aforementioned related works. To overcome the lack of $\sqrt{n}$-consistency under H_0 or H_1 in the Patra and Sen (2016) setup, and get a complete valid asymptotic theory, we decided to rethink from scratch the two-sample testing problem (3). For this purpose, we introduce a new so-called IBM (Inversion-Best Matching) approach resulting in a relaxed semiparametric

Cramér-von Mises type two-sample test requiring very minimal assumptions about the unknown distributions. The accomplishment of our work lies in the fact that we establish a joint functional central limit theorem about the proportion parameters along with the unknown cumulative distribution functions of the model thanks to some mutual-identifiability conditions involving the couples (F_i, G_i), $i = 1, 2$.

Our paper is organized as follows: In Section 2 we describe our testing methodology and pedagogically explain how we built it up. In Section 3 we analyse the two-sample identifiability of our testing problem (H_0 and H_1 separation) and state the basic assumptions ensuring the validity of our method. Let us point out that Section 3 also includes the important Remark 1 in which a practical interpretation of our identifiability conditions is provided. Section 4 is dedicated to technical lemmas and asymptotic results showing the theoretical validity of our testing procedure under the condition $G_1 \neq G_2$, whereas Section 5 is dedicated to a Monte Carlo assessment of our asymptotic results. In Section 6 we investigate the empirical levels and powers of our test while in Section 7 we present an original application in which we perform a pairwise comparison of the mortality excess due to COVID-19 across a panel of European countries. Finally, Section 8 contains a discussion in which we present two further leads of research based on dependent two-sample models: i) we introduce the coordinates independence testing for the unknown component of a multivariate contamination model along with the complete concordance/discordance testing problem arising in z-scores modelling, ii) we introduce the homogeneity testing problem in the so-called blending process (temporal contamination model). Some proofs and technical material are relegated to the Appendix Section of the Supplement. Note that Section D in Appendix is devoted to the non-identifiable situation where $G_1 = G_2$, in which the testing problem (3) can still surprisingly be addressed by using a re-parametrisation trick.

2. Testing problem and methodology

In this paper, the data of interest is made of two independent i.i.d. samples $X_1 = (X_{1,1}, \ldots, X_{1,n_1})$ and $X_2 = (X_{2,1}, \ldots, X_{2,n_2})$ with respective cdfs:

$$\begin{cases} L_1(x) = (1 - p_1)G_1(x) + p_1F_1(x), & x \in \mathbb{R} \\ L_2(x) = (1 - p_2)G_2(x) + p_2F_2(x), & x \in \mathbb{R}, \end{cases} \tag{4}$$

where p_1, p_2 are the unknown mixture proportions and F_1, F_2 are the unknown cdf components. For simplicity matters we suppose that $n = n_1 \leq n_2$ and consider the sample size ratio $\kappa = n_2/n_1 \geq 1$. In this work, similarly to Patra and Sen (2016), we will consider situations where the G_i's and F_i's distributions are: i) absolutely continuous with respect to the Lebesgue measure, supported over $\mathbb{R}$, $\mathbb{R}^+$ or intervals of $\mathbb{R}$; ii) finite discrete or $\mathbb{N}$-discrete distributions such as Binomial or Poisson; iii) a mixture of a discrete and an absolutely continuous distribution. All our results will still be valid in such frameworks. Given the above model, our goal is to tackle the testing problem (3) stated in the Introduction without assigning any specific parametric family to the F_i's.

The basic first idea of our paper consists in noticing, similarly to Bordes, Delmas and Vandekerkhove (2006), that expression (2) can be reinterpreted in order to isolate the cdf of the unknown component, *i.e.*

$$F(x) = \frac{L(x) - (1 - p)G(x)}{p}, \quad x \in \mathbb{R}. \tag{5}$$

Given the previous remark, we introduce the so-called *Inversion step* of our method which is the introduction of two parametric function families:

$$\mathcal{F}_i = \left\{ F_i(x, L_i, p_i) = \frac{L_i(x) - (1 - p_i)G_i(x)}{p_i}, \;\; p_i \in \Theta_i, \;\; x \in \mathbb{R} \right\}, \quad i = 1, 2, \tag{6}$$

to which the true unknown cdfs F_1, F_2 should belong. Indeed by picking, for $i = 1, 2$, the true value of the parameters $p_i^* \in \Theta_i$, we exactly retrieve

$$F_i(x) = F_i(x, L_i, p_i^*), \quad x \in \mathbb{R}. \tag{7}$$

For the sake of clarity and without loss of generality, we suppose $\Theta_1 = \Theta_2 = [\delta_1, \delta_2]$, where $0 < \delta_1 < 1 < \delta_2 < +\infty$ and denote $\theta = (p_1, p_2) \in \Theta = [\delta_1, \delta_2]^2$. Let us notice that the $\mathcal{F}_i$'s are not constrained to contain exclusively cumulative distribution functions and that the parametric space Θ_i associated to the p_i's is not necessarily a $[\delta, 1-\delta]$-type subset, $0 < \delta < 1$, of the natural $]0, 1[$ mixture proportion support. In practice δ_2 is a value greater than one (not too large) in order to manage a tolerance on the empirical contrast optimisation when certain components of the true parameter θ are close to one.

We consider now the discrepancy measure

$$d(\theta) = \int_{\mathbb{R}} D^2(x, L_1, L_2, \theta) dU(x), \tag{8}$$

where

$$D(x, L_1, L_2, \theta) = F_1(x, L_1, p_1) - F_2(x, L_2, p_2), \tag{9}$$

measuring possible departures between the functions $F_1(\cdot, L_1, p_1)$ and $F_2(\cdot, L_2, p_2)$ under the parametric location $\theta = (p_1, p_2) \in \Theta$. For simplicity matters, we will denote hereafter $D(x, \theta) = D(x, L_1, L_2, \theta)$ and $F_i(x, p_i) = F_i(x, L_i, p_i)$, $i = 1, 2$, except when the role of the L_i's is central in our study. The integrating cdf U should ideally be chosen in order to focus (assign a strong probability mass) on domains where the $F_i(\cdot, L_i, p_i^*)$'s clearly depart from each other to help on the final test decision. Nevertheless, since the $F_i(\cdot, L_i, p_i^*)$'s are unknown and the structure of the $F_i(\cdot, L_i, p_i)$'s is constantly changing as the p_i's vary in the parametric space, we propose in practice to consider for U rather flat distributions encompassing the support of the observations. A detailed discussion about this choice for U is carried out in Section F of the Supplement.

It is worth to notice that under H_0, there exist $p_1 = p_1^*$ and $p_2 = p_2^*$ such that $d(\theta^*) = d(p_1^*, p_2^*) = 0$. Suppose now that under some regularity and identifiability-type conditions we could prove that

$$\begin{cases} \arg\min_{\theta \in \Theta} d(\theta) = \theta^* \\ d(\theta^*) = 0, \end{cases} \quad \text{under } H_0, \quad \text{and} \quad \begin{cases} \arg\min_{\theta \in \Theta} d(\theta) = \theta^c \\ d(\theta^c) > 0, \end{cases} \quad \text{under } H_1. \tag{10}$$

Then, we would have, assuming that $0 < \delta_1 \leq \delta$, that under H_1

$$\inf_{\theta \in [\delta, 1-\delta]^2 :\, F_i(\cdot, L_i, p_i) \in \mathcal{F}, i=1,2} d(\theta) \;\geq\; \inf_{\theta \in \Theta} d(\theta) = d(\theta^c) \;>\; 0, \tag{11}$$

where $\mathcal{F}$ denotes the set of all cdfs. Note that the search of the infimum in the left-hand side of (11) matches what we would normally expect in a classical semiparametric estimation problem, *i.e.* mixing proportions in $]\delta, 1-\delta[^2 \subset]0, 1[^2$ and F_i's in the cdfs range $\mathcal{F}$, when the second infimum have much more relaxed constraints, *i.e.* mixing proportions in a compact set Θ of $(\mathbb{R}^+)^2$ embedding $]\delta, 1-\delta[^2$ and no specific constraints on the F_i's, that we claim to be sufficient to solve our two-sample testing problem. This relaxation in the optimisation problem allows us to by-pass what was the blocking point to achieve the $\sqrt{n}$-consistency of the estimator $\widehat{\mathsf{p}}_n$ in Patra and Sen (2016), see Section 3 for further details. As illustrated in Figure 2, we basically face two types of situations:

i) there exists a local minima of $d(\theta)$, θ^* under H_0 or θ^c under H_1, in the interior of Θ a close envelop of the natural parametric space $]0, 1[^2$, and then the testing problem is non-trivial and should be addressed,

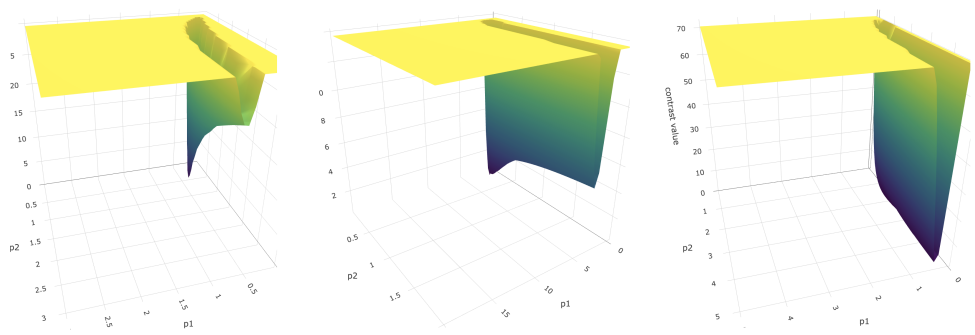

Figure 2. Examples of $d(\theta)$-surface for $\theta = (p_1, p_2)$ varying over Θ. From left to right: under H_0, under H_1 with a minimiser inside Θ a close envelope of the natural parametric space $]0,1[^2$, and under H_1 with a minimiser very likely to be far from the interval $]0,1[^2$.

ii) the optimisation of $d(\theta)$ shows that we bump into the boundaries of the parametric space, *i.e.* at least one of the component of θ^c is equal to δ_2 because the only/main way to reduce the contrast $d(\theta)$ is to make θ large, and then the testing problem is not even worth to be addressed because there is no "reasonable" $\theta^c = (p_1^c, p_2^c)$ close to the probability weights domain $]0,1[^2$ that make $F_1(x, L_1, p_1^c)$ close to $F_2(x, L_2, p_2^c)$.

Now, the empirical estimate $d_n(\cdot)$ of $d(\cdot)$ obtained with replacing the L_i's by the accessible empirical cdfs $\widehat{L}_i(x) = 1/n_i \sum_{j=1}^{n_i} \mathbb{I}_{X_{i,j} \leq x}$ in (8–9), would naturally lead us to find, respectively under H_0 or H_1 the *Best Matching* solution, *i.e.* the true value of the parameter θ^*, respectively the $(\mathcal{F}_1, \mathcal{F}_2)$-models distance minimiser θ^c, by considering:

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} d_n(\theta). \tag{12}$$

We so call *IBM-method* the semiparametric estimation strategy based on the "Inversion" step (6) and the "Best Matching" step (12) between the $\mathcal{F}_1$ and $\mathcal{F}_2$ families to look at the closest they can possibly be.

Next, by analysing closely the statistic $\mathsf{T}_n = n d_n(\widehat{\theta}_n)$, we can show, as stated in Theorem 2, the following asymptotic separation behaviour:

$$n d_n(\widehat{\theta}_n) = \mathsf{U}_n^0 \xrightarrow{\mathcal{L}} Z(\theta^*, L_1, L_2), \quad \text{under } H_0$$

$$n d_n(\widehat{\theta}_n) = \mathsf{U}_n^1 + \mathsf{V}_n^1, \quad \text{with} \quad \mathsf{U}_n^1 \xrightarrow{\mathcal{L}} Z(\theta^c, L_1, L_2) \quad \text{and} \quad \mathsf{V}_n^1 \xrightarrow{a.s.} +\infty, \quad \text{under } H_1,$$

where the random variables $Z(\theta^*, L_1, L_2)$ and $Z(\theta^c, L_1, L_2)$, corresponding to a parametrized closed-form stochastic integral, could be consistently sampled (and thus tabulated) under both H_0 or H_1 by generating $Z(\widehat{\theta}_n, \widehat{L}_1, \widehat{L}_2)$-type random variables. It is important to mention that these last limiting random variables are strictly connected to the *inner convergence* phenomenon arising either under H_0 or H_1, as expressed in Theorem 2, see convergences (24-26), based on notation (9), along with (39) and (40).

Finally, by considering an empirical sample-based $(1-\alpha)$-quantile of the stochastic integral $Z(\widehat{\theta}_n, \widehat{L}_1, \widehat{L}_2)$, denoted $\widehat{q}_{1-\alpha}$, we propose to retain the following H_0-rejection rule:

$$\mathsf{T}_n \geq \widehat{q}_{1-\alpha} \quad \Rightarrow \quad H_0 \text{ is rejected.} \tag{13}$$

The above decision rule expresses the following principle: if the test statistic T_n is too remote from the *inner convergence* regime we could legitimately suspect a difference between F_1 and F_2, as illustrated in the right-hand side of Figure 3, and then reject H_0.

3. Identifiability and assumptions under $G_1 \neq G_2$

In this section, we propose to investigate under which type of conditions our discrepancy function $d(\cdot)$ satisfies the crucial setup (10). As it will appear in this section, the condition $G_1 \neq G_2$ plays a central role in the proportion parameters identification under H_0. Nevertheless, as it will be shown in Section D of the Supplement, our testing problem can still be addressed under $G_1 = G_2$ using a re-parametrisation approach.

Consider models (4) with generic proportions parameter $\theta = (p_1, p_2) \in \Theta$ and denote by $\theta^* = (p_1^*, p_2^*) \in]0,1[^2$ the true proportions parameter value. By isolating the expressions of F_1 and F_2 under θ we obtain for all $x \in \mathbb{R}$:

$$F_1(x, L_1, p_1) = \frac{L_1(x) - (1-p_1)G_1(x)}{p_1} \quad \text{and} \quad F_2(x, L_2, p_2) = \frac{L_2(x) - (1-p_2)G_2(x)}{p_2}. \tag{14}$$

Let us investigate now the situations where possibly $F_1(x, L_1, p_1) = F_2(x, L_2, p_2)$. Since under the true parameter p_i^*: $L_i(x) = (1-p_i^*)G_i(x) + p_i^* F_i(x)$, $i = 1, 2$, we easily obtain

$$F_1(x, L_1, p_1) = F_2(x, L_2, p_2) \Leftrightarrow \frac{p_1 - p_1^*}{p_1} G_1(x) = \frac{p_2 - p_2^*}{p_2} G_2(x) + \frac{p_2^*}{p_2} F_2(x) - \frac{p_1^*}{p_1} F_1(x). \tag{15}$$

Under H_0, $F_1 = F_2 = F$, we simply obtain

$$\frac{p_1 - p_1^*}{p_1} G_1(x) = \frac{p_2 - p_2^*}{p_2} G_2(x) + \left(\frac{p_2^*}{p_2} - \frac{p_1^*}{p_1} \right) F(x).$$

Hence, if $G_1 \notin \text{span}(G_2, F)$, which at least requires $G_1 \neq G_2$ and frames our present study, we necessarily have $p_1 = p_1^*$ and $p_2 = p_2^*$. On the other hand, under H_1 (or $F_1 \neq F_2$), if the cdfs family $\{G_1, G_2, F_1, F_2\}$ is linearly independent, equation (15) is impossible since it would imply $p_1^* = 0$ and $p_2^* = 0$ which is in contradiction with $\theta^* \in]0,1[^2$, and therefore $F_1(x, p_1) \neq F_2(x, p_2)$ for all $\theta \in \Theta$. Given the above discussion, in order to consistently pick the right θ^* under H_0 and select under H_1 a θ such that $F_1(x, p_1) \neq F_2(x, p_2)$ (the property being actually true for all $\theta \in \Theta$), it is natural to investigate the location of the minimum contrast parameter θ^c defined in (10). In order to reflect the linear independence of the cdfs needed to solve our testing problem, we propose two simple mutual identifiability conditions inspired from the identifiability theorem of Teicher (1963, Theorem 1):

(I) Under H_0 ($F_1 = F_2 = F$), there exists $(x_1, x_2, x_3) \in \mathbb{R}^3$ such that

$$\det \begin{pmatrix} G_1(x_1) & G_2(x_1) & F(x_1) \\ G_1(x_2) & G_2(x_2) & F(x_2) \\ G_1(x_3) & G_2(x_3) & F(x_3) \end{pmatrix} \neq 0. \tag{16}$$

(II) Under H_1, there exists $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ such that

$$\det \begin{pmatrix} G_1(x_1) & G_2(x_1) & F_1(x_1) & F_2(x_1) \\ G_1(x_2) & G_2(x_2) & F_1(x_2) & F_2(x_2) \\ G_1(x_3) & G_2(x_3) & F_1(x_3) & F_2(x_3) \\ G_1(x_4) & G_2(x_4) & F_1(x_4) & F_2(x_4) \end{pmatrix} \neq 0. \tag{17}$$

Remark 1. The above conditions are sufficient to characterize that, under the continuous, discrete or discrete/continuous distributions, the cdfs family $\mathcal{E}_1 = \{G_1, G_2, F\}$, respectively $\mathcal{E}_2 = \{G_1, F_1, G_2, F_2\}$, is linearly independent. Since G_1 and G_2 are known and $G_1 \neq G_2$, possible difficulties could happen under H_0 if for example $F_1 = F_2 = \alpha G_1 + (1-\alpha)G_2$, for $\alpha \in]0,1[$, meaning that L_1 and L_2 are (G_1, G_2)-mixtures. Nevertheless the testing of such hypothesis is not really a concern since the pdfs G_1 and G_2 are known and that an adapted simple version of the IBM-test can be proposed. This identifiability pre-checking method will be closely addressed in an upcoming work. Similar difficulties could happen under H_1 if for example $F_1 = \beta G_2(x) + (1-\beta)F_2(x)$, for $\beta \in]0,1[$, which Patra and Sen (2016) qualifies for F_1 to have G_2 and F_2 in its background. To avoid these type of confusing situations, practitioners have to assess, given their own knowledge about the datasets collection, that the contamination phenomena they want to compare are endogenous and spontaneous relatively to each population (no "physical" porosity between the two-sample populations). For instance in our COVID-19 mortality excess comparison over a panel of European countries, the above delicate/odd situation would happen if a noticeable amount of people from a given country impacted by the pandemic migrated to another one (without knowing it) not impacted by the pandemic yet. These are typically the kind of situations we cannot address with our method. Thereby, our method does not rely on identifiability shape conditions for marginal model identification, which is the only case, to the best of our knowledge, where asymptotic normality results can be obtained in a test perspective, see Milhaud et al. (2022) for the symmetric case, but only on a clear separation of the sources composing the mixture models to be compared. Note also that condition **(II)** clearly states that at least 4 classes are required in the discrete cases to correctly address the testing problem (3), which condition just holds in the application Section 7.

In order to clarify why the Patra and Sen (2016) marginal model estimation is not suitable to solve the two-sample estimation problem (3), let us recall a few facts about the complexity of the semiparametric contamination model identifiability. For p fixed in $]0,1[$, consider $\lambda \in \mathbb{R}$ such that $p + \lambda \in]0,1[$, it then comes:

$$\begin{aligned} L(x) &= (1-p)G(x) + pF(x), \qquad (18) \\ &= (1-[p+\lambda])G(x) + [p+\lambda]\left(\frac{p}{p+\lambda}F(x) + \frac{\lambda}{p+\lambda}G(x)\right), \\ &= (1-\pi(\lambda))G(x) + \pi(\lambda)F_{\pi(\lambda)}(x), \qquad (19) \end{aligned}$$

which shows the non-uniqueness of the mixture model (2) representation. Now by "squizzing" the parameter π (by making λ decrease) in the expression, we are left with

$$F_{\pi(\lambda)}(x) = \frac{p}{p+\lambda}F(x) + \frac{\lambda}{p+\lambda}G(x), \quad \pi(\lambda) = p + \lambda\searrow. \tag{20}$$

Hence, we possibly face three type of situations:

- $\lambda > 0$: $F_{\pi(\lambda)}(x)$ is a valid cdf.
- $\lambda = 0$: $F_{\pi(\lambda)}(x) = F(x)$ is a valid cdf.

- $\lambda < 0$: $F_{\pi(\lambda)}(x) = F(x)$ is possibly not a valid cdf (difference of two increasing functions).

Based on this principle, Patra and Sen (2016) propose to well-pose the proportion parameter definition in the semiparametric model (2) as follows:

$$p_0 = \inf\left\{\pi \in]0,1[: \underbrace{\frac{L-(1-\pi)G}{\pi}}_{F(\cdot,\pi,L)} \quad \text{is a valid cdf}\right\},$$

and estimate this parameter by

$$\widehat{p}_{c_n} = \inf\left\{\pi \in]0,1[: \pi d_n(F(\cdot,\pi,\widehat{L}_n), \check{F}_n(\cdot,\pi)) \leq \frac{c_n}{\sqrt{n}}\right\},$$

where the tuning parameter $c_n \to +\infty$ not too fast, d_n is the empirical $L^2(\mathbb{L}_n)$ distance, $\check{F}_n(\cdot,\pi)$ is the closest cdf from $F(\cdot,\pi,\widehat{L}_n)$ in the $L^2(\mathbb{L}_n)$-sense obtained by isotonic regression and pool adjacent violators algorithm. Patra and Sen (2016, Theorem 3) show that their estimation method is consistent in probability with a certain rate but cannot achieve the $\sqrt{n}$-consistency, see Patra and Sen (2016, Theorem 4), which seriously compromises the chances to build a statistical test based on their estimators. According to us, the IBM approach provided with the cross-model identifiability conditions **(I)** and **(II)**, requiring no proper shape conditions, allows under H_0 to drastically simplify the parametric estimation, truly complex due to the infinite mixture representation (18), thanks to the targeted cross-model condition **(I)** and takes benefit of the inability to find any matching under H_1, because of condition **(II)**.

We finally assume an additional technical condition:

(A) The $d(\cdot)$ minimiser θ^c ($\theta^c = \theta^* \in]0,1[^2$ under H_0, or $\theta^c \neq \theta^*$ under H_1), belongs to $\overset{o}{\Theta}$ the interior of the compact parametric space Θ.

This condition is crucial to guarantee that a Taylor expansion of the empirical gradient $\dot{d}_n(\cdot)$ about point θ^c can be made, see expression (27) and its further convergence analysis. This precaution is connected to the artefact described in Section 2, see points i) and ii) along with Figure 2.

4. Asymptotic results

In the sequel, we denote by $\dot{\ell}(\vartheta)$ and $\ddot{\ell}(\vartheta)$ the gradient vector and Hessian matrix of any real function ℓ (when it makes sense) with respect to argument $\vartheta \in \mathbb{R}^2$. The notation A^T refers to the transpose matrix of A. To look at the proofs and technical results related to Lemma 1, the reader is referred to Section A in the Supplement.

Lemma 1. (*i*) *The mapping $\theta \mapsto d(\theta)$ is C^2 over Θ both under H_0 or H_1.*

(*ii*) *Assume that conditions* **(I)** *and* **(A)** *hold. If U is strictly increasing on an interval I_U that encompasses the support of the L_i's and G_i's, $i = 1, 2$, then under H_0, d is a contrast function, i.e. for all $\theta \in \Theta$, $d(\theta) \geq 0$ and $d(\theta) = 0$ if and only if $\theta = \theta^* \in \overset{o}{\Theta}$.*

(*iii*) *Assume that conditions* **(II)** *and* **(A)** *hold. If U is strictly increasing on an interval I_U that encompasses the support of the L_i's and G_i's, $i = 1, 2$, and if for any given case under H_1 there exists one single point $\theta^c \in \overset{o}{\Theta}$ such that $\theta^c = \arg\min_{\theta\in\Theta} d(\theta)$, then $d(\theta^c) > 0$.*

(*iv*) *The regular and empirical contrasts* $d(\cdot)$ *and* $d_n(\cdot)$ *are Lipschitz over* Θ *under* H_0 *or* H_1.
(*v*) *We have under* H_0 *or* H_1 *that*

$$\sup_{\theta\in\Theta} |d_n(\theta) - d(\theta)| = o_{a.s.}(n^{-1/2+\alpha}), \quad \text{for all } \alpha > 0. \tag{21}$$

(*vi*) *Assume that conditions* **(I)** *and* **(A)** *hold. Then under* H_0 *the Hessian matrix*

$$\ddot{d}(\theta^*) = 2\int_{\mathbb{R}} \dot{D}(x,\theta^*)\dot{D}^T(x,\theta^*)dU(x) \tag{22}$$

is symmetric and positive definite.

Since we could not find simple tractable conditions to prove the counter part of vi) under H_1 in the above lemma, we propose to consider the following additional condition.

(DP) Under H_1 the Hessian matrix $\ddot{d}(\theta^c) = 2\int_{\mathbb{R}} \ddot{D}(x,\theta^c)D(x,\theta^c) + \dot{D}(x,\theta^c)\dot{D}^T(x,\theta^c)dU(x)$ is symmetric and positive definite.

Note that the above condition can be checked numerically through the consistent estimate $\ddot{d}_n(\widehat{\theta}_n)$ of $\ddot{d}(\theta^c)$, see Theorem 1 i) and Section C.2 for closed form expression. It is indeed enough to check that the eigenvalues of $\ddot{d}_n(\widehat{\theta}_n)$ are real and positive. Also if $\ddot{d}_n(\widehat{\theta}_n)$ proves to be non symmetric positive definite, this must be interpreted as a serious warning about H_0 simply because if H_0 and **(I)** are true, then the Hessian matrix $\ddot{d}(\theta^*)$ should be precisely symmetric positive definite according to Lemma 1 vi).

Let us denote as $\|\cdot\|_2$ the Euclidean distance in $\mathbb{R}^2$, and $\theta^c = \theta^*$ if the assumption H_0 is specified.

Theorem 1. 1. *If conditions* **(I)**, *respectively* **(II)**, *and* **(A)** *hold, we have under* H_0, *resp.* H_1, $\widehat{\theta}_n \xrightarrow{P} \theta^c$ *as* $n \to +\infty$.
2. *If conditions* **(I)**, *respectively* **(II)** *and* **(DP)** , *and* **(A)** *hold, we have under* H_0, *resp.* H_1, *that* $\|\widehat{\theta}_n - \theta^c\|_2 = o_{a.s.}(n^{-1/4+\alpha})$ *for all* $\alpha > 0$.

Proof. i) Using the Lipschitz property on both $d_n(\cdot)$ and $d(\cdot)$ and uniform convergence of $d_n(\cdot)$ towards $d(\cdot)$ stated respectively in Lemma 1 iv) and v), the classical proof of consistency detailed in Butucea and Vandekerkhove (2014) applies and provides the result.

ii) By Lemma 1 v) or vi) there exists $\gamma > 0$ such that for all $v \in \mathbb{R}^2$, $v^T\ddot{d}(\theta^c)v > \gamma||v||_2^2$. By a second-order Taylor expansion of d at $\theta^c \in \overset{o}{\Theta}$ we can find $\eta > 0$ such that for all v satisfying $\|v\| < \eta$ and $\theta^c + v \in \overset{\circ}{\Theta}$, we have

$$d(\theta^c + v) \geq \frac{\gamma}{4}\|v\|_2^2. \tag{23}$$

Let us consider now $B(\theta^c, \eta_n)$ the Euclidean ball centred at θ^c with radius $\eta_n > 0$. Following the proof of Theorem 3.3 in Bordes, Mottelet and Vandekerkhove (2006), we show the following events inclusion:

$$\limsup_n \left\{\widehat{\theta}_n \notin B(\theta^c,\eta_n)\right\} \subseteq \limsup_n \left\{\inf_{\theta\in\Theta\setminus B(\theta^c,\eta_n)} d(\theta) < \xi_n\right\} \cup \limsup_n \left\{\xi_n \leq 2\sup_{\theta\in\Theta}|d_n(\theta) - d(\theta)|\right\},$$

for any arbitrary sequence ξ_n. Choosing now $\xi_n = n^{-1/2+\alpha}$ and $\eta_n = n^{-1/4+\beta/2}$, with $0 < \alpha < \beta$ taken arbitrarily small, it follows from (23) and the uniform almost sure rate of d_n given in Lemma 1 (iv), that

$$P\left(\limsup_n \left\{ \inf_{\theta \in \Theta \setminus B(\theta^c, \eta_n)} d(\theta) < \xi_n \right\}\right) = 0$$

and

$$P\left(\limsup_n \left\{ \xi_n \leq 2 \sup_{\theta \in \Theta} |d_n(\theta) - d(\theta)| \right\}\right) = 0.$$

In conclusion, $\widehat{\theta}_n$ converges almost surely towards θ^c at rate $n^{-1/4+\alpha}$, for $\alpha > 0$ chosen arbitrarily small. □

Let us denote as $D(\mathbb{R})$ the space of càd-làg functions on $\mathbb{R}$.

Theorem 2. 1. *If conditions* **(I)**, *respectively* **(II)** *and* **(DP)**, *and* **(A)** *hold, we have under* H_0, *resp.* H_1, *that*

$$\sqrt{n}\begin{bmatrix} \widehat{p}_1 - p_1^c \\ \widehat{p}_2 - p_2^c \\ D_n(\cdot) - D(\cdot) \end{bmatrix} \rightsquigarrow W(\theta^c, \cdot) \text{ in } \mathbb{R}^2 \times D(\mathbb{R}), \tag{24}$$

where $D_n(\cdot) = D(\cdot, \widehat{L}_1, \widehat{L}_2, \widehat{\theta}_n)$, *and* $W(\theta^c, \cdot) = (W_1(\theta^c), W_2(\theta^c), W_3(\theta^c, \cdot))^T$ *is a centered 3-dimensional Gaussian process with covariance matrix* $\Sigma_W = M(\theta^c, \cdot)\Sigma_L(\cdot, \cdot)M(\theta^c, \cdot)^T$ *where* $M(\theta^c, \cdot)$ *is defined in* (34) *and* $\Sigma_L(\cdot, \cdot)$ *in* (37).

2. *If conditions* **(I)**, *respectively* **(II)** *and* **(DP)**, *and* **(A)** *hold, we have that*

$$\mathsf{T}_n = \mathsf{U}_n^0 \xrightarrow{\mathcal{L}} Z(\theta^*) = \int_{\mathbb{R}} (W_3(\theta^*, x))^2 dU(x), \quad \text{under } H_0, \tag{25}$$

$$\mathsf{T}_n = \mathsf{U}_n^1 + \mathsf{V}_n^1, \quad \text{with} \quad \mathsf{V}_n^1 = n \int_{\mathbb{R}} D^2(x, L_1, L_2, \theta^c) dU(x) + o_{a.s.}(n)$$

$$\text{and } \mathsf{U}_n^1 \xrightarrow{\mathcal{L}} Z(\theta^c) = \int_{\mathbb{R}} (W_3(\theta^c, x))^2 dU(x), \quad \text{under } H_1. \tag{26}$$

Proof. i) By a Taylor expansion of $\dot{d}_n$ about $\theta^c \in \overset{o}{\Theta}$ we have

$$\ddot{d}_n(\widetilde{\theta}_n)\sqrt{n}(\widehat{\theta}_n - \theta^c) = -\sqrt{n}\dot{d}_n(\theta^c), \tag{27}$$

where $\widetilde{\theta}_n$ lies in the line segment with extremities $\widehat{\theta}_n$ and θ^c. Now writing that $\dot{d}(\theta) = (\dot{d}_1(\theta), \dot{d}_2(\theta))^T$,

$$\dot{d}_1(\theta) = 2E_U\left(\frac{(2-p_1)G_1L_1}{p_1^3} - \frac{(1-p_1)G_1^2}{p_1^3} - \frac{L_1^2}{p_1^3} - \frac{G_1L_2}{p_1^2p_2} + \frac{(1-p_2)G_1G_2}{p_1^2p_2} + \frac{L_1L_2}{p_1^2p_2} - \frac{(1-p_2)L_1G_2}{p_1^2p_2}\right)$$

$$\dot{d}_2(\theta) = 2E_U\left(\frac{(2-p_2)G_2L_2}{p_2^3} - \frac{(1-p_2)G_2^2}{p_2^3} - \frac{L_2^2}{p_2^3} - \frac{G_2L_1}{p_2^2p_1} + \frac{(1-p_1)G_2G_1}{p_2^2p_1} + \frac{L_2L_1}{p_2^2p_1} - \frac{(1-p_1)L_2G_1}{p_2^2p_1}\right),$$

we look at

$$\dot{d}_{1,n}(\theta) - \dot{d}_1(\theta) = 2\left(\frac{2-p_1}{p_1^3}T_{1,1} - \frac{1-p_1}{p_1^3}T_{1,2} - \frac{1}{p_1^3}T_{1,3} - \frac{1}{p_1^2 p_2}T_{1,4} + \frac{1-p_2}{p_1^2 p_2}T_{1,5} + \frac{1}{p_1^2 p_2}T_{1,6} - \frac{1-p_2}{p_1^2 p_2}T_{1,7}\right)$$

$$\dot{d}_{2,n}(\theta) - \dot{d}_2(\theta) = 2\left(\frac{2-p_2}{p_2^3}T_{2,1} - \frac{1-p_2}{p_2^3}T_{2,2} - \frac{1}{p_2^3}T_{2,3} - \frac{1}{p_2^2 p_1}T_{2,4} + \frac{1-p_1}{p_2^2 p_1}T_{2,5} + \frac{1}{p_2^2 p_1}T_{2,6} - \frac{1-p_1}{p_2^2 p_1}T_{2,7}\right),$$

where

$$\begin{aligned}
T_{1,1}(G_1, L_1) &= E_U\left(G_1\left(\widehat{L}_1 - L_1\right)\right)\\
T_{1,2}(G_1) &= E_U\left(G_1^2 - G_1^2\right) = 0\\
T_{1,3}(L_1) &= E_U\left(\widehat{L}_1^2 - L_1^2\right) = E_U\left((\widehat{L}_1 - L_1)(\widehat{L}_1 + L_1)\right) = E_U\left((\widehat{L}_1 - L_1)(2L_1 + o_{a.s.}(1))\right)\\
T_{1,4}(G_1, L_2) &= E_U(G_1(\widehat{L}_2 - L_2))\\
T_{1,5}(G_1, G_2) &= E_U(G_1 G_2) - E_U(G_1 G_2) = 0\\
T_{1,6}(L_1, L_2) &= E_U\left(\widehat{L}_1\widehat{L}_2 - L_1 L_2\right) = E_U\left(\widehat{L}_1(\widehat{L}_2 - L_2) + L_2(\widehat{L}_1 - L_1)\right)\\
&= E_U\left((L_1 + o_{a.s.}(1)(\widehat{L}_2 - L_2)\right) + E_U\left(L_2(\widehat{L}_1 - L_1)\right)\\
T_{1,7}(G_2, L_1) &= E_U\left(G_2(\widehat{L}_1 - L_1)\right) = T_{1,4}(G_2, L_1)\\
T_{2,1}(G_2, L_2) &= E_U\left(G_2(\widehat{L}_2 - L_2)\right) = T_{1,1}(G_2, L_2)\\
T_{2,2}(G_2) &= E_U\left(G_2^2\right) - E_U\left(G_2^2\right) = 0\\
T_{2,3}(L_2) &= E_U\left((\widehat{L}_2 - L_2)(\widehat{L}_2 + L_2)\right) = E_U\left((\widehat{L}_2 - L_2)(2L_2 + o_{a.s.}(1))\right) = T_{1,3}(L_2)\\
T_{2,4}(L_1, G_2) &= E_U\left(G_2(\widehat{L}_1 - L_1)\right) = T_{1,4}(L_1, G_2)\\
T_{2,5}(G_1, G_2) &= E_U(G_2 G_1) - E_U(G_2 G_1) = 0\\
T_{2,6}(L_1, L_2) &= E_U\left(\widehat{L}_2\widehat{L}_1 - L_2 L_1\right) = T_{1,6}(L_1, L_2)\\
T_{2,7}(G_1, L_2) &= E_U\left(G_1(\widehat{L}_2 - L_2)\right).
\end{aligned}$$

For a generic cdf Y and a generic N-sample based empirical process $\mathbb{V} = \sqrt{N}(\widehat{V} - V)$, define $\varphi(Y, \mathbb{V}) = \int_{\mathbb{R}} Y(x)\mathbb{V}(x)dU(x)$. Introducing $\mathcal{S} = (G_1, G_2, L_1, L_2)$, let us consider

$$\Psi_{1,1}(\mathcal{S}, \theta) = 2\left(\frac{2-p_1}{p_1^3}G_1 - \frac{2}{p_1^3}L_1 + \frac{1}{p_1^2 p_2}L_2 - \frac{1-p_2}{p_1^2 p_2}G_2\right)$$

$$\Psi_{1,2}(\mathcal{S}, \theta) = 2\left(\frac{1}{p_1^2 p_2}(L_1 - G_1)\right)$$

$$\Psi_{2,1}(\mathcal{S}, \theta) = 2\left(\frac{2-p_2}{p_2^3}G_2 - \frac{2}{p_2^3}L_2 + \frac{1}{p_2^2 p_1}L_1 - \frac{1-p_1}{p_2^2 p_1}G_1\right)$$

$$\Psi_{2,2}(\mathcal{S}, \theta) = 2\left(\frac{1}{p_2^2 p_1}(L_2 - G_2)\right)$$

and

$$\Phi_{1,1}(\mathbb{L}_1,\theta)=\varphi(\Psi_{1,1}(S,\theta),\mathbb{L}_1),\quad \Phi_{1,2}(\mathbb{L}_2,\theta)=\varphi(\Psi_{1,2}(S,\theta),\mathbb{L}_2),$$
$$\Phi_{2,1}(\mathbb{L}_2,\theta)=\varphi(\Psi_{2,1}(S,\theta),\mathbb{L}_2),\quad \Phi_{2,2}(\mathbb{L}_1,\theta)=\varphi(\Psi_{2,2}(S,\theta),\mathbb{L}_1).$$

Note that the first and fourth, respectively the second and third, expression depends only on the randomness of $\mathbb{L}_1$, resp. $\mathbb{L}_2$. We summarize the above remarks into the following basic expression:

$$\sqrt{n}(\dot{d}_n(\theta)-\dot{d}(\theta))=\Phi(\mathbb{L}_1,\mathbb{L}_2,\theta)+o_{a.s}(1), \tag{28}$$

where, according to $\sqrt{n}=\sqrt{\kappa n}/\sqrt{\kappa}=\zeta\sqrt{n_2}$ with $\zeta=1/\sqrt{\kappa}$:

$$\Phi(\mathbb{L}_1,\mathbb{L}_2)=\begin{bmatrix}\Phi_{1,1}(\mathbb{L}_1,\theta)+\zeta\Phi_{1,2}(\mathbb{L}_2,\theta)\\ \zeta\Phi_{2,1}(\mathbb{L}_2,\theta)+\Phi_{2,2}(\mathbb{L}_1,\theta)\end{bmatrix}. \tag{29}$$

Since the empirical processes $\mathbb{L}_1$ and $\mathbb{L}_2$ are independent, by the Donsker Theorem, see Van der Vaart (2000, Theorem 19.3, p. 266) the vector $[\mathbb{L}_1,\mathbb{L}_2]$ converges in distribution to a bi-dimensional zero-mean Gaussian process $\mathcal{B}$, i.e.

$$\begin{bmatrix}\mathbb{L}_1\\ \mathbb{L}_2\end{bmatrix}\rightsquigarrow\mathcal{B}=\begin{bmatrix}\mathcal{B}_1\\ \mathcal{B}_2\end{bmatrix}\quad\text{in}\quad D(\mathbb{R})\times D(\mathbb{R}), \tag{30}$$

where $\mathcal{B}$ is a bi-dimensional Gaussian process with diagonal correlation matrix $\rho=\mathrm{diag}(\rho_1,\rho_2)$, where $\rho_1(x,y)=L_1(x\wedge y)(1-L_1(x\vee y))$ and $\rho_2(x,y)=L_2(x\wedge y)(1-L_2(x\vee y))$.

Moreover,

$$\begin{aligned}\sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n)-D(x,L_1,L_2,\theta^c)]&=\sqrt{n}[F_1(x,\widehat{L}_1,\widehat{p}_1)-F_1(x,L_1,p_1^c)]\\ &\quad-\sqrt{n}[(F_2(x,\widehat{L}_2,\widehat{p}_2)-F_2(x,L_2,p_2^c))].\end{aligned} \tag{31}$$

Let us decompose, for $i=1,2$, the terms $F_i(x,\widehat{L}_i,\widehat{p}_i)-F_i(x,L_i,p_i^c)$:

$$\begin{aligned}\sqrt{n}[F_i(\cdot,\widehat{L}_i,\widehat{p}_i)-F_i(\cdot,L_i,p_i^c)]&=\sqrt{n}\left[\left(\frac{\widehat{L}_i}{\widehat{p}_i}-\frac{L_i}{p_i^c}\right)-\left(\frac{1-\widehat{p}_i}{\widehat{p}_i}-\frac{1-p_i^c}{p_i^c}\right)G_i\right]\\ &=\sqrt{n}\left[\frac{\widehat{L}_i-L_i}{\widehat{p}_i}\right]+\sqrt{n}\left(\frac{p_i^c-\widehat{p}_i}{p_i^c\widehat{p}_i}\right)(L_i-G_i)\\ &=\zeta_i\frac{1}{p_i^c}\mathbb{L}_i-\left(\frac{L_i-G_i}{(p_i^c)^2}\sqrt{n}[\widehat{p}_i-p_i^c]\right)+o_P(1),\end{aligned} \tag{32}$$

where by convention $\zeta_1=1$ and $\zeta_2=\zeta=\frac{1}{\sqrt{\kappa}}$. It is also easy to prove that $\ddot{d}_n(\widetilde{\theta}_n)\overset{a.s.}{\to}\ddot{d}(\theta^c)>0$, as $n\to+\infty$. Indeed, considering the decompositions (49), (50) and (51) in Appendix C, we have

$$\left|[\ddot{d}_n(\widetilde{\theta}_n)-\ddot{d}(\theta^c)]_{i,j}\right|\le\int_{\mathbb{R}}\left|M_{i,j}(x,\widehat{L}_1,\widehat{L}_2,G_1,G_2,\widetilde{\theta}_n)-M_{i,j}(x,L_1,L_2,G_1,G_2,\theta)\right|dU(x)$$

$$\leq \int_{\mathbb{R}} \left| M_{i,j}(x, \widehat{L}_1, \widehat{L}_2, G_1, G_2, \widetilde{\theta}_n) - M_{i,j}(x, L_1, L_2, G_1, G_2, \widetilde{\theta}_n) \right| dU(x)$$

$$+ \int_{\mathbb{R}} \left| M_{i,j}(x, L_1, L_2, G_1, G_2, \widetilde{\theta}_n) - M_{i,j}(x, L_1, L_2, G_1, G_2, \theta) \right| dU(x)$$

$$\leq C \left(\mathcal{P}(\widetilde{\theta}_n) \left[\sum_{i=1}^{2} \|\widehat{L}_i - L_i\|_\infty \right] + |\mathcal{P}(\widetilde{\theta}_n) - \mathcal{P}(\theta^c)| \right), \tag{33}$$

where $\mathcal{P}(\theta) = \sum_{k=0}^{4} p_1^{-k} p_2^{-4+k}$ is a $\mathbb{R}^2 \to \mathbb{R}$ continuous mapping. Now by using on (33) the Glivenko-Cantelli theorem and the $a.s.$ convergence of $\widehat{\theta}_n$ towards θ^c stated in Theorem 1 ii), we obtain the wanted result.

In order to synthetically summarize results (27), (28), (29) and (31–32) for the Central Limit Theorem relative to our quantities of interest, we define the following matrix-type relation:

$$\sqrt{n} \begin{bmatrix} \widehat{p}_1 - p_1^c \\ \widehat{p}_2 - p_2^c \\ D_n(\cdot) - D(\cdot) \end{bmatrix} = M(\theta^c, \cdot) \begin{bmatrix} \Phi_{1,1}(\mathbb{L}_1, \theta^c) \\ \Phi_{2,2}(\mathbb{L}_1, \theta^c) \\ \mathbb{L}_1 \\ \Phi_{2,1}(\mathbb{L}_2, \theta^c) \\ \Phi_{1,2}(\mathbb{L}_2, \theta^c) \\ \mathbb{L}_2 \end{bmatrix} + o_P(1), \tag{34}$$

with $M(\theta^c, \cdot) = L(\cdot, \theta^c)\, J^{-1}(\theta^c)\, C$ and

$$C = \begin{bmatrix} -1 & 0 & 0 & 0 & -\zeta & 0 \\ 0 & -1 & 0 & -\zeta & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad J(\theta) = \begin{bmatrix} \ddot{d}(\theta) & 0_{2\times 2} \\ 0_{2\times 2} & \mathrm{Id}_{2\times 2} \end{bmatrix},$$

$$L(\cdot, \theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\dfrac{L_1(\cdot) - G_1(\cdot)}{p_1^2} & \dfrac{L_2(\cdot) - G_2(\cdot)}{p_2^2} & \dfrac{1}{p_1} & -\dfrac{\zeta}{p_2} \end{bmatrix}. \tag{35}$$

We finally have

$$\begin{bmatrix} \Phi_{1,1}(\mathbb{L}_1, \theta^c) \\ \Phi_{2,2}(\mathbb{L}_1, \theta^c) \\ \mathbb{L}_1 \\ \Phi_{2,1}(\mathbb{L}_2, \theta^c) \\ \Phi_{1,2}(\mathbb{L}_2, \theta^c) \\ \mathbb{L}_2 \end{bmatrix} \rightsquigarrow Z = \begin{bmatrix} \Phi_{1,1}(\mathcal{B}_1, \theta^c) \\ \Phi_{2,2}(\mathcal{B}_1, \theta^c) \\ \mathcal{B}_1 \\ \Phi_{2,1}(\mathcal{B}_2, \theta^c) \\ \Phi_{1,2}(\mathcal{B}_2, \theta^c) \\ \mathcal{B}_2 \end{bmatrix} \quad \text{in} \quad (\mathbb{R}^2 \times D(\mathbb{R}))^2, \tag{36}$$

where Z is a Gaussian random vector of $\mathbb{R}^6$ with covariance matrix $\Sigma_L = E(ZZ^T)$. Since $\mathcal{B}_1$ and $\mathcal{B}_2$ are two independent (limiting) Gaussian processes, we have

$$\Sigma_L(x, y) = \begin{bmatrix} \Sigma_1(x, y) & 0_{3\times 3} \\ 0_{3\times 3} & \Sigma_2(x, y) \end{bmatrix}. \tag{37}$$

Because

$$E\left(\int_{\mathbb{R}} f_1(x)\mathcal{B}(x)dU(x)\times\int_{\mathbb{R}} f_2(y)\mathcal{B}(y)dU(y)\right)=E\left(\int_{\mathbb{R}^2} f_1(x)f_2(y)\mathcal{B}(x)\mathcal{B}(y)dU(x)dU(y)\right)$$
$$=\int_{\mathbb{R}^2} f_1(x)f_2(y)E(\mathcal{B}(x)\mathcal{B}(y))dU(x)dU(y)$$
$$=\int_{\mathbb{R}^2} f_1(x)f_2(y)\mathrm{cov}(\mathcal{B}(x),\mathcal{B}(y))dU(x)dU(y)$$
$$=\int_{\mathbb{R}^2} f_1(x)f_2(y)\rho(x,y)dU(x)dU(y),$$

it comes that $\Sigma_1(x,y)=(\sigma_1(i,j;x,y))_{1\le i,j\le 3}$ where

$$\sigma_1(1,1;x,y)=\sigma_1(1,1)=\int_{\mathbb{R}^2}\Psi_{1,1}(S,\theta,u)\Psi_{1,1}(S,\theta,v)\rho_1(u,v)dU(u)dU(v)$$
$$\sigma_1(1,2;x,y)=\sigma_1(1,2)=\sigma_1(2,1)=\int_{\mathbb{R}^2}\Psi_{1,1}(S,\theta,u)\Psi_{2,2}(S,\theta,v)\rho_1(u,v)dU(u)dU(v)$$
$$\sigma_1(1,3;x,y)=\sigma_1(1,3;y)=\int_{\mathbb{R}}\Psi_{1,1}(S,\theta,u)\rho_1(u,y)dU(u)$$
$$\sigma_1(2,2;x,y)=\sigma_1(2,2)=\int_{\mathbb{R}^2}\Psi_{2,2}(S,\theta,u)\Psi_{2,2}(S,\theta,v)\rho_1(u,v)dU(u)dU(v)$$
$$\sigma_1(2,3;x,y)=\sigma_1(2,3;y)=\int_{\mathbb{R}}\Psi_{2,2}(S,\theta,u)\rho_1(u,y)dU(u)$$
$$\sigma_1(3,3;x,y)=\rho_1(x,y)$$
$$\sigma_1(3,1;x,y)=\sigma_1(3,1;x)=\sigma_1(1,3;x)$$
$$\sigma_1(3,2;x,y)=\sigma_1(3,2;x)=\sigma_1(2,3;x),$$

and $\Sigma_2=(\sigma_2(i,j;x,y))_{1\le i,j\le 3}$ where

$$\sigma_2(1,1;x,y)=\sigma_2(1,1)=\int_{\mathbb{R}^2}\Psi_{2,1}(S,\theta,u)\Psi_{2,1}(S,\theta,v)\rho_2(u,v)dU(u)dU(v)$$
$$\sigma_2(1,2;x,y)=\sigma_2(1,2)=\sigma_2(2,1)=\int_{\mathbb{R}^2}\Psi_{2,1}(S,\theta,u)\Psi_{1,2}(S,\theta,v)\rho_2(u,v)dU(u)dU(v)$$
$$\sigma_2(1,3;x,y)=\sigma_2(1,3;y)=\int_{\mathbb{R}}\Psi_{2,1}(S,\theta,u)\rho_2(u,y)dU(u)$$
$$\sigma_2(2,2;x,y)=\sigma_2(2,2)=\int_{\mathbb{R}^2}\Psi_{1,2}(S,\theta,u)\Psi_{1,2}(S,\theta,v)\rho_2(u,v)dU(u)dU(v)$$
$$\sigma_2(2,3;x,y)=\sigma_2(2,3;y)=\int_{\mathbb{R}}\Psi_{1,2}(S,\theta,u)\rho_2(u,y)dU(u)$$
$$\sigma_2(3,3;x,y)=\rho_2(x,y)$$
$$\sigma_2(3,1;x,y)=\sigma_2(3,1;x)=\sigma_2(1,3;x)$$

$$\sigma_2(3,2;x,y) = \sigma_2(3,2;x) = \sigma_2(2,3;x).$$

Note that all the above matrices can be estimated consistently, see Appendix C.2.

ii) Let us now decompose $\mathsf{T}_n = nd_n(\widehat{\theta}_n)$:

$$\begin{aligned} nd_n(\widehat{\theta}_n) &= \int_{\mathbb{R}} n(D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^c) + D(x,L_1,L_2,\theta^c))^2 dU(x) \\ &= \int_{\mathbb{R}} (\sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^c)])^2 dU(x) \\ &\quad +2\sqrt{n}\int_{\mathbb{R}} \sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^c)]D(x,L_1,L_2,\theta^c)dU(x) \\ &\quad +n\int_{\mathbb{R}} D^2(x,L_1,L_2,\theta^c)dU(x). \end{aligned} \tag{38}$$

Note that under H_0, $\theta^c = \theta^*$ and we simply obtain

$$\begin{aligned} nd_n(\widehat{\theta}_n) &= \int_{\mathbb{R}} (\sqrt{n}D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n))^2 dU(x) \\ &= \int_{\mathbb{R}} (\sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^*)]^2 dU(x) \\ &= \mathsf{U}_n^0, \end{aligned} \tag{39}$$

since $D(\cdot,L_1,L_2,\theta^*) = 0$ almost everywhere. Next, denoting $\mathbb{D}_n(\cdot) = \sqrt{n}(D_n(\cdot) - D(\cdot))$ it is easy to show that the mapping $\mathbb{D} \mapsto \int(\mathbb{D}(x))^2 dU(x)$ is Hadamard differentiable from the domain of càd-làg functions of bounded variation into $\mathbb{R}$, see Van der Vaart (2000, Theorem 20.10). This combined with the weak convergence result about the process $\mathbb{D}_n$, see the third row of (24), yields the desired result.

Note now that under H_1, we have $\int_{\mathbb{R}} D^2(x,L_1,L_2,\theta^c)dU(x) > 0$ and that the cross-term in (38) can be controlled, according to decompositions (31) and (32), by the law of the iterated Logarithm for empirical processes and the almost sure rate of convergence established in Theorem 1 ii), by:

$$\begin{aligned} &2\sqrt{n}\int_{\mathbb{R}} \sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^c)]D(x,L_1,L_2,\theta^c)dU(x) \\ &\leq 2\sqrt{n}\frac{1+\widetilde{\delta}}{\delta_1}\|\sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^c)]\|_\infty \\ &\leq 2\sqrt{n}\frac{1+\widetilde{\delta}}{\delta_1}\left(\sum_{i=1}^{2}\frac{\xi_i}{p_i^c}\|\mathbb{L}_i\|_\infty + \frac{2}{p_i^c}\sqrt{n}|\widehat{p}_i - p_i^c|\right) \\ &= \sqrt{n}\left(O_{a.s.}\left(\sqrt{\log\log(n)}\right) + o_{a.s}(n^{1/4+\alpha})\right) = o_{a.s.}(n). \end{aligned}$$

Using the above remark we obtain

$$nd_n(\widehat{\theta}_n) = \int_{\mathbb{R}} (\sqrt{n}[D(x,\widehat{L}_1,\widehat{L}_2,\widehat{\theta}_n) - D(x,L_1,L_2,\theta^c)])^2 dU(x)$$

$$+n\int_{\mathbb{R}} D^2(x, L_1, L_2, \theta^c)dU(x) + o_{a.s.}(n)$$
$$= \mathsf{U}_n^1 + \mathsf{V}_n^1. \tag{40}$$

Given the asymptotic convergence analysis under both H_0 or H_1, the random variable within brackets involved in the first term of (38) and (40) can be analysed closely. □

Let us remind that the stochastic integrals distribution in Theorem 2 can be simulated by standard Monte Carlo methods, see for instance Higham (2001), and thus be fully tabulated. As detailed in Section 1, the use of the above theorem in our testing perspective consists in rejecting H_0 if the statistic $\mathsf{T}_n = nd_n(\widehat{\theta}_n)$ exceeds $\widehat{q}_{1-\alpha}$, where $\widehat{q}_{1-\alpha}$ is the approximated $(1-\alpha)$-quantile of the limiting random variable $\int_{\mathbb{R}}(W_3(\theta^c, x))^2 dU(x)$ (given that by convention $\theta^c = \theta^*$ under H_0).

We propose, in order to identify if the testing problem is relevant or not, see third situation in Figure 3, to check if a $(1-\alpha)$-region of confidence of θ^c, denoted $\mathcal{R}_{1-\alpha}$, intersects somehow the $]0,1[^2$ proportions domain. Such a case could possibly mean that, due to estimation uncertainty, the parameter θ^c could be located close to the $]0,1[^2$ border. To build the desired $(1-\alpha)$-region of confidence we notice that we have:

$$(\widehat{\theta}_n - \theta^c)^T [\ddot{d}_n(\widehat{\theta}_n)]^{-1}(\widehat{\theta}_n - \theta^c) \xrightarrow{\mathcal{L}} \chi^2(2), \quad \text{as} \quad n \to \infty,$$

which implies, denoting by $\chi^2_{1-\alpha}(2)$ the $(1-\alpha)$-quantile of the χ^2-distribution with two degrees of freedom, that

$$P((\widehat{\theta}_n - \theta^c)^T [\ddot{d}_n(\widehat{\theta}_n)]^{-1}(\widehat{\theta}_n - \theta^c) \leq \chi^2_{1-\alpha}(2)) \simeq 1-\alpha, \quad \text{as} \quad n \to \infty.$$

We deduce from the above convergence the $(1-\alpha)$-asymptotic elliptical region of confidence:

$$\mathcal{R}_{1-\alpha} = \left\{\theta \in \mathbb{R}^2 : \ (\widehat{\theta}_n - \theta)^T [\ddot{d}_n(\widehat{\theta}_n)]^{-1}(\widehat{\theta}_n - \theta) \leq \chi^2_{1-\alpha}(2)\right\}.$$

Finally a *green light criterion* to proceed to the test could be the checking of the condition:

$$\min_{\theta \in \{1\}\times]0,1[\cup]0,1[\times\{1\}} (\widehat{\theta}_n - \theta)^T [\ddot{d}_n(\widehat{\theta}_n)]^{-1}(\widehat{\theta}_n - \theta) \leq \chi^2_{1-\alpha}(2),$$

ensuring that the region $\mathcal{R}_{1-\alpha}$ intersects the $]0,1[^2$ proportions domains. Practically speaking, this green light criterion happens to be very useful for time saving purposes, since if it is not satisfied no tabulation of the limiting random variable $\int_{\mathbb{R}}(W_3(\theta^c, x))^2 dU(x)$ is then required.

Remark 2. Although the same underlying ideas can be used, from the identifiability and asymptotic results perspective, the case where $G_1 = G_2$ is slightly different and requires a picking trick as the parameters are no longer identifiable even under H_0 (see Section D in the Supplement for further details).

5. Convergence Monte Carlo assessment

In this section along with the next Sections 6 and 7 we propose, based on the rule of thumb discussion Section F of the Supplement, to use $U = U_{Unif}$ where U_{Unif} denotes the uniform distribution over the

observations range interval $[\min(X_{1,(1)}, X_{2,(1)}), \max(X_{1,(n_1)}, X_{2,(n_2)})]$. Note that in order to be compliant with our theoretical results the weight function U should not depend on n but since our numerical experiments are done for n small to moderately large we think this integration strategy is acceptable in a finite sample setup (Cauchy or Gaussian distributions with large variance are compliant with the theory and numerically good alternatives).

Consider now the random vector $(P_1, P_2, D_z)^T = \sqrt{n}\,(\widehat{p}_1 - p_1^c, \widehat{p}_2 - p_2^c, D_n(z) - D(z))^T$ at any point $z \in \text{support}(X_1, X_2)$. Recall that z represents one single location point of the empirical process trajectory. Theorem 2 states that this vector is asymptotically Gaussian, and that its first two components are consistent towards θ^c (both under H_0 or H_1). Our goal in this section is to check/illustrate this asymptotic result by comparing numerical approximations of our theoretical expressions to Monte Carlo experiments. For this aim, we consider $K = 200$ simulations of two samples X_1 and X_2 with cdfs given by (4), both generated from two-component mixtures of Gaussian distributions. More precisely, the k-th simulation provides X_1^k and X_2^k, $k = 1, ..., K$, where X_1^k and X_2^k are respectively drawn from mixtures with parameters $n_1 = n_2 = 5{,}000$, $p_1^* = 0.4$, $p_2^* = 0.6$, $F_1 = F_2$ are $\mathcal{N}(1, 1)$ cdfs, when G_1, G_2 are respectively $\mathcal{N}(2, 0.7)$ and $\mathcal{N}(3, 1.2)$ cdfs. Note that we are here under the null, but keep in mind that such comparisons were also made on very different setups involving H_1-type setups, with $n_1 \neq n_2$ and distributions supported over $\mathbb{R}^+$, $\mathbb{N}$ or bounded intervals of $\mathbb{R}$, see Figures 4 and 5. Estimating $\theta = (p_1, p_2)$ by $\widehat{\theta}_n^k = (\widehat{p}_1^k, \widehat{p}_2^k)$ from each of the K simulated couples (X_1, X_2), we obtain an empirical average $\bar{\theta}_n^K = n^{-1} \sum_{k=1}^K \widehat{\theta}_n^k$ which observed value is equal to $(0.402, 0.601)$ and an empirical variance equal to $1/\sqrt{n}(2.03, 1.63) = (0.0287, 0.0230)$, see Section B of the Supplement, illustrating the asymptotic consistency of our estimators. A Kolmogorov-Smirnov tests on the components of the vector $(P_1, P_2, D_z)^T$ validates that the three estimators are asymptotically Gaussian, with p-values always greater than 0.7. To validate the explicit covariance structure between the estimators, it is necessary to fix z and to compare the empirical covariances (computed from the Monte Carlo simulations) to the theoretical ones. Appendix B shows the obtained results in the aforementioned parametric setup for $z = 2$. Clearly, all tests made through these comparisons for different values of z show the validity of formulas (34)-(37), which confirms the theoretical consistency given in part i) of Theorem 2.

It now remains to have a closer look at the behaviour of the statistic $\mathsf{T}_n = n d_n(\widehat{\theta}_n)$, see formulas (25) and (26). The theorem states that the empirical distribution of U_n^0 under H_0 (obtained through the Monte Carlo procedure providing as many realisations of $n d_n(\widehat{\theta}_n)$ as the K experiments) should converge to some explicit random variable $Z(\theta^*)$. Also, the same kind of regime for U_n^1 should be observed under H_1. However, in the latter case, the discrepancy measure dramatically increases due to the term V_n^1, pointing the departure from the null hypothesis. This phenomenon is well illustrated in Figure 3, where one can see that the empirical distributions suit the expected behaviours provided by the random

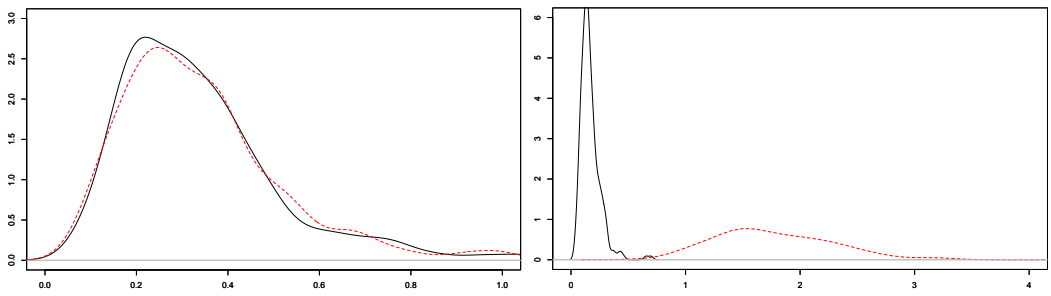

Figure 3. On the left panel, theoretical distribution $Z(\theta^*)$ (solid) and empirical version U_n^0 (dotted). On the right panel, distribution $Z(\theta^c)$ (solid) and empirical contrast distribution $U_n^1 + V_n^1$ (dotted).

variables $Z(\theta^*)$ and $Z(\theta^c)$. Indeed, under H_1, the empirical distribution of $nd_n(\widehat{\theta}_n)$ is far from $Z(\theta^c)$, showing the impact of the drift V_n^1, see Equation (40). This way, the tabulated distributions of the limiting random variables $Z(\theta^*)$ and $Z(\theta^c)$ and their respective $(1-\alpha)$-quantile can be used to fruitfully answer our testing problem.

6. Test performances

In this section, we study the empirical levels and powers of the test in various situations. In that perspective, we generate X_1 and X_2 from (4) on various supports, and the behaviour of our test is investigated over a range of (more or less) challenging setups. Depending on the case under study, mixture components can be easily detected or not, either because of the importance of the mixture weight p_i, $i = 1, 2$, or due to the specified mixture components features. The idea is to get some insights about the strengths and weaknesses of our testing procedure. Each time we evaluate the empirical level, respectively power, of the test, the 95-th percentile of the test is previously estimated using 150 trajectories of the Gaussian process embedded in the stochastic integral appearing in the right-hand side of (25) and (26). Then the testing procedure (13) is performed K times to get the result, through the K simulations of the k-th samples X_1^k and X_2^k and the associated test statistic $nd_n(\widehat{\theta}_n^k)$, $k = 1, \ldots, K$. Here, we take $K = 100$, fix equal sample sizes $n_1 = n_2 = n$ for conciseness and make n vary.

6.1. Empirical levels ($F_1 = F_2 = F$)

The distributions considered here are Gaussian-Gaussian mixtures on $\mathbb{R}$, Gamma-Exponential on $\mathbb{R}^+$, Negative-Binomial-Poisson on $\mathbb{N}$, and Logit-Uniform on $[0, 1]$. To figure out whether the test remains significant in real-life situations, we have chosen to make the component weights p_i, $i = 1, 2$, vary from 10% to 60%. The asymptotic properties of the test can be checked by considering different values of the sample size n, ranging from 500 to 10,000. However, our experiments show that the number of observations does not have a big impact on the level of the test, provided that there are at least around 300 observations for the mixture component to be tested. This is why we decided to display only the results corresponding to $n = 2,000$ observations, which lightens the presentation.

For each support ($\mathbb{R}, \mathbb{R}^+, \mathbb{N}, [0, 1]$), four very different setups are studied (see Figure 1 in Section E.1 of the Supplement, with corresponding mixture parameters stored in Table 1). We will denote from (a) to (d) these four different cases, corresponding to: (a) G_1 not so different from G_2, and G_1 and G_2 close to F; (b) G_1 very different from G_2 with F "in between"; (c) G_1 not so different from G_2, with both distributions far from F; (d) G_1 very different from G_2, with G_1 close to F and G_2 far from F. The global simulation scheme thus encompasses overall 144 different setups (4 supports, 4 cases, and 9 combinations for p_1 and p_2). We recall that for each of these 144 possibilities, the testing procedure (13) is performed 100 times, which provides an approximation of the empirical level of the test in all of the aforementioned situations.

The overall results are summarized thanks to the heatmap displayed in Figure 4, with dark zones pointing to unsatisfactory results. For one given support, the four panels from top left to bottom right correspond to cases (a) to (d). For instance, case (c) with mixtures of Gaussian distributions is the bottom left 3x3 square of the heatmap. One can see that most setups under study lead to satisfactory empirical levels of the test, close to the theoretical 5%. Indeed, since each simulation enables to compare the empirical test statistic to the 95-th percentile of the calibrated distribution $Z(\theta^*)$, it is expected that the level of the test fluctuates around 5%. In practice only 12 over 144 approximations of the level exceed 10%, which means that less than 9% of the setups under study provide mixed-up conclusions.

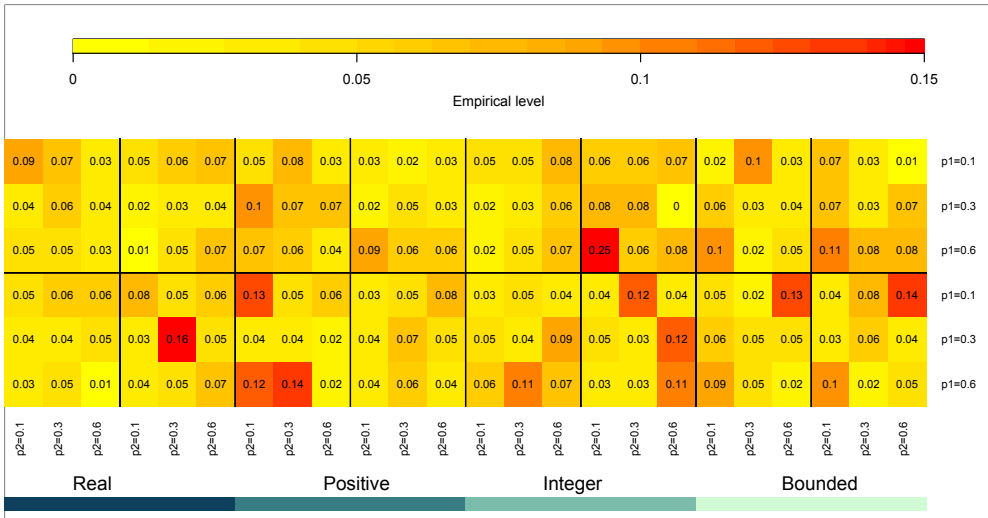

Figure 4. Heatmap of empirical level (under H_0) for different supports, different component weights, and different parameters for component distributions. For each support, cases (a) to (d) are given from top left to bottom right.

Looking more carefully at the results, the concerning situations mostly arise when at least one of the proportions p_i equals 10%. It is very likely that the main reason explaining this drop of efficiency is the lack of observations to perform the test about the unknown components. The low component weight assigned to the unknown part of the distribution leads to under-represent the observations useful for the test to be enough informative. In some very rare setups, although p_1 and p_2 equal at least 30%, the empirical level remains "high" (e.g. the case of mixing Negative Binomial and Poisson distributions, case (d), with $p_1 = 0.3$ and $p_2 = 0.6$, where the empirical level equals 12%). In such cases, the choice of the mixture components parameters (Table 1 in Section E.1 of the Supplement) has a crucial impact and can affect the estimation of the component weights, which impacts the overall quality of the test.

6.2. Empirical powers

In the same spirit, one can analyse the heatmap that illustrates the empirical power of the test in Figure 5, still considering the same previous mixture distributions. However, the difference here lies in the different considered type of departure setups, as illustrated in Figure 2 of Section E.2. Hereafter, we denote them as follows: case (a) F_1 and F_2 have the same distribution, with very different means; case (b) F_1 and F_2 have the same distribution, with close means; case (c) F_1 and F_2 have the same distribution, with same means but very different variances; case (d) F_1 and F_2 have the same distribution, with same means and close variances. We obviously expect here that the most difficult case to be detected is the latter one.

Here, the sample size has a major impact on the results, which explains why the heatmap is provided for results corresponding to a sensitively higher sample size $n = 3,000$. To understand how crucial the number of observations is, Figure 6 depicts the connection between the empirical power of the test and the sample size n. In fact we can observe very heterogeneous behaviours depending on the support and component weights, especially in the case where alternatives are very difficult to distinguish, which happens when F_1 and F_2 have the same two first order moments (see case (e) in Table 2 of Section E.2

for further details about mixture distributions and parameters). Not surprisingly, departures from the null hypothesis can be detected provided that the number of observations is large enough, otherwise the power of the test remains pretty low (especially when F_1 and F_2 are very similar, see cases (b) and (d)). Indeed, low proportions p_i, $i = 1, 2$, lead to deteriorate the accuracy of the estimates $\widehat{p_i}$, which favour situations where $\widehat{\theta}_n$ can be remote from θ^c (minimisation of the contrast is solved by escaping from $]0, 1[^2$). The consequence of this phenomenon is that extreme quantiles (e.g. 95-th percentile) of the tabulated random variable representing $Z(\theta^c)$ tend to be larger, which mechanically lowers the power of the test.

7. Application to COVID-19 excess mortality

There is an abundant literature investigating the impact of the COVID-19 on the mortality across countries, see for instance Beaney et al. (2020) or Kontis et al. (2020). We generally witness a wide variation in mortality across countries, leading to questioning the extent to which one can proceed to pairwise comparative studies. In our application, we will be looking at the nodular impact of the COVID-19 and compare the latter across a panel of European countries. Formally, we investigate the age distribution of deaths (the distribution of the proportion of deaths per age group among all deaths) and study the changes between 2019 and 2020 for France, Belgium, Germany, Italy, Netherlands and Spain from the Short-Term Mortality Fluctuations (STMF) data series compiled by the Human Mortality Database (HMD). The datasets contain death records aggregated over age groups: 0-14, 15-64, 65-74, 75-85 and 85+. We restrain our study to the four last age classes (given that experts agree to consider that the first one 0-14 was clearly not affected by the pandemic), and to the first 25 weeks of each considered year as shown in the first graph of Figure 1 when the second graph of Figure 1 shows the distribution of the proportion of deaths per age class for years 2019 and 2020 (total of proportions equals to 1), indicating the empirical probability for a death to happen in each age class.

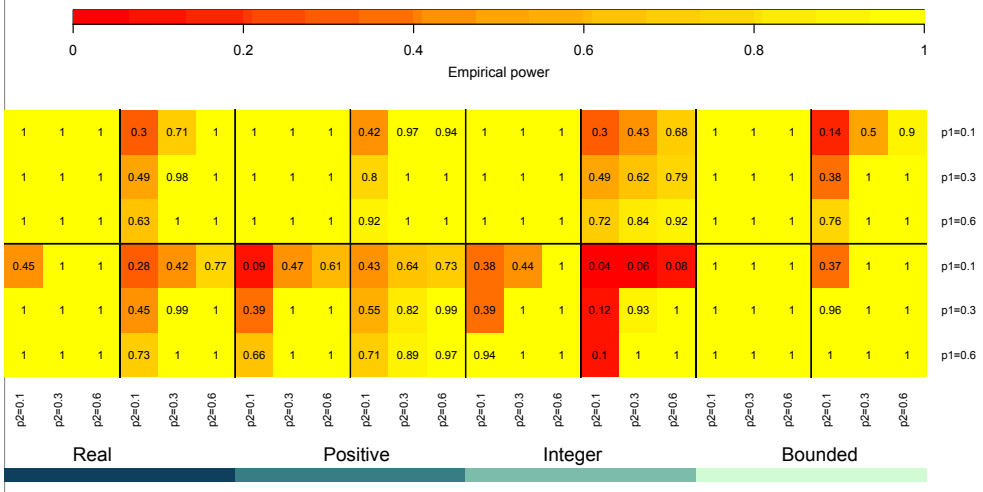

Figure 5. Heatmap of empirical power (under H_1) for different supports, different component weights, and different parameters for component distributions. For each support, cases (a) to (d) are given from top left to bottom right.

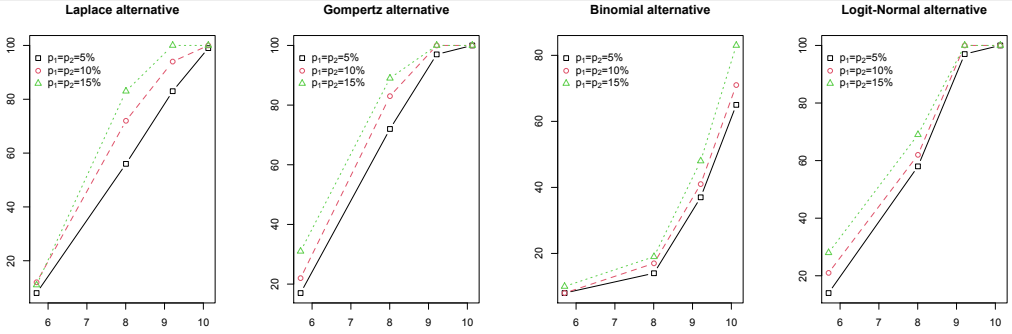

Figure 6. Empirical power depending on sample size n (300; 3,000; 10,000; 25,000) in logarithmic scale, on various supports ($\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{N}$, $[0,1]$), when F_1 and F_2 have same mean and variance (parameters listed in Table 2 of Section E.2, case (e), see also Figure 3).

It is assumed, as explained in expression (1), that the differences in the observed mortality between 2019 and 2020 is imputed (directly or indirectly) to the COVID-19. The 2020 population is then a two-component mixture composed by the previous 2019 population plus a latent one subject to the impact of the COVID-19 crisis. In other words, model (2) has an appealing application to capture the excess of mortality due the COVID-19. It is then legitimate to assume a second unknown nodular component driving the mortality due to the COVID-19 during the considered period. More precisely, we will assume that the known cdf is the one observed over 2019, *i.e.* the multinomial distribution G, and aim to compare the distribution F of the mortality excess across countries. This mortality excess can be regarded as a measure that encompasses all causes of death and provides a metric of the overall mortality impact in 2020.

In Table 1 we report the outputs of the testing procedure developed in this paper for the aforementioned countries. The known component G_i is described as the multinomial distribution computed in 2019 for each country. We shall stress out that in this application we are clearly in presence of two distinct known cdfs, *i.e.* $G_1 \neq G_2$, which is our basic assumption to implement our procedure, and will

Population		p_1	p_2	Green light	Test statistic	95% quantile	p-value	Decision
1	2							
Belgium	Spain	0.1453	0.1447	valid	0.1152	10.3854	0.95	H_0
Belgium	France	3	0.9556	non valid	-	-	-	H_1
Belgium	Germany	0.7263	0.1518	valid	0.7337	0.8922	0.08	H_0
Belgium	Italy	0.2620	1.0181	non valid	-	-	-	H_1
Belgium	Netherlands	0.8425	3	non valid	-	-	-	H_1
Spain	France	3	0.2956	non valid	-	-	-	H_1
Spain	Germany	2.0088	0.1189	non valid	-	-	-	H_1
Spain	Italy	0.3561	3.0000	valid	-	-	-	H_1
Spain	Netherlands	3	3	non valid	-	-	-	H_1
France	Germany	0.2209	0.1036	valid	14.0715	12.6294	0.05	H_1
France	Italy	0.1508	3	non valid	-	-	-	H_1
France	Netherlands	0.3530	3	non valid	-	-	-	H_1
Germany	Italy	0.1003	3	non valid	-	-	-	H_1
Germany	Netherlands	0.1943	3	non valid	-	-	-	H_1
Italy	Netherlands	0.1944	0.1876	valid	0.7946	5.4099	0.63	H_0

Table 1. Pairwise testing of the mortality excess across a panel of European countries.

assume that conditions **(I)** and **(II)** are satisfied for x_i's picked within $\{1, 2, 3, 4\}$. Also, in this case, we choose the discrete uniform distribution for the integrating cdf U, see (8), and we set the upper bound of the parametric space equal to $\delta_2 = 3$, which seems to be large enough given our previous simulation results.

First, we can see from Table 1 that some estimated proportions in the pairwise analysis bump into this boundary, which clearly indicates that we must reject the null hypothesis H_0. Among these countries, we see that only Belgium and Italy (with p-value equal to 95%), Belgium and Germany (8%) and Italy and the Netherlands (63%) possibly share the same mortality excess profile. Although the French and German pairwise test passes the green light criterion, the equality between their nodular effect is rejected. These countries, all showed, historically, a significant peak in excess mortality around early April with a return to normal levels of deaths by mid-May, see the left panel of Figure 1. However, as a result of our testing hypothesis, the first wave has not the same impact in terms of excess of mortality, over the underlying population.

When we look at the decontaminated distributions based on the cdfs, *i.e.* $\widehat{F}_i(\cdot) = F_i(\cdot, \widehat{L}_i, \widehat{p}_i)$, $i = 1, 2$, in Figure 7, we can see the very close patterns of the two multinomial distributions between Belgium and Spain, on one hand, and Italy and the Netherlands, on the other hand; which is consistent with the hypothesis testing outputs reported in Table 1 and in particular the level of the p-value, *i.e.* 95% and 63% respectively. Eventually, the proportion for the impact of the crisis are consistent with the reported statistics over the first wave, see Beaney et al. (2020) and Mannucci, Nreu and Monami (2020).

On the other hand, when we compare these estimated cdfs for Belgium and France against Germany, we can see that the estimated impact of the COVID-19 is far from being homogeneous between these countries. Indeed, Belgium and Germany exhibit the same proportion of individuals impacted by the crisis within the age bands $[75, 84]$ and +85, but have disparities for the other age classes. Although the null hypothesis is accepted, this notable difference may explain the low level of the corresponding p-value, *i.e.* 8%. Similar conclusions can be drawn for the comparison between France and Germany but with a more pronounced difference over ages less than 74, which may explains that the null hypothesis is rejected in this case. These disparities may be explained by demographic variables and most notably the age pyramid of the underlying population. This should can also be regarded as a consequence of the measures put in place to contain the impact of the COVID-19. Indeed, Germany has been cited for early widespread testing with less restrictive lock-down measures. Also, the German health care was one the most well-equipped on intensive care units among European countries and thus can be seen as a key element for explaining the impact of the pandemic over the age band $[65, 74]$. The fact that France, Germany and Spain exhibit similar pattern for ages beyond 75 should be explored. The discussion of such a behaviour is, however, beyond the scope of this paper. Instead, we can refer to the various discussions in the literature that intended to understand the differential impact of the COVID-19 crisis over countries looking at the socioeconomic, demographic variables and the measures put in place to contain the pandemic.

For information, all the results presented in this application are reproducible using the source code available at https://cran.r-project.org/web/packages/admix/index.html, after having installed the R package `admix`.

8. IBM-method and further models

In the next two sections we propose to highlight on the range of our method, to describe challenging situations (involving dependencies) in which our semiparametric IBM-method could provide interesting results. These are ongoing works which are beyond the scope of the current paper.

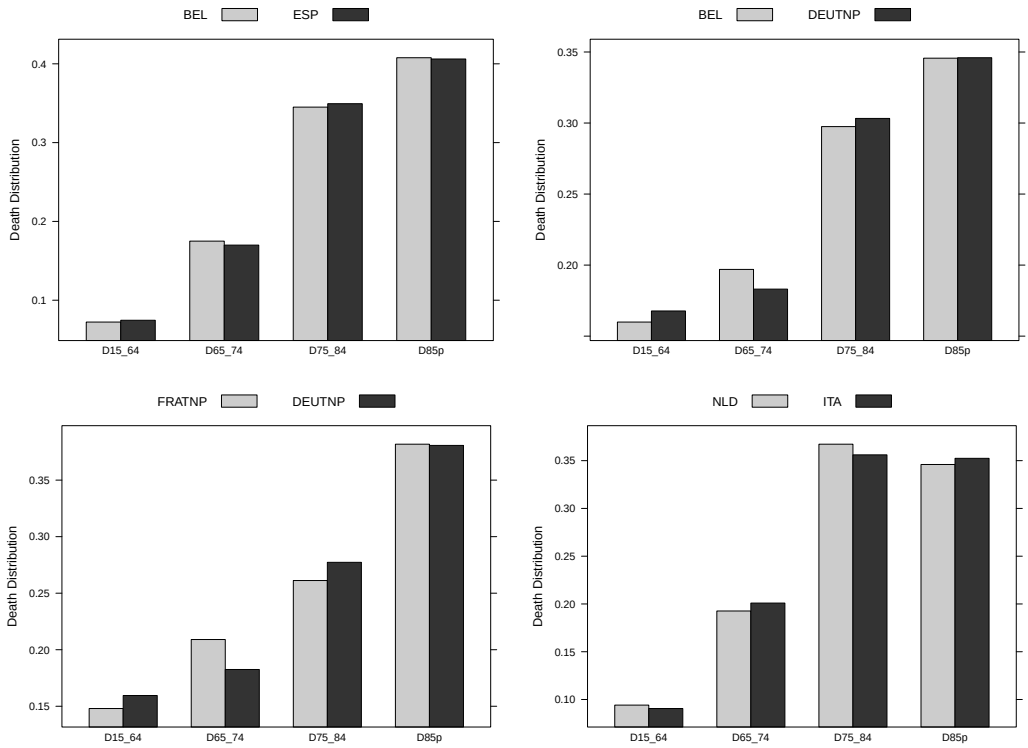

Figure 7. Pairwise COVID-19 nodular distribution estimates deduced from the $\widehat{F_i}$, $i = 1, 2$, for the couple of countries satisfying the green light criteria and/or the hypothesis test.

8.1. Independence, concordance and discordance

Let us consider for simplicity a bivariate contamination model (extension to the d-variate setup, $d \geq 3$, being straightforward):

$$L(x_1, x_2) = pG(x_1, x_2) + (1 - p)F(x_1, x_2), \quad (x_1, x_2) \in \mathbb{R}^2, \tag{41}$$

where L is the common cdf of an i.i.d. sample $(X_1, \ldots, X_n)$, G is a known cdf when the mixture proportion p and the cdf F are both unknown. By splitting the observation vector X into 2 components $X = (X_1, X_2)^T$, we have respective marginal cdfs

$$L_i(x) = pG_i(x) + (1 - p)F_i(x), \quad x \in \mathbb{R}, \quad i = 1, 2. \tag{42}$$

An interesting problem is then to test the mutual independence of the nodular components X_1 and X_2, *i.e.*

$$H_0 : F = F_1 \otimes F_2 \qquad \text{against} \qquad H_0 : F \neq F_1 \otimes F_2, \tag{43}$$

where $G \neq G_1 \otimes G_2$ on a $\mu^{\otimes 2}$-non null set to avoid trivial testing situations (otherwise independence on the L-components would then reflect the independence on the F-components). Given the above

remarks we can define two parametric families (Inversion step):

$$\mathcal{F}_1 = \left\{ F(u_1,u_2;p) = \frac{L(u_1,u_2) - pG(u_1,u_2)}{1-p}, \quad p \in]0,1[\right\}, \qquad \text{and}$$

$$\mathcal{F}_2 = \left\{ F_{1\times 2}(u_1,u_2;p) = F_1(u_1;p)F_2(u_2;p), F_i(\cdot;p) = \frac{L_i(\cdot) - pG_i(\cdot)}{1-p},\ i=1,2,\ p \in]0,1[\right\},$$

and build a contrast function (Best Matching step) in the spirit of (8–9)

$$d(p) = \int_{\mathbb{R}\times\mathbb{R}} (F(u_1,u_2;p) - F_{1\times 2}(u_1,u_2;p))^2 dU(u_1,u_2).$$

Using copula techniques to handle global and marginal empirical processes, as it is classically done in the "direct" (not mixture component testing) Cramér-von Mises independence testing literature, see Genest et al. (2019) for recent results and bibliography, we reasonably think that asymptotic decision results similar to ii) in Theorem 2 could be established on the test statistic $nd_n(\widehat{p}_n)$ where d_n is the empirical version of d and $\widehat{p}_n$ is the minimum argument of d_n over $]0,1[$. Note that such accomplishment would also help in answering/testing the complete concordance/discordance problem arising in z-score analysis, see Lai et al. (2007, 2017), where basically model (41) can take (among others) two basic forms:

$$L = pG_1(x_1) \otimes G_2 + (1-p)F_1 \otimes F_2, \qquad \textit{(complete concordance)},$$
$$L = (p_1G_1 + (1-p_1)F_1) \otimes (p_2G_2 + (1-p_2)F_2), \quad \textit{(complete discordance)}.$$

In fact in the above models, slightly more complex contrast functions d, based on F-inversions and comparison inspired from the previous independence testing strategy, can also be proposed and proved to provide a fully tractable Cramér-von Mises test in the spirit of Theorem 2.

8.2. Blending process

As mentioned earlier, the testing methodology we introduce in this paper can be extended to temporal contamination models we propose to name *blending* process. This type of model are especially interesting to analyse situations in which a phenomenon has been observed with a good stability for a long period of time but turns out to be contaminated by a new trend which importance becomes more and more prominent. This type of model would be especially relevant to analyse temporal mortality datasets during the COVID-19 crisis as described in Section 7 (collections of mortality datasets over time would be required instead of one single sample collected during a given period of time). By denoting G the cdf of the well-known phenomenon and by p_t, respectively F_t, the proportion, respectively the cdf, of the new trend at time t, the distribution of a generic i.i.d. sample $X^t = (X_1^t, \ldots, X_{n_t}^t)$ at time $t \in \mathbb{N}$ could be expressed as follows:

$$L_t(x) = p(t)G(x) + (1-p(t))F_t(x), \quad x \in \mathbb{R}. \tag{44}$$

In that setup it could be interesting, following the identifiability and parameter picking strategy presented in Section D of the Supplement, when $G_1 = G_2$, to test the consistency in time of the trend distribution, *i.e.*

$$H_0(t_i,t_j) : F_{t_i} = F_{t_j} \qquad \text{against} \qquad H_1(t_i,t_j) : F_{t_i} \neq F_{t_j}, \quad i \neq j \in \{1,\ldots,T\}. \tag{45}$$

Note that if the testing problem (45) is very mostly answered positively we could possibly assess that F is independent from t and then estimate nonparametrically the mixing proportion function $p(t)$ based on the condition $F_t = F$, $t \in \mathbb{N}$ and the Remark in Section D of the Supplement. The main technical difficulty here is to handle correctly the possible dependencies between samples X^{t_i} and X^{t_j}, for $i \neq j$, especially when t_i and t_j are close. Note that the analogue of (65) and (72) will certainly involve a limiting bivariate Gaussian process $\mathcal{B}$ with no longer independent coordinates since the source samples X^{t_i} and X^{t_j} are dependent. The paper by Gribkova and Lopez (2015) on nonparametric copula estimation under bivariate censoring looks to provide interesting ideas to investigate this problem.

9. Conclusion

In this work, we address the comparison testing of the unknown components of a two-sample contamination model. We introduce for this purpose the so-called IBM (Inversion-Best Matching) approach that results into a relaxed and tuning parameter-free semiparametric Cramér-von Mises type two-sample test with very minimal assumptions about the unknown components. Indeed, we do not require any shape constraints on the unknown distributions, such as symmetry, tail conditions etc. which are commonly key technical identifiability conditions arising in univariate semiparametric mixture models. We establish in particular a functional joint central limit theorem on the proportion parameters (with consistency under H_0) along with the best fitted differences between the unknown cdfs, which is unachievable in the basic univariate case as shown by Patra and Sen (2016). An intensive numerical study has been carried out from a large range of simulation setups to illustrate the asymptotic properties of our test. This includes examples using Gaussian distributions but also more challenging distributions supported by $\mathbb{R}^+$, $\mathbb{N}$ or $[0, 1]$ which are considered as very non-standard in the mixture models literature. Finally, our testing procedure is applied to a real-life case attempting to fill the gap in understanding the disparities of the excess of mortality during the COVID-19 crisis, which allows to test pairwise the mortality excess across a panel of European countries.

This work could be extended in many interesting ways such as: i) the coordinate independence testing in the multivariate contamination model along with the concordance or discordance hypothesis testing crucial in z-scores modelling; ii) the homogeneity in time testing for temporal contamination models (blending processes). Also, returning to the COVID-19 case, it is important to develop a more adapted scheme for pairwise contamination comparison at large scale in case of a massive amount of countries to deal with. In fact, a clustering procedure would be beneficial along with a K-sample testing procedure based on the results we developed in this paper. This could bring in a new challenging model-based clustering problem. Finally, given the ability of the test to accommodate very different frameworks, we developed the `admix` R package (https://cran.r-project.org/web/packages/admix/index.html) implementing a wide variety of two-sample testing methods for contamination/admixture models available to researchers and practitioners.

References

BEANEY, T., CLARKE, J. M., JAIN, V., GOLESTANEH, A. K., LYONS, G., SALMAN, D. and MAJEED, A. (2020). Excess mortality: the gold standard in measuring the impact of COVID-19 worldwide? *J. R. Soc. Med.* **113** 329–334.

BORDES, L., DELMAS, C. and VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Stat.* **33** 733–752.

BORDES, L., MOTTELET, S. and VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Stat.* **34** 1204–1232.

BORDES, L. and VANDEKERKHOVE, P. (2010). Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Math. Methods Stat.* **19** 22–41.

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.

BUTUCEA, C. and VANDEKERKHOVE, P. (2014). Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.* **41** 227–239.

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.

CELISSE, A. and ROBIN, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Stat. Plan. Inference* **140** 3132–3147.

CHAKRABORTY, R. and WEISS, K. M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *PNAS* **85** 9119–9123.

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet.* **23** 70–86.

GENEST, C., NEŠLEHOVÁ, J., RÉMILLARD, B. and MURPHY, O. (2019). Testing for independence in arbitrary distributions. *Biometrika* **106** 47–68.

GRIBKOVA, S. and LOPEZ, O. (2015). Non-parametric Copula Estimation Under Bivariate Censoring. *Scand. J. Stat.* **42** 925–946.

HIGHAM, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review* **43** 525–546.

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.

KONTIS, V., BENNETT, J. E., RASHID, T., PARKS, R. M., PEARSON-STUTTARD, J., GUILLOT, M., ASARIA, P., ZHOU, B., BATTAGLINI, M., CORSETTI, G. et al. (2020). Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. *Nat. Med.* 1–10.

LAI, Y., ADAM, B.-L., PODOLSKY, R. and SHE, J.-X. (2007). A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups. *Bioinformatics* **23** 1243–1250.

LAI, Y., ZHANG, F., NAYAK, T. K., MODARRES, R., LEE, N. H. and MCCAFFREY, T. A. (2017). An efficient concordant integrative analysis of multiple large-scale two-sample expression data sets. *Bioinformatics* **33** 3852–3860.

LOH, P.-R., LIPSON, M., PATTERSON, N., MOORJANI, P., PICKRELL, J. K., REICH, D. and BERGER, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193** 1233–1254.

MANNUCCI, E., NREU, B. and MONAMI, M. (2020). Factors associated with increased all-cause mortality during the COVID-19 pandemic in Italy. *Int. J. Infect. Dis.* **98** 121–124.

MCLACHLAN, G. J., BEAN, R. and JONES, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22** 1608–1615.

MILHAUD, X., POMMERET, D., SALHI, Y. and VANDEKERKHOVE, P. (2022). Semiparametric two-sample admixture components comparison test: The symmetric case. *J. Stat. Plan. Inference* **216** 135–150.

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.

PATRA, R. K. and SEN, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc., B: Stat. Methodol.* 869–893.

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.

POMMERET, D. and VANDEKERKHOVE, P. (2019). Semiparametric density testing in the contamination model. *Electron. J. Stat.* **13** 4743–4793.

SUESSE, T., RAYNER, J. C. and THAS, O. (2017). Assessing the fit of finite mixture distributions. *Aust. N.Z. J. Stat.* **59** 463–483.

TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* 1265–1269.

VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

WALKER, M. G., MATEO, M., OLSZEWSKI, E. W., 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.

XIANG, S., YAO, W. and YANG, G. (2019). An overview of semiparametric extensions of finite mixture models. *Stat. Sci.* **34** 391–404.