Bernoulli **18**(2), 2012, 606–634
DOI: 10.3150/11-BEJ351

Similarity of samples and trimming

PEDRO C. ÁLVAREZ-ESTEBAN[1,*], EUSTASIO DEL BARRIO[1,**], JUAN A. CUESTA-ALBERTOS[2] and CARLOS MATRÁN[1,†]

[1]*Dept. de Estadística e Investigación Operativa, Universidad de Valladolid, Prado de la Magdalena s.n., 47005 VALLADOLID, Spain. E-mail: *pedroc@eio.uva.es; **tasio@eio.uva.es; †matran@eio.uva.es*
[2]*Dept. Matemáticas, Estadística y Computación, Universidad de Cantabria, Avda. los Castros s.n., 39005 SANTANDER, Spain. E-mail: juan.cuesta@unican.es*

We say that two probabilities are similar at level α if they are contaminated versions (up to an α fraction) of the same common probability. We show how this model is related to minimal distances between sets of *trimmed probabilities*. Empirical versions turn out to present an *overfitting* effect in the sense that trimming beyond the similarity level results in trimmed samples that are closer than expected to each other. We show how this can be combined with a bootstrap approach to assess similarity from two data samples.

Keywords: asymptotics; bootstrap; consistency; mass transportation problem; over-fitting; robustness; similarity of distributions; trimmed probability; Wasserstein distance

1. Similarity vs. homogeneity

Classical goodness of fit deals with the problem of assessing whether the unknown random generator, P, of a data object, X, belongs to a given class, $\mathcal{F}$. This includes two-sample problems in which two different random objects are observed. We focus on checking whether a certain feature of the corresponding random generators coincides. The case in which X_1 is a collection of i.i.d. random variables $X_1^1, \ldots, X_n^1$ with common distribution P_1, X_2 is another sequence of i.i.d. random variables $X_1^2, \ldots, X_m^2$ with law P_2 and the goal is to assess whether $\theta(P_1) = \theta(P_2)$ for some function $\theta(\cdot)$ (including, for instance, $\theta(P) = P$) is a *homogeneity* problem, to which a large amount of literature has been devoted. Our starting point is that it is often the case that the researcher is not really interested in checking whether $P \in \mathcal{F}$ or whether $P_1 = P_2$. Imagine the case of a pharmaceutical company trying to introduce a new (and cheaper) alternative to some reference drug. The regulatory authorities will approve the new drug if its performance with respect to a certain biological magnitude does not differ from that of the standard drug. Both drugs could produce a similar outcome on most patients. However, if there is a fraction of them for whom the results are clearly different, then the new drug is very likely to be rejected by a homogeneity test, while, in fact, it has a similar performance for most individuals. As another example, consider the comparison of two human populations that were initially equal but have received immigration with different patterns. In these situations the relevant assumption to check is not homogeneity, but rather *similarity* in the following sense.

Definition 1. *Two probability measures P_1 and P_2 on the same sample space are α-similar if there exist probability measures P_0, P_1', P_2' such that*

$$\begin{cases} P_1 = (1-\varepsilon_1)P_0 + \varepsilon_1 P_1', \\ P_2 = (1-\varepsilon_2)P_0 + \varepsilon_2 P_2' \end{cases} \tag{1}$$

with $0 \le \varepsilon_i \le \alpha$, $i = 1, 2$.

Definition 1 measures the overlap between P_1 and P_2, in agreement with other possible measures of similarity (see the section "Similarity between Populations" in [14]). Beware that smaller values of α in Definition 1 correspond to more similar distributions (the case $\alpha = 0$ being equivalent to $P_1 = P_2$).

A related situation, for one-sample problems, would be the case when we observe some random object X with law P_1. Ideally, P_1 should equal P_0 (some gold standard), but the presence of noise means that, in fact,

$$P_1 = (1-\varepsilon)P_0 + \varepsilon N, \qquad \varepsilon \le \alpha \tag{2}$$

for some unspecified N if we assume that the noise level does not exceed α. We would say that P_1 is *similar* to P_0 at level α if (2) holds (observe that P_1 and P_0 do not play a symmetric role in this definition). In two-sample problems, we want to assess whether the two samples can be assumed to be noisy realizations of some unkown gold standard, as in Definition 1. Model (2) corresponds to the 'contamination neighborhoods' introduced in Huber [15,16] in a robust testing setup. We discuss further connections to these and other related references in Section 2.2 below. Our goal in this work is to present a method for assessing similarity of the unknown random generators P_1, P_2 of two independent i.i.d. samples. Our procedure also yields an estimate of the *common core* of the two distributions.

Our approach is based on trimming. Trimming procedures are of frequent use in robust statistics as a way of downplaying the influence of contaminating data in our inferences. The introduction of data-dependent versions of trimming, often called impartial trimming, allows us to overcome some limitations of earlier versions of trimming that simply removed extreme observations at tails. Generally, impartial trimming is based on some optimization criterion, keeping the fraction of the sample (of a prescribed size) that yields the least possible deviation with respect to a theoretical model. Today, impartial trimming constitutes one of the main tools in the robust approach to a variety of statistical settings (see [9,12,18,23]). The first approach to model validation based on impartial trimming is (to the best of our knowledge) the one in Álvarez-Esteban *et al.* [1,3]. The problem considered there can be rephrased as follows. Given two independent i.i.d. samples of univariate data with unknown random generators P_1, P_2, we want to assess whether $P_i = \mathcal{L}(\varphi_i(Z))$, $i = 1, 2$, for some random variable Z defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and non-decreasing functions, φ_1, φ_2, such that

$$\mathbb{P}\big(\varphi_1(Z) \ne \varphi_2(Z)\big) \le \alpha$$

(see Section 2.2 for further discussion). Despite the interest of this approach, we believe that the similarity model given by Definition 1 is often more natural and useful in applications. Some technically related results and the connection with the optimal transportation problem have been reported in Álvarez-Esteban *et al.* [2]. A related approach based on density estimation can be found in Martínez-Camblor *et al.* [19].

As we will show in Section 2, the similarity model of Definition 1 can be expressed in terms of a minimal distance between the sets of *trimmings* of the probabilities P_i, $i = 1, 2$. These are the sets of probabilities that one obtains from a fixed one by removing or downplaying (to some degree) the weight assigned by the original probability. When we look for the minimal distance between trimmings of the empirical measures based on two samples, we are highlighting the part of the data that, hopefully, comes from the common core P_0. From a descriptive point of view, this gives an interesting tool for the comparison of data samples.

A distinctive feature of our proposal concerns the rates of convergence. If P_n, Q_n are the empirical distributions based on two samples of univariate data (of equal size for simplicity), we will trim up to an α-fraction of data from both samples in order to minimize some distance, $d(\cdot,\cdot)$; and if we write $P_{n,\alpha}$, $Q_{n,\alpha}$ for the optimally trimmed empirical distributions, we will have $d(P_{n,\alpha}, Q_{n,\alpha}) \leq d(P_n, Q_n)$. Trimming procedures generally give a balanced compromise between efficiency and robustness, and increasing the level of trimming has a moderate effect on the efficiency. Thus, for univariate i.i.d. data coming from equal random generators, we typically have $d(P_n, Q_n) = \mathrm{O}_P(n^{-1/2})$ and $d(P_{n,\alpha}, Q_{n,\alpha}) = \mathrm{O}_P(n^{-1/2})$, but it is not true that $d(P_{n,\alpha}, Q_{n,\alpha}) = \mathrm{o}_P(n^{-1/2})$ (see, e.g., Theorem A.1 in [1]). However, for our procedure, overtrimming (i.e., trimming beyond the similarity level) will produce an *over-fitting effect*, namely, $d(P_{n,\alpha}, Q_{n,\alpha}) = \mathrm{o}_P(n^{-1/2})$. That will be the key for the statistical application of the procedure. Roughly speaking, if two random samples are trimmed more than required to delete contamination, then two samples far more similar than expected are obtained and it is feasible to distinguish this pair of trimmed samples from any other pair of non-trimmed, non-contaminated samples. We formalize this idea in Section 2. As in Álvarez-Esteban *et al.* [1], our choice for the metric d is the L_2 Wasserstein distance.

This over-fitting effect can be combined with a bootstrap procedure to consistently decide if the underlying distributions of two i.i.d. samples are similar in the sense of Definition 1 as we show in Section 3. This statistical procedure should also be useful in other frameworks of model validation. The consistency of our procedure is independent of the kind of contaminations. However, as expected, inliers are harder to detect than outliers. In this proposal, we have to consider small resampling sizes in the presence of inliers. This is discussed in Section 4, where we present some simulations showing the performance of our bootstrap procedure over finite samples. We also include the analysis of a real data set.

For the sake of readability we have moved most of the proofs to an Appendix, together with some additional results on rates of convergence.

Throughout the paper $\mathcal{P}$ will be the set of Borel probability measures on the real line, $\mathbb{R}$, while $\mathcal{F}_p$ will denote the set of distributions in $\mathcal{P}$ with finite pth absolute moment. If F is a distribution function, F^{-1} will denote its generalized inverse or quantile function. Given $P, Q \in \mathcal{P}$, by $P \ll Q$ we will denote absolute continuity of P with respect to Q, and by $\frac{\mathrm{d}P}{\mathrm{d}Q}$ the corresponding Radon–Nikodym derivative. Unless otherwise stated, the random variables will be assumed to be defined on the same probability space (Ω, σ, ν). Weak convergence of probabilities will be denoted by $\rightarrow_w$ and $\mathcal{L}(X)$ (resp., EX) will denote the law (resp., the mean) of the variable X. The indicator function of a set A will be I_A and ℓ will denote the Lebesgue measure.

2. Trimming and over-fitting

2.1. Trimmings of a distribution

Trimming an α-fraction of data in a sample of size n can be understood as replacing the empirical measure by a new one in which the data are reweighted so that the trimmed points now have zero probability while the remaining points will have weight $1/n(1-\alpha)$. By analogy we can define the trimming of a distribution as follows.

Definition 2. *Given* $\alpha \in (0,1)$, *we define the set of* α*-trimmed versions of* P *by*

$$\mathcal{R}_\alpha(P) := \left\{ Q \in \mathcal{P}\colon\ Q \ll P,\ \frac{\mathrm{d}Q}{\mathrm{d}P} \le \frac{1}{1-\alpha},\ P\text{-}a.s. \right\}. \tag{3}$$

This definition has been considered by several authors (see [1,7,13]). It allows the consideration of partial removal of the points in the support of the probability. This flexibility results in nice properties of the sets of trimmings, making $\mathcal{R}_\alpha(P)$ a convex set, compact for the topology of weak convergence (see Proposition 2.1 in [2]).

In this paper we use the quadratic Wasserstein distance, $\mathcal{W}_2$, namely, the minimal quadratic transportation cost between probabilities with finite second moment. $\mathcal{W}_2$ metrizes weak convergence plus convergence of second moments. We refer the reader to Section 8 of Bickel and Freedman [4] for further details on $\mathcal{W}_2$. On the real line $\mathcal{W}_2$ is simply the L_2 distance between quantile functions, that is, $\mathcal{W}_2^2(P_1,P_2) = \int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2\,\mathrm{d}t$ if F_i^{-1} is the quantile function of P_i. Trimmings are also well behaved with respect to $\mathcal{W}_2$, as shown in Álvarez-Esteban *et al.* [2]. For instance, for $P \in \mathcal{F}_2$, $\mathcal{R}_\alpha(P)$ is a compact subset of $\mathcal{F}_2$ for $\mathcal{W}_2$ (see Proposition 2.8 in [2]). A simple consequence is that in

$$\mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)) := \min_{R_i \in \mathcal{R}_\alpha(P_i)} \mathcal{W}_2(R_1, R_2) \tag{4}$$

the minimum is indeed attained. A remarkable result is that the minimizer is unique under mild assumptions. This is Theorem 2.16 in Álvarez-Esteban *et al.* [2], which generalizes related results in Caffarelli and McCann [6] and Figalli [11].

Proposition 1. *If* $P_1, P_2 \in \mathcal{F}_2$, $0 < \alpha < 1$ *and* P_1 *or* P_2 *has a density, then there exists a unique pair* $(P_{1,\alpha}, P_{2,\alpha}) \in \mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2)$ *such that*

$$\mathcal{W}_2(P_{1,\alpha}, P_{2,\alpha}) = \mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)),$$

provided $\mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)) > 0$.

The connection between trimmings and the similarity model of Definition 1 is given by the next result. Here d_{TV} denotes the distance in total variation, namely, $d_{\mathrm{TV}}(P_1,P_2) = \sup_B |P_1(B) - P_2(B)|$, where B ranges among all Borel sets.

Proposition 2. *For $\alpha \in [0,1)$ the following are equivalent*:

(a) *P_1 and P_2 are α-similar.*
(b) $\mathcal{R}_\alpha(P_1) \cap \mathcal{R}_\alpha(P_2) \neq \varnothing$.
(c) $d_{\mathrm{TV}}(P_1, P_2) \leq \alpha$.

If $P_1, P_2 \in \mathcal{F}_2$, then (a), (b) *or* (c) *is equivalent to*

(d) $\mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)) = 0$.

Finally, the common core distribution, P_0, in Definition 1 is unique if and only if $d_{\mathrm{TV}}(P_1, P_2) = \alpha$. In this case, P_0 is given by the density $f_0 = (f_1 \wedge f_2)/(1-\alpha)$ with respect to μ if μ is a common σ-finite dominating measure for P_1 and P_2 and f_1 and f_2 are the corresponding densities and we have the canonical decomposition $P_i = (1-\alpha)P_0 + \alpha P_i'$, $i = 1, 2$, P_i' having density $\frac{1}{\alpha}(f_i - f_1 \wedge f_2)$ with respect to μ.

Proof. If (a) holds, then $P_0(A) \leq \frac{1}{1-\alpha} P_i(A)$ for all Borel A. In particular, $P_0 \ll P_i$ and, if $A_i = \{\frac{\mathrm{d}P_0}{\mathrm{d}P_i} > (1-\alpha)^{-1}\}$, obviously $P_0(A_i) = 0$ and $P_0 \in \mathcal{R}_\alpha(P_1) \cap \mathcal{R}_\alpha(P_2)$, showing (b). Assume now (b) and take $P_0 \in \mathcal{R}_\alpha(P_1) \cap \mathcal{R}_\alpha(P_2)$. Then $(1-\alpha)P_0(A) \leq P_i(A)$ for all A. If $\alpha = 0$, then (c) holds trivially. Otherwise define $P_i'(A) = (P_i(A) - (1-\alpha)P_0(A))/\alpha$. Then P_i' is a probability and $d_{\mathrm{TV}}(P_1, P_2) = \alpha d_{\mathrm{TV}}(P_1', P_2') \leq \alpha$, that is, (c) holds. Finally, we assume that (c) holds and take μ to be a common σ-finite dominating measure for P_1 and P_2 and write f_1 and f_2 for the corresponding densities. Then (see Lemma 2.20 in [20]) $d_{\mathrm{TV}}(P_1, P_2) = 1 - \int (f_1 \wedge f_2)\,\mathrm{d}\mu$ (where $a \wedge b$ means $\min(a, b)$). Write $\varepsilon = d_{\mathrm{TV}}(P_1, P_2)$ and assume $\varepsilon > 0$ (the case $\varepsilon = 0$ is trivial). We set $f_i' = (f_i - f_1 \wedge f_2)/\varepsilon$, $i = 1, 2$, and $f_0 = (f_1 \wedge f_2)/(1-\varepsilon)$. f_0, f_1', f_2' are densities with respect to μ. We write P_0, P_1', P_2' for the associated probabilities. Then (1) holds with $\varepsilon_1 = \varepsilon_2 = \varepsilon \leq \alpha$. Equivalence of (b) and (d) follows from compactness of the sets of trimmings. The last claim follows easily from the arguments above. □

Remark 1. It follows from Proposition 2 that $\mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)) > 0$ if and only if $d_{\mathrm{TV}}(P_1, P_2) > \alpha$, that is, $d_{\mathrm{TV}}(P_1, P_2)$ is the minimal level of trimming required to make P_1 and P_2 equal. Also, if $d_{\mathrm{TV}}(P_1, P_2) = \alpha$, then the probability P_0 with density $f_0 = (f_1 \wedge f_2)/(1-\alpha)$ with respect to μ (as in the proof above) is the unique element in $\mathcal{R}_\alpha(P_1) \cap \mathcal{R}_\alpha(P_2)$. This means that, as in Proposition 1, there is also a unique pair, namely, $(P_0, P_0) \in \mathcal{R}_\alpha(P_1) \times \mathcal{R}_\alpha(P_2)$ such that

$$\mathcal{W}_2(P_0, P_0) = \mathcal{W}_2(\mathcal{R}_\alpha(P_1), \mathcal{R}_\alpha(P_2)) = 0.$$

This extends the result in Proposition 1 to the case $d_{\mathrm{TV}}(P_1, P_2) \geq \alpha$.

Proposition 2 shows that the similarity model (1) can be expressed in terms of different metrics. In fact, (d) would remain true if $\mathcal{W}_2$ were replaced by any other metric for which the sets of trimmings are compact. With applications in mind, $\mathcal{W}_2$ turns out to be a more convenient choice. In order to assess (1) from two samples of i.i.d. data with empirical distributions $P_{1,n}$ and $P_{2,m}$, say, we will have $d_{\mathrm{TV}}(P_{1,n}, P_{2,m}) = 1$ almost surely (provided P_1 and P_2 have densities) and we cannot use (at least in a naïve fashion) formulation (c). On the other hand, $\mathcal{W}_2$ is well behaved in this respect and empirical versions of both the minimal distances and the minimizers are consistent estimators of their theoretical counterparts. This is the content of the following result (Theorem 2.17 in [2]). We quote it here for completeness.

Theorem 1 (Consistency). *Let $\{X_n\}_n$, $\{Y_n\}_n$ be two sequences of i.i.d. random variables with $\mathcal{L}(X_n) = P$, $\mathcal{L}(Y_n) = Q$, $P, Q \in \mathcal{F}_2$, and write P_n, Q_m for the empirical distributions based on the samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$, respectively. Then, if* $\min(m, n) \to \infty$,

$$\mathcal{W}_2(\mathcal{R}_\alpha(P_n), \mathcal{R}_\alpha(Q_m)) \to \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) \qquad a.s.$$

Further, if P or $Q \ll \ell$ and $d_{\mathrm{TV}}(P, Q) \ge \alpha$, then

$$\mathcal{W}_2(P_{n,\alpha}, P_\alpha) \to 0 \quad \textit{and} \quad \mathcal{W}_2(Q_{m,\alpha}, Q_\alpha) \to 0 \qquad a.s.,$$

where $(P_\alpha, Q_\alpha) = \arg\min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2)$ and $(P_{n,\alpha}, Q_{m,\alpha})$ are defined similarly from P_n, Q_m.

2.2. Related concepts and works

The similarity model (1) is obviously related to the so-called ‘contamination neighborhoods’ of a probability P_0, defined as

$$\begin{aligned} \mathcal{V}_\varepsilon(P_0) &:= \{(1-\varepsilon)P_0 + \varepsilon P' \colon\ P' \in \mathcal{P}\} \\ &= \{Q \in \mathcal{P} \colon\ Q(A) \le (1-\varepsilon)P_0(A) + \varepsilon \text{ for every Borel set } A\}, \end{aligned} \tag{5}$$

which have been widely used in the theory of robust statistics after the pioneering works by Huber [15,16]. In particular, Huber [16] introduced these neighborhoods in robust testing, providing a robust version of the Neyman–Pearson lemma for simple hypothesis versus simple alternative. This theory was completed for more general sets of hypotheses and alternatives, additionally considering more flexible neighborhoods in Huber and Strassen [17], Rieder [22] and Buja [5]. In fact, Rieder’s neighborhoods of a probability P_0, defined as

$$\mathcal{V}^R_{\varepsilon,\delta}(P_0) := \{Q \in \mathcal{P} \colon\ Q(A) \le (1-\varepsilon)P_0(A) + \varepsilon + \delta \text{ for every Borel set } A\}, \tag{6}$$

comprise contamination as well as total variation norm neighborhoods (taking $\delta = 0$ or $\varepsilon = 0$, resp.).

It can be easily shown (see also Proposition 2.1 in [2]) that $P \in \mathcal{V}_\varepsilon(P_0)$ is equivalent to $P_0 \in \mathcal{R}_\varepsilon(P)$. Thus, our statement P_1 *and* P_2 *are α-similar* can also be expressed, in terms of contamination neighborhoods, as *there exists a probability* P_0 *such that* $P_1, P_2 \in \mathcal{V}_\alpha(P_0)$. However, there are different possibilities for such P_0, and the model considered in this paper, given through any one of the equivalent statements in Proposition 2, cannot be expressed in terms of a neighborhood, like (5) or (6) of a fixed probability.

Further related work includes Álvarez-Esteban *et al.* [1], where it is shown, for a probability, P, on the real line, that $\mathcal{R}_\alpha(P)$ can be expressed in terms of the trimmings of the uniform law on $(0, 1)$, $U(0, 1)$. This set can be identified with the set $\mathcal{C}_\alpha$ of absolutely continuous functions $h : [0, 1] \to [0, 1]$ such that $h(0) = 0$, $h(1) = 1$, with derivative h' such that $0 \le h' \le \frac{1}{1-\alpha}$. For function h, it is useful to write P_h for the probability measure with distribution function $h(P(-\infty, t])$. Then

$$\mathcal{R}_\alpha(P) = \{P_h \colon h \in \mathcal{C}_\alpha\}. \tag{7}$$

Hence, we can measure the deviation between the sets of trimmings of P and Q through

$$\mathcal{T}_\alpha(P, Q) := \min_{h \in \mathcal{C}_\alpha} \mathcal{W}_2(P_h, Q_h).$$

We call $\mathcal{T}_\alpha(P, Q)$ the *common trimming* distance between P and Q. If P and Q have quantile functions F^{-1} and G^{-1}, then a simple change of variable shows

$$\begin{aligned}\mathcal{W}_2(P_h, Q_h) &= \int_0^1 \bigl(F^{-1}(h^{-1}(x)) - G^{-1}(h^{-1}(x))\bigr)^2 \, \mathrm{d}x \\ &= \int_0^1 \bigl(F^{-1}(y) - G^{-1}(y)\bigr)^2 h'(y) \, \mathrm{d}y.\end{aligned}$$

Thus, $\mathcal{T}_\alpha(P, Q) = 0$ if and only if $\ell(\{y \in (0, 1) \colon F^{-1}(y) \ne G^{-1}(y)\}) \le \alpha$. It follows easily from this that $\mathcal{T}_\alpha(P, Q) = 0$ if and only if there is a random variable Z defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and non-decreasing, left-continuous functions, φ_1, φ_2, with $\mathcal{L}(\varphi_1(Z)) = P$, $\mathcal{L}(\varphi_2(Z)) = Q$ such that

$$\mathbb{P}\bigl(\varphi_1(Z) \ne \varphi_2(Z)\bigr) \le \alpha. \tag{8}$$

In contrast, since $d_{\mathrm{TV}}(P, Q) = \min\{\mathbb{P}(X \ne Y) \colon \mathcal{L}(X) = P, \mathcal{L}(Y) = Q\}$ (see Lemma 2.20 in [20]), we see that $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$ if and only if $\mathcal{L}(\varphi_1(Z)) = P$, $\mathcal{L}(\varphi_2(Z)) = Q$ for some random variable Z and measurable (not necessarily monotonic) φ_i such that (8) holds. In summary, *two random objects are α-similar if and only if they are different transforms of a common random signal and the transforms differ from each other with probability at most* α; they are equivalent in terms of common trimming if and only if they are different *monotonic* transforms of a common random signal and the transforms differ from each other with probability at most α. In the somewhat artificial event that we believe that our two samples come from a monotonic, possibly different, transform of some original signal, then the common trimming similarity model is reasonable. Otherwise, the similarity model (1) is the natural choice. For a less technical illustration of this idea we show in Figure 1 the different effect of independent and common trimming. We have taken $P = N(0, 1)$, $Q = 0.8N(0, 1) + 0.2N(4, 1)$ and three values of the trimming level, α. In the first row we show the densities of P_α (blue line) and Q_α (red line), with $(P_\alpha, Q_\alpha) = \arg\min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2)$. In this case, trimming $\alpha = 0.2$ results in $P_\alpha = Q_\alpha$, that is, trimming removes contamination. The second row shows the densities of P_{h_α} (blue line) and Q_{h_α} (red line), with $h_\alpha = \arg\min_{h \in \mathcal{C}_\alpha} \mathcal{W}_2(P_h, Q_h)$. Clearly, P_{h_α} and Q_{h_α} are different and this remains true no matter how close to 1 we choose α. If trimming is used with the goal of removing contamination and assessing that the core of the two distributions are equal, then it is clear that the common trimming approach fails to do so.

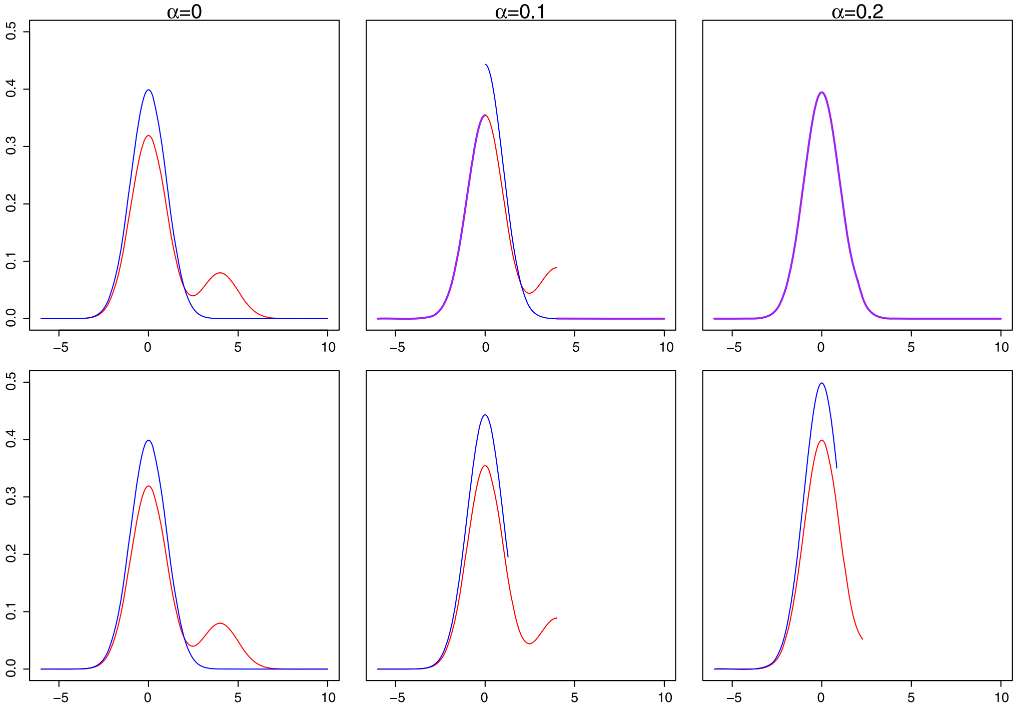

Figure 1. Densities of optimally trimmed P and Q with independent trimming (first row) and common trimming (second row).

In Álvarez-Esteban *et al.* [3] we have considered, under this common trimming setup, the problem of testing whether a random sample can be considered 'mostly normal', that is, if the generator of the sample is similar to a normal distribution with unknown parameters.

Finally let us mention the application in Álvarez-Esteban *et al.* [2] of some asymptotic results for a related two-sample problem: Given $X_1, \ldots, X_n$ i.i.d. P and $Y_1, \ldots, Y_m$ i.i.d. Q, we consider testing the related null hypotheses

$$H_1\colon\ \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) \le \Delta_0 \quad \text{vs.} \quad \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) > \Delta_0,$$

$$H_2\colon\ \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) \ge \Delta_0 \quad \text{vs.} \quad \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) < \Delta_0$$

for a given threshold $\Delta_0 > 0$ to be chosen by the practitioner. Observe that rejecting the null hypothesis H_2 allows us to conclude that, with high confidence, the unknown random generators P and Q are not far from similarity.

2.3. The over-fitting effect of trimming

In this subsection we keep the notation of Theorem 1 and assume that we deal with two independent samples, $X_1, \ldots, X_n$ i.i.d. P and $Y_1, \ldots, Y_m$ i.i.d. Q. We write P_n, Q_m for the empirical measures and $P_{n,\alpha}$, $Q_{m,\alpha}$ are minimizers of the $\mathcal{W}_2$ distance between trimmings of the empirical distributions P_n, Q_m.

It follows from Theorem 1 that $\mathcal{W}_2(P_{n,\alpha}, Q_{m,\alpha}) \to 0$ a.s. when the similarity model (1) holds true and we may wonder about the rate of convergence in this limit. Note that under homogeneity, that is, if $P = Q$ and taking $n = m$ for simplicity, we have under integrability assumptions

$$\sqrt{n}\mathcal{W}_2(P_n, Q_n) \to_w \left(2\int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))}\,\mathrm{d}t\right)^{1/2}, \tag{9}$$

where B is a Brownian bridge and f and F^{-1} are the density and quantile functions of P (this follows easily, for instance, from Theorem 4.6 in [10]). Thus, random samples from homogeneous generators have empirical distributions at $\mathcal{W}_2$-distance of exact order $n^{-1/2}$, while, for non-homogeneous random generators $\mathcal{W}_2(P_n, Q_n) \to \mathcal{W}_2(P, Q)$, a positive constant. Likewise, in the common trimming model of Section 2.2, if $h_{n,\alpha}$ is such that $\mathcal{T}_\alpha(P_n, Q_n) = \mathcal{W}_2((P_n)_{h_{n,\alpha}}, (Q_n)_{h_{n,\alpha}})$ and we write $\tilde{P}_{n,\alpha} = (P_n)_{h_{n,\alpha}}$, $\tilde{Q}_{n,\alpha} = (Q_n)_{h_{n,\alpha}}$ (the optimal trimmings of the empirical measures), then, under $\mathcal{T}_\alpha(P, Q) = 0$, we have that $\sqrt{n}\mathcal{W}_2(\tilde{P}_{n,\alpha}, \tilde{Q}_{n,\alpha})$ converges in law to a non-null limit (Theorem A.1 in [1]), whereas if $\mathcal{T}_\alpha(P, Q) > 0$, then $\mathcal{W}_2(\tilde{P}_{n,\alpha}, \tilde{Q}_{n,\alpha})$ converges a.s. to a positive constant.

In the similarity model (1) the gap between the null and the alternative is of higher order. If P and Q are not similar at level α, then $\mathcal{W}_2(P_{n,\alpha}, Q_{m,\alpha}) \to \mathcal{W}_2(P_\alpha, Q_\alpha) > 0$ (Theorem 1). On the other hand, if $d_{\mathrm{TV}}(P, Q) < \alpha$, then our next result shows that $\sqrt{n}\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha}) \to 0$ in probability. To avoid integrability issues, we assume P and Q to have bounded support; this is enough for applications, since a monotonic transformation of the data could achieve boundedness while preserving the distance in total variation. Furthermore, it ensures that the conditions $d_{\mathrm{TV}}(P, Q) \le \alpha$ and $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$ are equivalent.

Theorem 2. *Assume $P, Q \in \mathcal{F}_2$ are supported in a common bounded interval and have densities bounded away from zero and with bounded derivatives. Assume further that $n/(n+m) \to \lambda \in (0,1)$. If $\alpha_n \in (0,1)$ satisfies $\alpha_n \ge d_{\mathrm{TV}}(P, Q) + \frac{r_n}{\sqrt{n}}$ for some $r_n \to \infty$, then*

$$\sqrt{n}\mathcal{W}_2(P_{n,\alpha_n}, Q_{m,\alpha_n}) \to 0 \qquad \textit{in probability.} \tag{10}$$

We give a proof of Theorem 2 in the Appendix. A similar over-fitting effect is observed if a sample is over-trimmed to optimally fit a given model: If $X_1, \ldots, X_n$ are i.i.d. P, $P_{n,\alpha} = \arg\min_{R\in\mathcal{R}_\alpha(P_n)} \mathcal{W}_2(R, Q)$ and $\mathcal{W}_2(\mathcal{R}_{\alpha_0}(P), Q) = 0$ for some $\alpha_0 < \alpha$, then (see Theorem 5 in the Appendix)

$$\sqrt{n}\mathcal{W}_2(P_{n,\alpha}, Q) \to 0 \qquad \text{in probability.}$$

Empirical evidence of this over-fitting effect is shown in Figure 2. A random sample of size $n = 1000$ from a $U(0,1)$ distribution was taken. This sample was trimmed using the proportions $\alpha = 0, 0.1, 0.3$ in order to obtain a sample as close to the $U(0,1)$ as possible. We denote by F_n^α the distribution function of $P_{n,\alpha}$ and in Figure 2, we represent the empirical processes $D_n^\alpha(t) = n^{1/2}(F_n^\alpha(t) - t)$, $t \in [0,1]$ for $\alpha = 0, 0.1, 0.3$.

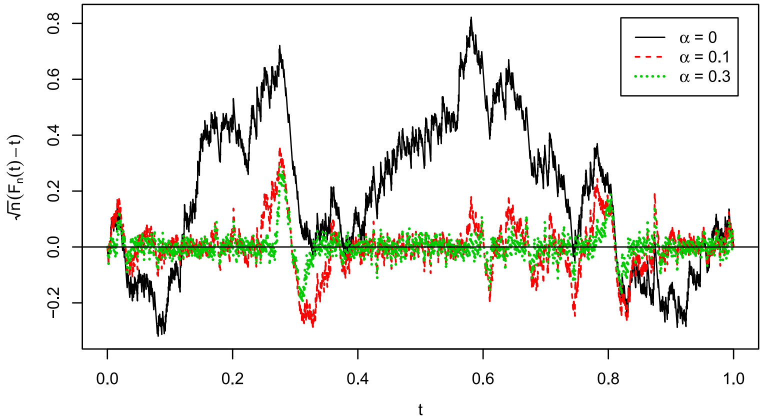

Figure 2. Trajectories of the uniform empirical process (solid line) and two variants based on trimming. The trimming levels are $\alpha = 0.1$ and $\alpha = 0.3$ (dashed and dotted lines).

Since the true random generator and the target are the same, no trimming is required in this case to remove contamination and, for $\alpha > 0$, we are over-trimming. Observe that $D_n^{0.1}$ and $D_n^{0.3}$ do not differ too much from each other, while they are quite far from the untrimmed version.

3. A bootstrap assessment of similarity

We show in this section how we can use the over-fitting effect of trimming for the assessment of the similarity model (1). Again, we will assume that we observe two independent random samples $X_1, \dots, X_n$ i.i.d. P, $Y_1, \dots, Y_m$ i.i.d. Q. We would like to test the null hypothesis H_0: $d_{\mathrm{TV}}(P, Q) \le \alpha$. Theorem 2 says that trimming beyond the similarity level kills randomness and results in (trimmed) samples that are more similar to each other than random samples coming from the same generator. We will use a bootstrap approach to generate suitable random samples from a common generator and compare the optimally trimmed distance to the distance computed on the bootstrap replicates.

We write P_n, Q_m for the empirical distributions and, given $\alpha_n \in (0, 1)$,

$$(P_{n,\alpha_n}, Q_{m,\alpha_n}) = \operatorname*{arg\,min}_{R_1 \in \mathcal{R}_{\alpha_n}(P_n), R_2 \in \mathcal{R}_{\alpha_n}(Q_m)} \mathcal{W}_2(R_1, R_2),$$

so that $\mathcal{W}_2(P_{n,\alpha_n}, Q_{m,\alpha_n}) = \mathcal{W}_2(\mathcal{R}_{\alpha_n}(P_n), \mathcal{R}_{\alpha_n}(Q_m))$.

We consider now the pooled probability

$$R_{n,m} = \frac{n}{n+m} P_{n,\alpha_n} + \frac{m}{n+m} Q_{m,\alpha_n}.$$

$R_{n,m}$ is a random probability measure concentrated on $\{Z_1, \dots, Z_{n+m}\}$, where $Z_j = X_j$ for $j = 1, \dots, n$, and $Z_j = Y_{j-n}$ for $j = n+1, \dots, n+m$.

Conditionally, given the data, we draw new random variables, $X_1^*, \dots, X_{n'}^*, Y_1^*, \dots, Y_{m'}^*$ i.i.d. $R_{n,m}$, with $m' = [n'm/n]$ and n' to be chosen later. We will use the notation $\mathbb{P}^*$ for the bootstrap probability, that is, the conditional probability given the original data $\{X_n\}_n$, $\{Y_m\}_m$. Finally, by $P_{n'}^*$ and $Q_{m'}^*$ we will denote the empirical measures based on $X_1^*, \dots, X_{n'}^*$ and $Y_1^*, \dots, Y_{m'}^*$, respectively. Now, we define

$$p_{n,m}^* := \mathbb{P}^* \left\{ \sqrt{\frac{n'm'}{n'+m'}} \mathcal{W}_2(P_{n'}^*, Q_{m'}^*) > \sqrt{\frac{nm}{n+m}} \mathcal{W}_2(P_{n,\alpha_n}, Q_{m,\alpha_n}) \right\}. \tag{11}$$

$p_{n,m}^*$ is the bootstrap p-value for the similarity model (1), with rejection for small values of it. In practice $p_{n,m}^*$ can be approximated by Monte Carlo simulation. We note that if $n\alpha_n$ and $m\alpha_n$ are integer, typically the trimming process will not produce partially trimmed points and P_{n,α_n} and Q_{m,α_n} will be the empirical measures on the sets of non-trimmed data. If we take $\alpha_n \to \alpha$, then if the similarity model fails, $\mathcal{W}_2(P_{n,\alpha_n}, Q_{m,\alpha_n})$ will be large while $\mathcal{W}_2(P_{n'}^*, Q_{m'}^*)$ will vanish. On the other hand, for similar distributions $\mathcal{W}_2(P_{n,\alpha_n}, Q_{m,\alpha_n})$ will vanish at a faster rate than $\mathcal{W}_2(P_{n'}^*, Q_{m'}^*)$ and rejection for small bootstrap p-values will result in a consistent rule. We make this precise in our next result.

Theorem 3. *With the above notation, assume that P, Q have densities satisfying the assumptions of Theorem 2. Assume further that $n/(n+m) \to \lambda \in (0,1)$ and take $\alpha_n = \alpha + K/\sqrt{n \wedge m}$ with $K > 0$. Then, if $n' \to \infty$ and $n' = \mathrm{O}(n)$,*

(i) *if $d_{\mathrm{TV}}(P, Q) < \alpha$, then $p_{n,m}^* \to 1$ in probability,*
(ii) *if $d_{\mathrm{TV}}(P, Q) > \alpha$, then $p_{n,m}^* \to 0$ in probability.*

A proof of Theorem 3 is given in the Appendix. It roughly says that a test of the similarity model (1) that rejects α-similarity for values of $p_{n,m}^*$ above a fixed threshold $L \in (0,1)$ is a consistent rule. In order to make a sensible choice of the threshold, L, as well as of the constant, K, in Theorem 3, we still need to control the probability of rejection at the boundary of the null hypothesis; that is, in the case $d_{\mathrm{TV}}(P, Q) = \alpha$. In this case we write again P_0 for the common part of P and Q in the canonical decomposition in Remark 1. If $\tilde{P}_n \in \mathcal{R}_{\alpha_n}(P)$ and $\tilde{Q}_n \in \mathcal{R}_{\alpha_n}(Q)$, with α_n as in Theorem 3, are such that $\mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) \to 0$, then, by uniqueness, we have $\mathcal{W}_2(\tilde{P}_n, P_0) \to 0$. We introduce the following assumption about rates in this convergence: If $\tilde{P}_n \in \mathcal{R}_{\alpha_n}(P)$, $\tilde{Q}_n \in \mathcal{R}_{\alpha_n}(Q)$ (and $\alpha_n = d_{\mathrm{TV}}(P, Q) + \frac{K}{\sqrt{n}}$), then, for some $\rho \in (0, 1]$,

$$\mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) = \mathrm{O}(n^{-1/2}) \quad \Rightarrow \quad \mathcal{W}_2(\tilde{P}_n, P_0) = \mathrm{O}(n^{-\rho/2}). \tag{12}$$

Under this assumption we can control the type I error probability using our next result.

Theorem 4. *Under the assumptions and notation of Theorem 3, if P and Q are such that $d_{\mathrm{TV}}(P, Q) = \alpha$ and satisfy (12), taking $n' \to \infty$, $n' = \mathrm{o}(n^\rho)$ and*

$$\alpha_n = \alpha + \frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{n \wedge m}} \Phi^{-1}\left(\sqrt{1-\gamma}\right)$$

with $\gamma \in (0,1)$, then $\limsup_n \mathbb{P}(p_{n,m}^* \le \beta) \le \beta + \gamma$.

The main consequence is that we can test the similarity model (1) at a given level $\beta+\gamma\in(0,1)$. To be precise, if we replace our ideal H_0: $d_{\mathrm{TV}}(P,Q)\le\alpha$ by $\tilde{H}_0$ consisting of pairs (P,Q) satisfying the assumptions in Theorem 3 and $d_{\mathrm{TV}}(P,Q)<\alpha$ or $d_{\mathrm{TV}}(P,Q)=\alpha$ plus Condition (12), then, if we reject for $p^*_{n,m}\le\beta$, Theorems 3 and 4 ensure

$$\sup_{(P,Q)\in\tilde{H}_0}\limsup_n \mathbb{P}_{(P,Q)}(p^*_{n,m}\le\beta)\le\beta+\gamma,$$

where $\mathbb{P}_{(P,Q)}$ denotes probability assuming the laws of the X's and the Y's are P and Q, respectively. It is in this sense that we can say that the procedure is conservative, having an asymptotic level of, at most, $\beta+\gamma$; nevertheless, the test will consistently reject the similarity model if it fails. In the next section we show the performance in practice of this procedure. Of course, one would like to control

$$\limsup_n \sup_{(P,Q)\in H_0}\mathbb{P}_{(P,Q)}(p^*_{n,m}\le\beta)$$

instead of the bound given by our results. Some of the limitations of our procedure come from the smoothness requirements posed by our choice of metric, $\mathcal{W}_2$. This could, perhaps, be overcome with the use of the L_1 Wasserstein metric (but we would lose the uniqueness and consistency results given in Proposition 1 and Theorem 1) and consideration of a less restrictive null hypothesis, $\tilde{H}_0$. Uniformity in $(P,Q)\in H_0$ is a more delicate issue, since one can take P and Q at an arbitrary (but positive) Wasserstein distance from each other, but such that they are at distance one in total variation. Perhaps a different choice of metric could lead to some type of uniform bound. We believe this issue is worth further research.

Turning to the meaning of Condition (12), observe that the contaminations, P_1', P_2', in the canonical decomposition in Proposition 2 have disjoint support but can be arbitrarily close in Wasserstein distance. With Condition (12) we avoid pathological cases in which some inconvenient distribution of the contaminations allows that some trimmings of P and Q, with trimming size slightly above the similarity level, are close to each other without being too close to the common core. Rather than pursuing an involved technical analysis we include a couple of illustrative examples that show that the best possible rate ρ depends on the *degree of separation* between the contaminating distributions P_1', P_2' in the canonical decomposition. In the well-separated case (when the distance between the supports of P_1' and P_2' is positive), under additional technical conditions we can take $\rho=1$ and we have that the optimal trimming, P_{n,α_n}, approaches the common part, P_0, at the parametric rate: $\mathcal{W}_2(P_{n,\alpha_n},P_0)=\mathrm{O}_P(n^{-1/2})$. Without this separation we cannot take ρ greater than 4/5 and we have a nonparametric rate of convergence: $\mathcal{W}_2(P_{n,\alpha_n},P_0)=\mathrm{O}_P(n^{-2/5})$. Again, in our examples we assume P and Q to have bounded support since this is enough for applications.

Example 1 (The well-separated case). *Assume P and Q are probabilities on the real line with quantile functions, F^{-1} and G^{-1}, such that $G^{-1}(t)=F^{-1}(t+\alpha)$, $0<t<1-\alpha$ and F^{-1} has a bounded derivative* (*as in Figure* 3(a)). *Then $d_{\mathrm{TV}}(P,Q)=\alpha$ and, taking $\alpha_n=\alpha+\frac{K}{\sqrt{n}}$ for some $K>0$ and writing P_0 for the common part in the canonical decomposition for P and Q, we have that if $\tilde{P}_n\in\mathcal{R}_{\alpha_n}(P)$, $\tilde{Q}_n\in\mathcal{R}_{\alpha_n}(Q)$, then*

$$\mathcal{W}_2(\tilde{P}_n,\tilde{Q}_n)=\mathrm{O}(n^{-1/2})\quad\Rightarrow\quad\mathcal{W}_2(\tilde{P}_n,P_0)=\mathrm{O}(n^{-1/2}).$$

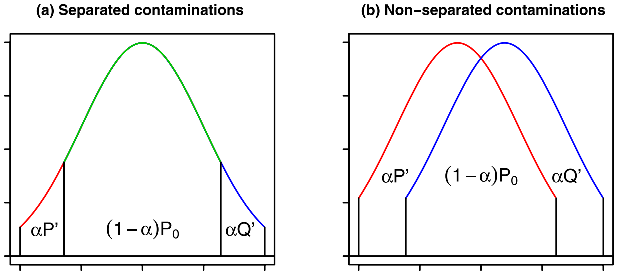
(a) Separated contaminations
αP'
(1 − α)P0
αQ'
(b) Non−separated contaminations
αP'
(1 − α)P0
αQ'

Figure 3. Canonical decomposition in the separated (left) and non-separated (right) cases.

Example 2 (The non-separated case). *We assume now that P and Q differ only in location and have a symmetric, unimodal density. Without loss of generality, we write $F(\cdot + \mu/2)$ and $F(\cdot - \mu/2)$ for the distribution functions of P and Q, respectively, and f for the density associated to F. We suppose that F has bounded support and f is strictly positive on it. Further, we assume f to be continuously differentiable with $f' < 0$ in $(0, \sup(\mathrm{supp}(F)))$. If μ and α satisfy $1 - \alpha = 2(1 - F(\mu/2)) = 2F(-\mu/2)$, then $d_{\mathrm{TV}}(P, Q) = \alpha$ (see Figure 3(b)). If $\tilde{P}_n \in \mathcal{R}_{\alpha_n}(P)$, $\tilde{Q}_n \in \mathcal{R}_{\alpha_n}(Q)$, then*

$$\mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) = \mathrm{O}(n^{-1/2}) \quad \Rightarrow \quad \mathcal{W}_2(\tilde{P}_n, P_0) = \mathrm{O}(n^{-2/5}).$$

A proof of the claims in the last two examples is sketched in the Appendix.

While this work is concerned mainly with testing α-similarity in two-sample problems, in many real problems the interest could be focused on the estimation on the common core P_0. The results in Section 2 ensure that the pooled probability, $P_{n,m}$, in our bootstrap procedure is a consistent estimator of P_0 if α equals the (unknown) distance in total variation between P and Q. Our simulations in Section 4 (see Figure 5 and the related comments) suggest that the bootstrap p-value curves (the values of $p^*_{n,m}$ as a function of α) change sharply from 0 to 1 around the true similarity level. Maybe this rapid growth could be used to give some estimation of the similarity level and, as a result, of the common core. Further research is needed.

We conclude this section by presenting a simple upper bound for the transportation cost between empirical measures. This result, together with Theorem 2, is the key in our proofs of Theorems 3 and 4 and has some independent interest. The proof is also included in the Appendix. Here $X_{1,1}, \ldots, X_{1,n}$; $X_{2,1}, \ldots, X_{2,m}$ are i.i.d. $\mathbb{R}^k$-valued random vectors with common distribution P and $Y_{1,1}, \ldots, Y_{1,n}$; $Y_{2,1}, \ldots, Y_{2,m}$ are i.i.d. Q. We write $P_{n,1}$ and $P_{m,2}$ for the empirical measures based on $X_{1,1}, \ldots, X_{1,n}$ and $X_{2,1}, \ldots, X_{2,m}$, respectively, and, similarly, $Q_{n,1}$ and $Q_{m,2}$ for the empirical measures based on the $Y_{i,j}$. Let us define

$$S_{n,m} := \mathcal{W}_p(P_{n,1}, P_{m,2}) \quad \text{and} \quad T_{n,m} := \mathcal{W}_p(Q_{n,1}, Q_{m,2}).$$

Proposition 3. *With the above notation, if $p \geq 1$, then*

$$\mathcal{W}_p(\mathcal{L}(S_{n,m}), \mathcal{L}(T_{n,m})) \leq 2\mathcal{W}_p(P, Q).$$

4. Empirical analysis of the procedure

In this section we explore the performance of the procedure for finite samples. The section is divided in two subsections that address the analysis of a planned simulation study and of a case study, respectively. To simplify our exposition we will assume equal sizes in the two samples through the first subsection. All the computations have been carried out with the programs available at http://www.eio.uva.es/~pedroc/R.

4.1. A simulation study

We consider first an example that illustrates the over-fitting effect on the bootstrap p-values. We generate 200 pairs of samples of size $n = 1000$ obtained from the $N(0, 1)$ and the $0.9N(0, 1) + 0.1N(10, 3)$ distributions. Then, for each pair of samples, we carry out the bootstrap procedure (1000 bootstrap replicates in each run) for trimming levels $\alpha = 0.09$ and 0.11. At this point an important caution when dealing with mixtures should be made, namely the distinction between the level (0.1 in our case) of the "contaminating" distribution in the mixture and the similarity level between the non-contaminated and contaminated distributions. Of course, both distributions are similar at level 0.1, but they are also similar at a lower level (recall the canonical decomposition in Remark 1). For example, since the supports of the $U(0, 1)$ and $U(1, 2)$ distributions are disjoint, then the minimum level of similarity between the $U(0, 1)$ and $0.9U(0, 1) + 0.1U(1, 2)$ distributions is 0.1; but between the $N(0, 1)$ and $0.9N(0, 1) + 0.1N(\mu, 3)$ distributions, it is strictly lower for every μ. For instance, this level is 0.0484 if $\mu = 0$, 0.0653 for $\mu = 3$; or 0.0989 when $\mu = 10$.

Figure 4 shows the absolute frequencies of the bootstrap p-values, $p^*_{n,n}$, obtained in this example.

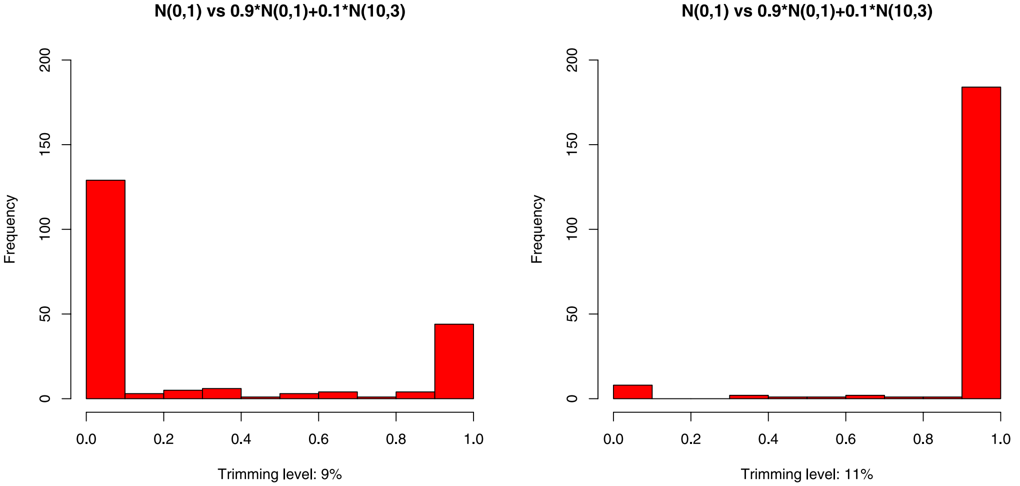

Figure 4. Histograms, for different sizes of trimming, of the bootstrap p-values obtained from 200 pairs of samples from $P = N(0, 1)$ and $Q = 0.9N(0, 1) + 0.1N(10, 3)$ distributions.

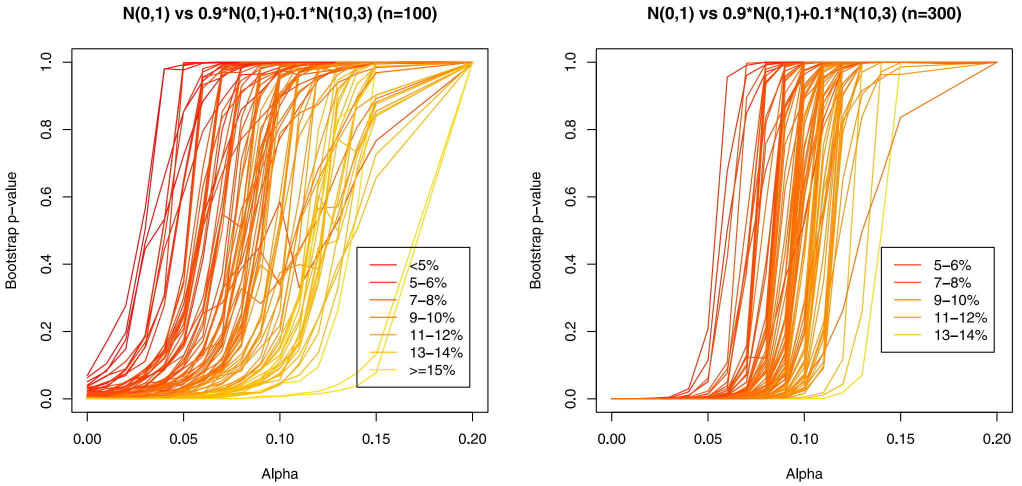

Figure 5. Curves of bootstrap p-values obtained by varying the trimming level (α). Colors depend on the real proportion of data coming from the $N(10, 3)$ distribution in each particular sample.

As stated above, the similarity level between the considered distributions is 0.0989. Thus, the probability of obtaining an observation from the non-common part in the mixture is 0.0989. Taking into account sample sizes and the number of samples considered, the expected number of times in which we obtain at most 110 ‘contaminating’ observations in both samples is 158.13. In these cases, after 0.11 trimming, we will be comparing similar samples and should have no evidence against similarity. We note that 158 is slightly below the observed frequency in the right bar of the right histogram in Figure 4. On the other hand, the expected number of times in which the amount of ‘contaminating’ data exceeds 90 in both samples is 132.02. In this event, 0.09 trimming is unable to remove contamination and we should have strong evidence against similarity. We can check that 132 is close to the observed frequency in the left bar of the left histogram in Figure 4.

The comments above suggest that the p-values are very sensitive to the effective proportion of contamination in the data. This is further illustrated with the plots in Figure 5, which show the curves of bootstrap p-values conditioned to different ranges of contaminating proportion in the second sample (the amount of data coming from the $N(10, 3)$ distribution). In this figure we observe that the transition from p-values close to 0 to p-values close to 1 is very fast along the trimming level. In other words, the effect of under-/over-trimming becomes apparent very quickly.

We show next a simulation study to illustrate the power performance for finite samples of the bootstrap procedure introduced in Section 3, when the trimming level, α_n, is determined as in Theorem 4. We consider two different cases, comparing samples of the same size, n, of $P = N(0, 1)$ versus $Q_i, i = 1, 2$. In the first case, $Q_1 = (1 - \varepsilon)N(0, 1) + \varepsilon N(10, 1)$; the contamination is due to outliers. In the second case, the contamination is due to inliers and $Q_2 = (1 - \varepsilon)N(0, 1) + \varepsilon N(0, 3)$. In both cases, the null hypothesis is H_0: $d_{\mathrm{TV}}(P, Q_i) \leq 0.1$ and we use 1000 bootstrap pairs of samples to obtain $p^*_{n,n}$, rejecting H_0 if $p^*_{n,n} \leq 0.05 = \beta$. Then we

Table 1. Observed rejection frequencies for H_0: $d_{\mathrm{TV}}(P, Q_1) \le 0.1$, $P = N(0,1)$, $Q_1 = (1-\varepsilon)N(0,1) + \varepsilon N(10,1)$, where $\nu = d_{\mathrm{TV}}(P, Q_1)$ and $\beta = 0.05$

		ρ:	1		4/5		2/3		1/2	
ν	n	γ:	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
0.10	100		0.008	0.001	0.016	0.003	0.043	0.006	0.047	0.007
$\varepsilon \simeq 0.10$	300		0.030	0.007	0.040	0.015	0.059	0.017	0.065	0.019
	1000		0.052	0.009	0.092	0.016	0.098	0.018	0.114	0.022
0.15	100		0.130	0.044	0.207	0.090	0.246	0.130	0.252	0.170
$\varepsilon \simeq 0.15$	300		0.587	0.386	0.648	0.458	0.687	0.507	0.703	0.556
	1000		0.996	0.980	0.998	0.985	0.998	0.986	0.999	0.990
0.20	100		0.576	0.403	0.685	0.515	0.732	0.585	0.738	0.624
$\varepsilon \simeq 0.20$	300		0.990	0.973	0.992	0.981	0.993	0.985	0.993	0.986
	1000		1	1	1	1	1	1	1	1
0.25	100		0.919	0.842	0.953	0.893	0.969	0.917	0.970	0.929
$\varepsilon \simeq 0.25$	300		1	1	1	1	1	1	1	1
	1000		1	1	1	1	1	1	1	1

compute the rejection frequencies in 1000 iterations of the procedure, obtaining the values shown in Tables 1 and 2. We do this for different values of ε (then different values of $\nu = d_{\mathrm{TV}}(P, Q_i)$) and different resampling orders $n' = n^{\rho}$. The simulation shows that the bound given in Theorem 4 is approached for moderate sizes in the first case (see Table 1, $\nu = 0.10$). However, in the second case, the procedure is conservative. The main conclusion is that in both cases the contamination is detected, but detection is more difficult in the case in which the contamination comes from inliers.

We close this subsection with a comparison to classical testing procedures that could be adapted to the setup of similarity testing. We recall from Proposition 2 that testing α-similarity of P and Q is equivalent to testing whether $\sup_A |P(A) - Q(A)| \le \alpha$, with A ranging among all (measurable) sets. If we focus on sets of type $A = (-\infty, x]$, then we could test the null hypothesis H_0: $\sup_{x\in\mathbb{R}} |F(x) - G(x)| \le \alpha$ using the Kolmogorov–Smirnov statistic: $D_n = \sup_{x\in\mathbb{R}} |F_n(x) - G_n(x)|$, where F_n and G_n denote the empirical distribution functions (d.f.'s) based on the X_i and the Y_j, respectively (and we have assumed for simplicity samples of equal size). It is known (see [21]) that, provided $\sup_{x\in\mathbb{R}} |F(x) - G(x)| = \lambda > 0$, $\sqrt{n}(D_n - \lambda)$ converges weakly to $Z_\lambda(F, G) = \max(Z_1, Z_2)$ with

$$Z_1 = \sup_{\{x: F(x)-G(x)=\lambda\}} B_1\bigl(G(x) + \lambda\bigr) - B_2(G(x)),$$

$$Z_2 = \sup_{\{x: G(x)-F(x)=\lambda\}} B_2(G(x)) - B_1\bigl(G(x) - \lambda\bigr),$$

where B_1, B_2 are independent Brownian bridges on $(0, 1)$. With standard arguments it can be shown that $P(Z_\lambda(F, G) > t) \le P(Z_\lambda > t)$ for $t > 0$, with $Z_\lambda = \sup_{0\le x\le 1-\lambda} B_1(x + \lambda) - B_2(x)$.

Table 2. Observed rejection frequencies for H_0: $d_{\mathrm{TV}}(P, Q_2) \le 0.1$, $P = N(0, 1)$, $Q_2 = (1-\varepsilon)N(0, 1) + \varepsilon N(0, 3)$, where $\nu = d_{\mathrm{TV}}(P, Q_2)$ and $\beta = 0.05$

		ρ:	1		4/5		2/3		1/2	
ν	n	γ:	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
0.10	100		0	0	0	0	0	0	0	0
$\varepsilon \simeq 0.21$	300		0	0	0	0	0	0	0	0
	1000		0	0	0	0	0	0	0	0
0.15	100		0.002	0.000	0.002	0.001	0.002	0.001	0.003	0.001
$\varepsilon \simeq 0.31$	300		0.013	0.003	0.016	0.005	0.017	0.006	0.027	0.008
	1000		0.185	0.089	0.196	0.100	0.210	0.103	0.235	0.120
0.20	100		0.037	0.017	0.048	0.022	0.060	0.023	0.065	0.027
$\varepsilon \simeq 0.41$	300		0.397	0.253	0.418	0.279	0.437	0.293	0.490	0.330
	1000		0.992	0.979	0.994	0.979	0.995	0.982	0.994	0.983
0.25	100		0.254	0.146	0.277	0.163	0.301	0.189	0.324	0.195
$\varepsilon \simeq 0.52$	300		0.924	0.846	0.928	0.856	0.936	0.866	0.949	0.888
	1000		1	1	1	1	1	1	1	1
0.30	100		0.565	0.426	0.599	0.456	0.629	0.484	0.654	0.508
$\varepsilon \simeq 0.62$	300		0.996	0.993	0.998	0.993	0.998	0.993	0.999	0.995
	1000		1	1	1	1	1	1	1	1

Hence, if we choose $z_\alpha^{(\beta)}$ such that $P(Z_\alpha > z_\alpha^{(\beta)}) = \beta$, then the test that rejects when

$$D_n > \alpha + \frac{1}{\sqrt{n}} z_\alpha^{(\beta)}$$

is asymptotically of level β for testing H_0: $\sup_{x\in\mathbb{R}} |F(x) - G(x)| \le \alpha$. The critical value $z_\alpha^{(\beta)}$ can be approximated by Monte Carlo simulation. We could try to use this procedure for testing the α-similarity model. Though, since we can find distributions that are arbitrarily close in Kolmogorov–Smirnov distance but far from each other in total variation distance, this alternative procedure can fail badly. We show this in our last simulation study (see Table 3). We have taken $P = N(0, 1)$ and $Q = 0.70N(0, 1) + 0.15N(2.35, 1) + 0.15N(-2.35, 1)$, a mixture with three normal components. Here we have $\sup_{x\in\mathbb{R}} |P(-\infty, x] - Q(-\infty, x]| = 0.1$ and $d_{\mathrm{TV}}(P, Q) = 0.2$ and we test H_0: $d_{\mathrm{TV}}(P, Q) \le 0.1$ at level 0.05. We show the observed frequencies of rejection

Table 3. Observed rejection frequencies for H_0: $d_{\mathrm{TV}}(P, Q) \le 0.1$, $P = N(0, 1)$, $Q = 0.70N(0, 1) + 0.15N(2.35, 1) + 0.15N(-2.35, 1)$ at level 0.05

n	100	300	500	1000
D_n	0.007	0.004	0.003	0.002
$\mathcal{W}_2$	0.007	0.091	0.320	0.875

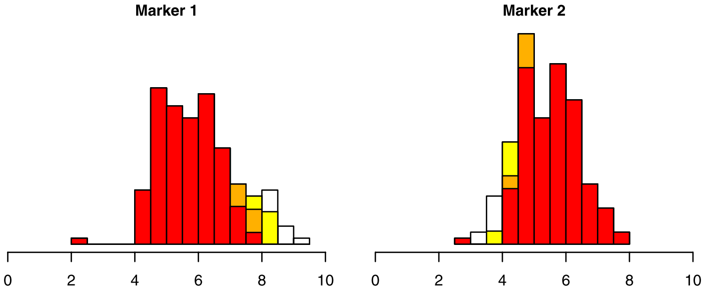

Figure 6. Best trimmings between markers 1 and 2, in the example of Section 4.2, $\alpha = 0.05$ (white), $\alpha = 0.10$ (white + yellow) and $\alpha = 0.15$ (white + yellow + orange).

for D_n and our bootstrap procedure based on $\mathcal{W}_2$ as in Theorem 4 with $\rho = 4/5$, $\gamma = 0.01$. In this case we reject for bootstrap p-values larger than 0.04 to make the asymptotic probability of type I error less than 0.05. We have considered sampling sizes $n = 100, 300, 500$ and 1000 and have produced 10,000 replicates of the tests in each case. We see that the Kolmogorov–Smirnov test fails to detect the dissimilarity, even for large sample sizes, while the bootstrap procedure suggested in this paper works reasonably for moderate sizes.

4.2. A case study

The data from this case study come from an admission exam to the Universidad de Valladolid. 308 exams on the same subject were randomly assigned to 2 markers. The distribution of the exams was not exactly balanced and markers received 152 and 156 exams, respectively. Each exam was given a grade between 0 and 10 points. In the admission exams some marking criteria are given to the markers with the goal of making the grading process "homogeneous". The main goal of this study is to determine whether the markers are using the same common criteria. Some degree of deviation from this common pattern is allowed for each marker. Therefore, we would like to assess the similarity of the samples of marks for the different markers.

The use of nonparametric methods strongly rejects, at level 0.05, homogeneity between the considered marking distributions (Wilcoxon–Mann–Whitney, p-value $= 0.000$; and Kolmogorov–Smirnov, p-value $= 0.003$). In Figure 6 we show the histograms corresponding to the full data sets and the progressive effects of best trimming, minimizing the Wasserstein distance between the remaining subsample distributions. The white portions of the bars represent the trimmed observations when the trimming size is $\alpha = 0.05$, the union of the white and yellow portions are the trimmed observations when $\alpha = 0.1$ and the orange portions complete the trimming corresponding to $\alpha = 0.15$. Notice that the best trimming is far from being symmetric.

In Table 4 we have included the p-values corresponding to the bootstrap procedure introduced in Section 3. In every case, for fixed $\beta = 0.05$ and taking α_n as in Theorem 4, we used 1000 bootstrap samples to compute the p-values for the null hypothesis H_0: $d_{\mathrm{TV}}(P, Q) \leq \alpha$. In general terms, these p-values show that both samples are not 0.05-similar, but they can be considered 0.10-similar. The considerations made in Section 3 about Condition (12) show the convenience

Table 4. Bootstrap p-values arising from the introduced bootstrap methodology, applied to the similarity analysis between markers ($\beta = 0.05$)

	ρ:	1		4/5		2/3		1/2	
α	γ:	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
0		0	0	0	0	0	0	0	0
0.05		0.059	0.133	0.016	0.058	0.007	0.034	0.005	0.019
0.10		0.884	0.975	0.717	0.865	0.567	0.708	0.371	0.597
0.15		1	1	1	1	1	1	0.997	0.999
0.20		1	1	1	1	1	1	1	1

of using resampling orders less than or equal to $n^{4/5}$, as we don't know if the supports of the contaminating distributions are well separated or not.

Appendix

A.1. Proof of Theorem 2

Our proof is based on a parallel result for the one-sample case. Let P_n be the empirical measure based on i.i.d. random variables $X_1, \ldots, X_n$ with common distribution P. In the particular case $P = Q$ and $\alpha = 0$ we have $n\mathcal{W}_2^2(P_n, Q) = \mathrm{O}_P(1)$ under sufficient integrability assumptions (see [10]). From the obvious bound $\mathcal{W}_2(\mathcal{R}_\alpha(P_n), Q) \le \mathcal{W}_2(P_n, Q)$ we see that $n\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q) = \mathrm{O}_P(1)$. Our first result here shows that $n\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q) = \mathrm{o}_P(1)$ even if $P \neq Q$.

Theorem 5. *Assume that $Q \in \mathcal{R}_{\alpha_0}(P)$ for some $\alpha_0 \in [0, 1)$, where Q is supported in a bounded interval, having a density function that is bounded away from zero on its support, and with a bounded derivative. If $\alpha_n \ge \alpha_0 + r_n/\sqrt{n}$ for some sequence $0 \le r_n \to \infty$, then*

$$\sqrt{n}\mathcal{W}_2(\mathcal{R}_{\alpha_n}(P_n), Q) \to 0 \qquad \text{in probability as } n \to \infty.$$

Proof. Arguing as in the proof of Proposition 2 we can check that $Q \in \mathcal{R}_{\alpha_0}(P)$ is equivalent to $P = (1-\alpha_0)Q + \alpha_0 P'$ for some distribution P'. Hence, we can assume $X_n = (1-U_n)Y_n + U_n Z_n$, where $\{Y_n\}_n$, $\{Z_n\}_n$ and $\{U_n\}_n$ are independent i.i.d. sequences with laws Q, P' and Bernoulli with mean α_0, respectively. Write $N_n = \sum_{i=1}^n I(U_i = 1)$. Then N_n follows a binomial distribution with parameters n and α_0. Hence, $\sqrt{n}(N_n/n - \alpha_0) \to \sqrt{\alpha_0(1-\alpha_0)}Z$, with Z standard normal. We assume w.l.o.g. that convergence holds, in fact, a.s. Write $n' = n - N_n$, $\tilde{X}_1, \ldots, \tilde{X}_{n'}$ for the Y_i's in the sample with associated $U_i = 0$ (the uncontaminated fraction of the sample: $\tilde{X}_1, \ldots, \tilde{X}_{n'}$ are i.i.d. Q) and $\tilde{P}_{n'}$ for the empirical measure on the $\tilde{X}_i$'s. Observe that $\tilde{P}_{n'} \in \mathcal{R}_{\tilde{\alpha}_n}(P_n)$ with $\tilde{\alpha}_n = N_n/n$. Now we note that given $\alpha, \beta \in [0, 1)$, if $Q \in \mathcal{R}_\alpha(P)$, then $\mathcal{R}_\beta(Q) \subset \mathcal{R}_{\alpha+\beta-\alpha\beta}(P)$. Hence, $\mathcal{R}_{\hat{\alpha}_n}(\tilde{P}_{n'}) \subset \mathcal{R}_{\alpha_n}(P_n)$ for $\hat{\alpha}_n = (\alpha_n - \tilde{\alpha}_n)/(\tilde{\alpha}_n)$ provided

$\alpha_n > \tilde{\alpha_n}$, which eventually holds. Consequently,

$$\mathcal{W}_2(\mathcal{R}_{\alpha_n}(P_n), Q) \le \mathcal{W}_2(\mathcal{R}_{\hat{\alpha}_n}(\tilde{P}_{n'}), Q).$$

Thus, the result will follow if we prove it in the particular case $P = Q$ and $\alpha_0 = 0$.

We proceed in this case writing F and f for the distribution and density functions of P. Recalling the parametrization in (7) we have

$$\mathcal{W}_2^2(\mathcal{R}_{\alpha_n}(P_n), P) = \min_{h \in \mathcal{C}_{\alpha_n}} \mathcal{W}_2^2((P_n)_h, P) = \min_{h \in \mathcal{C}_{\alpha_n}} \int_0^1 \big(F_n^{-1}(h^{-1}(t)) - F^{-1}(t)\big)^2 \,\mathrm{d}t$$

and we see that $n\mathcal{W}_2^2(\mathcal{R}_{\alpha_n}(P_n), P) = \min_{h \in \mathcal{C}_{\alpha_n}} M_n(h)$, where

$$M_n(h) = \int_0^1 \left(\frac{\rho_n(t)}{f(F^{-1}(t))} - \sqrt{n}\big(F^{-1}(h(t)) - F^{-1}(t)\big) \right)^2 h'(t) \,\mathrm{d}t$$

and $\rho_n(t) = \sqrt{n} f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$ is the weighted quantile process. Without loss of generality, we can assume that $\{X_n\}_n$ are defined in a sufficiently rich probability space in which there exist Brownian bridges, B_n, satisfying

$$n^{1/2-\nu} \sup_{1/n \le t \le 1-1/n} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} = \begin{cases} \mathrm{O}_P(\log n), & \text{if } \nu = 0, \\ \mathrm{O}_P(1), & \text{if } 0 < \nu \le 1/2 \end{cases} \tag{13}$$

(this is guaranteed by Theorem 6.2.1 in [8]). Now, defining

$$\tilde{N}_n(h) = \int_0^1 \left(\frac{B_n(t)}{f(F^{-1}(t))} - \sqrt{n}\big(F^{-1}(h(t)) - F^{-1}(t)\big) \right)^2 h'(t) \,\mathrm{d}t,$$

and assuming w.l.o.g. that $\alpha_n \le 1 - \delta$ for some $\delta > 0$ we have that

$$\sup_{h \in \mathcal{C}_\alpha} |M_n(h)^{1/2} - \tilde{N}_n(h)^{1/2}| \le \left(\frac{1}{\delta} \int_0^1 \left(\frac{\rho_n(t) - B_n(t)}{f(F^{-1}(t))} \right)^2 \mathrm{d}t \right)^{1/2} = \mathrm{o}_P(1).$$

The last equality follows from (13), taking $\nu = 0$, because, since f is bounded below

$$\int_{1/n}^{1-1/n} \left(\frac{\rho_n(t) - B_n(t)}{f(F^{-1}(t))} \right)^2 \mathrm{d}t \le \frac{\log n}{\sqrt{n}} \int_0^1 \frac{1}{f^2(F^{-1}(t))} \,\mathrm{d}t \mathrm{O}_P(1) = \mathrm{o}_P(1).$$

Thus, the conclusion will follow if we show $\min_{h \in \mathcal{C}_{\alpha_n}} \tilde{N}_n(h) \to 0$ in probability or, equivalently, if we show that $\min_{h \in \mathcal{C}_{\alpha_n}} N_n(h) \to 0$ in probability, where

$$N_n(h) = \int_0^1 \left(\frac{B(t)}{f(F^{-1}(t))} - \sqrt{n}\big(F^{-1}(h(t)) - F^{-1}(t)\big) \right)^2 h'(t) \,\mathrm{d}t$$

and B is a fixed Brownian bridge. To check that $\min_{h\in\mathcal{C}_{\alpha_n}} N_n(h) \to 0$ in probability, we observe that $\min_{h\in\mathcal{C}_{\alpha_n}} N_n(h) \le \frac{1}{\delta} \min_{k\in\mathcal{G}_n} R_n(k)$, where

$$R_n(k) = \int_0^1 \left(\frac{B(t)}{f(F^{-1}(t))} - \sqrt{n}\big(F^{-1}\big(t + k(t)/\sqrt{n}\big) - F^{-1}(t)\big) \right)^2 \mathrm{d}t$$

and $\mathcal{G}_n$ is the set of real-valued, absolutely continuous functions on [0, 1] such that $k(0) = k(1) = 0$ and $-\sqrt{n} \le k'(t) \le r_n$ for almost every t. We assume w.l.o.g. $r_n \le r_{n+1}$ for every n. Then $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ for every n and $\mathcal{G} := \bigcup_{n\ge 1} \mathcal{G}_n$ is the set of all absolutely continuous functions on [0, 1] such that $k(0) = k(1) = 0$ and k' is (essentially) bounded. From our hypotheses it follows easily that, for $k \in \mathcal{G}$,

$$R_n(k) \to R(k) := \int_0^1 \left(\frac{B(t) - k(t)}{f(F^{-1}(t))} \right)^2 \mathrm{d}t$$

and hence $\min_{k\in\mathcal{G}_n} R_n(k) \to 0$ (therefore $n\mathcal{W}_2^2(\mathcal{R}_{\alpha_n}(P_n), P) \to 0$) will follow if we show that $\inf_{k\in\mathcal{G}} R(k) = 0$. But this can be checked easily by noting, for instance, that if k_n is the function that interpolates $B(t)$ at knots i/n, $i = 0, \dots, n$, and is linear in between, then we have $k_n \in \mathcal{G}$ and $R(k_n) \to 0$. □

Proof of Theorem 2. We write $\alpha_0 = d_{\mathrm{TV}}(P, Q)$ and take P_0 as in the canonical decomposition in Proposition 2 (we take μ to be the Lebesgue measure there). Then $P_0 \in \mathcal{R}_{\alpha_0}(P)$ holds with P and P_0 playing the roles of P and Q and the density of P_0 satisfies the assumptions in Theorem 5 (in fact $f_0 = (f \wedge g)/(1 - \alpha_0)$ has a bounded derivative a.e., but this suffices for the strong approximation in the proof of Theorem 5). Hence, $\sqrt{n}\mathcal{W}_2(\mathcal{R}_{\alpha_n}(P_n), P_0) \to 0$ in probability and similarly for $\sqrt{n}\mathcal{W}_2(\mathcal{R}_{\alpha_n}(Q_n), P_0)$. The triangle inequality for $\mathcal{W}_2$ yields the conclusion. □

A.2. Asymptotic theory for the bootstrap

The behavior of the bootstrap p-value under the alternative follows from the next result.

Proposition 4. *Assume $X_{n,1}, \dots, X_{n,n'}$; $Y_{n,1}, \dots, Y_{n,m'}$ are i.i.d. random variables with common distribution $P_n \in \mathcal{F}_2$ such that $\mathcal{W}_2(P_n, P) \to 0$. If $P^*_{n'}$ and $Q^*_{m'}$ denote the empirical measures on $X_{n,1}, \dots, X_{n,n'}$ and $Y_{n,1}, \dots, Y_{n,m'}$, respectively, and $n', m' \to \infty$, then*

$$\mathcal{W}_2(P^*_{n'}, Q^*_{m'}) \to 0 \qquad \textit{in probability.}$$

Proof. By Proposition 3 it is enough to consider the case $P_n = P$ for all n. But then $P_{n'} \to_w P$ a.s. by the Glivenko–Cantelli theorem while the law of large numbers gives convergence of second-order moments. These two facts imply that $\mathcal{W}_2(P^*_{n'}, P) \to 0$ (and for $\mathcal{W}_2(Q^*_{m'}, P)$ as well). □

Now we take care of the null hypothesis. The next result will be useful for P and Q away from the boundary. Its proof is analogous to that of Theorem 2.1 in [4].

Proposition 5. *Assume $X_{n,1}, \dots, X_{n,n'}$ are i.i.d. random variables with common distribution $P_n \in \mathcal{F}_2$ such that $\mathcal{W}_2(P_n, P) \to 0$. If $\bar{X}_{n,n'} := \frac{1}{n'}\sum_{i=1}^{n'} X_{n,i}$, then*

$$\sqrt{n'}(\bar{X}_{n,n'} - \mu_n) \to_w N(0, \sigma^2),$$

where $\mu_n = E(\bar{X}_{n,n'})$ and σ^2 is the variance of P.

Proof of Theorem 3. We will assume for simplicity $n = m$ and $n' = m'$. The general case can be handled with straightforward modifications. We consider first the case $d_{\mathrm{TV}}(P, Q) > \alpha$. In this case we have (Theorem 1) that $\mathcal{W}_2(P_{n,\alpha_n}, P_\alpha) \to 0$ and $\mathcal{W}_2(Q_{n,\alpha_n}, Q_\alpha) \to 0$ a.s. Since

$$\mathcal{W}_2^2\bigl(aP_1 + (1-a)P_2, aQ_1 + (1-a)Q_2\bigr) \le a\mathcal{W}_2^2(P_1, Q_1) + (1-a)\mathcal{W}_2^2(P_2, Q_2)$$

for probabilities $P_i, Q_i \in \mathcal{F}_2$ and $a \in [0,1]$ (see [2]) it follows that $\mathcal{W}_2(R_{n,n}, \lambda P_\alpha + (1 - \lambda)Q_\alpha) \to 0$ a.s. Note that

$$p_{n,n}^* = \mathbb{P}^*\left(\mathcal{W}_2(P_{n'}^*, Q_{n'}^*) > \sqrt{\frac{n}{n'}}\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n})\right).$$

Now, Theorem 1 implies that $\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n}) \to \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) > 0$, while n/n' is bounded away from 0 by assumption. This, together with Proposition 4, gives (ii).

We assume now that $d_{\mathrm{TV}}(P, Q) < \alpha$. Then Theorem 2 ensures that $\sqrt{n}\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n}) \to 0$ in probability. Now, if P_1, P_2 are probabilities in $\mathcal{F}_2$ with means μ_1, μ_2 and $\bar{P}_1, \bar{P}_2$ are their centered versions, then it is easy to check that $\mathcal{W}_2^2(P_1, P_2) = (\mu_1 - \mu_2)^2 + \mathcal{W}_2^2(\bar{P}_1, \bar{P}_2)$ and, therefore, $\mathcal{W}_2^2(P_1, P_2) \ge (\mu_1 - \mu_2)^2$. Let $\bar{X}_{n'}^*$ and $\bar{Y}_{n'}^*$, respectively, denote the means corresponding to the X's and Y's bootstrap samples, and μ_n be the mean of the parent bootstrap distribution, $R_{n,n}$. Then

$$n'\mathcal{W}_2^2(P_n^*, Q_m^*) \ge n'(\bar{X}_{n'}^* - \bar{Y}_{n'}^*)^2 = \bigl(\sqrt{n'}(\bar{X}_{n'}^* - \mu_n) - \sqrt{n'}(\bar{Y}_{n'}^* - \mu_n)\bigr)^2.$$

From the Glivenko–Cantelli theorem we have a.s. tightness of $\{P_n\}_n$ and $\{Q_n\}_n$ and, as a consequence, of P_{n,α_n} and Q_{n,α_n} (see Proposition 2.1 in [2]). We can assume, taking subsequences if necessary, that $P_{n,\alpha_n} \to_w P_0$ and $Q_{n,\alpha_n} \to_w Q_0$ for some probabilities P_0, Q_0. A little thought shows that, necessarily, $P_0 \in \mathcal{R}_\alpha(P)$ and $Q_0 \in \mathcal{R}_\alpha(Q)$. Since $\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n}) \to 0$, necessarily, $P_0 = Q_0 \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$. Also, since $P, Q \in \mathcal{F}_2$, the strong law of large numbers shows that the map x^2 is uniformly integrable with respect to $\{P_n\}_n$ and $\{Q_n\}_n$ a.s., hence also with respect to $\{P_{n,\alpha_n}\}_n$ and $\{Q_{n,\alpha_n}\}_m$. Thus, perhaps through subsequences, $\mathcal{W}_2(P_{n,\alpha_n}, P_0) \to 0$ and $\mathcal{W}_2(Q_{n,\alpha_n}, P_0) \to 0$, hence $\mathcal{W}_2(R_{n,n}, P_0) \to 0$ for some $P_0 \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$.

The function that sends P to its variance is continuous in $\mathcal{F}_2$ for the $\mathcal{W}_2$ metric. Hence, since $\mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$ is compact, the variance attains its minimum there. Let us write $\sigma_0^2 = \min_{R \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)} \mathrm{Var}(R)$. Then $\sigma_0 > 0$ (a trimming of a probability with a density has a density, hence, cannot have null variance) and if we write σ^2 for the variance of P_0, we have

$$\begin{aligned}p^*_{n,n} &= \mathbb{P}^*\big(\sqrt{n'}\mathcal{W}_2(P^*_{n'}, Q^*_{n'}) > \sqrt{n}\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n})\big)\\ &\geq \mathbb{P}^*\left(\left|\frac{\sqrt{n'}}{2\sigma}(\bar{X}^*_{n'} - \bar{Y}^*_{n'})\right| > \frac{\sqrt{n}}{2\sigma}\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n})\right)\\ &\geq \mathbb{P}^*\left(\left|\frac{\sqrt{n'}}{2\sigma}(\bar{X}^*_{n'} - \bar{Y}^*_{n'})\right| > \frac{\sqrt{n}}{2\sigma_0}\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n})\right).\end{aligned}$$

Thus, Proposition 5 and the fact that $\sqrt{n}\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n}) \to 0$ yield that $p^*_{n,n} \to 1$ in probability, showing (i). □

Proof of Theorem 4. As in the proof of Theorem 2, we assume that $X_n = (1 - U_n)A_n + U_n B_n$, $Y_n = (1 - V_n)C_n + V_n D_n$ with $\{A_n\}_n$, $\{B_n\}_n$, $\{C_n\}_n$, $\{D_n\}_n$, $\{U_n\}_n$, $\{V_n\}_n$ independent i.i.d. sequences of which $\{A_n\}_n$ and $\{C_n\}_n$ have common distribution P_0 while $\{U_n\}_n$ and $\{V_n\}_n$ are Bernoulli with mean α. We write $N_n = \sum_{i=1}^n I(U_i = 1)$ and $M_n = \sum_{i=1}^n I(V_i = 1)$. Also we put $n'_1 = n - N_n$, $n'_2 = n - M_n$ and write $\tilde{X}_1, \ldots, \tilde{X}_{n'_1}$ and $\tilde{Y}_1, \ldots, \tilde{Y}_{n'_2}$ for the data corresponding to $U_i = 0$ and $V_i = 0$, respectively.

On the set $E_n := (N_n \leq n\alpha_n, M_n \leq n\alpha_n)$, the empirical measures on $\tilde{X}_1, \ldots, \tilde{X}_{n'_1}$ and $\tilde{Y}_1, \ldots, \tilde{Y}_{n'_1}$ (which we denote $\tilde{P}_{n'_1}$ and $\tilde{Q}_{n'_2}$) satisfy $\tilde{P}_{n'_1} \in \mathcal{R}_{\alpha_n}(P_n)$ and $\tilde{Q}_{n'_2} \in \mathcal{R}_{\alpha_n}(Q_n)$. Hence, we have $\mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n}) \leq \mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2})$. Thus,

$$\mathbb{P}(p^*_{n,n} \leq \beta) \leq \mathbb{P}(E_n^C) + \mathbb{P}\big((\tilde{p}^*_n \leq \beta) \cap E_n\big),$$

where

$$\tilde{p}^*_n = \mathbb{P}^*\big(\sqrt{n'}\mathcal{W}_2(P^*_{n'}, Q^*_{n'}) > \sqrt{n(1-\alpha)}\mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2})\big).$$

By the central limit theorem (CLT) we have $\mathbb{P}(E_n^C) \to \gamma$. Hence it suffices to control $\mathbb{P}((\tilde{p}^*_n \leq \beta) \cap E_n)$. If $J_1, \ldots, J_{n'}, L_1, \ldots, L_{n'}$ are i.i.d. random variables with law P_0, independent of the data (both original and bootstrap) and $\mu_{n'}$, $\nu_{n'}$ are the empirical measures, then Theorem 3 and the fact that $\mathcal{W}_2(\mathcal{L}(aX), \mathcal{L}(aY)) = a\mathcal{W}_2(\mathcal{L}(X), \mathcal{L}(Y))$ for $a > 0$ imply

$$\mathcal{W}_2\big(\mathcal{L}^*\big(\sqrt{n'}\mathcal{W}_2(P^*_{n'}, Q^*_{n'})\big), \mathcal{L}\big(\sqrt{n'}\mathcal{W}_2(\mu_{n'}, \nu_{n'})\big)\big) \leq 2\sqrt{n'}\mathcal{W}_2(R_{n,n}, P_0).$$

By Lemma 1 below $\sqrt{n'}\mathcal{W}_2(R_{n,n}, P_0)I_{E_n} \to 0$ in probability. The assumptions on P and Q yield that $\sqrt{n'}\mathcal{W}_2(\mu_{n'}, \nu_{n'})$ converges weakly to a non-null limiting distribution as in (9) (with a proof as in Theorem 4.6 in [10]). We call η the limit probability measure. Then

$$\big|\tilde{p}^*_n - \eta\big(\big(\sqrt{n(1-\alpha)}\mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2}), \infty\big)\big)\big|I_{E_n} \to 0$$

in probability. As a consequence,

$$\mathbb{P}\big((\tilde{p}^*_n \leq \beta) \cap E_n\big) - \mathbb{P}\big(\big(\eta\big(\big(\sqrt{n(1-\alpha)}\mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2}), \infty\big)\big) \leq \beta\big) \cap E_n\big) \to 0.$$

But

$$\mathbb{P}\Big(\big(\eta\big(\big(\sqrt{n(1-\alpha)}\mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2}), \infty\big)\big) \leq \beta\big) \cap E_n\Big)$$
$$\leq \mathbb{P}\Big(\big(\eta\big(\big(\sqrt{n(1-\alpha)}\mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2}), \infty\big)\big) \leq \beta\big)\Big) \to \beta,$$

since, as above, $\sqrt{n(1-\alpha)}\mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2})$ converges weakly to η. This completes the proof. □

The following technical result has been used in the proof of Theorem 4.

Lemma 1. *With the notation and assumptions of Theorem 4,*

$$\sqrt{n'}\mathcal{W}_2(R_{n,n}, P_0)I_{E_n} = \mathrm{o}_P(1).$$

Proof. We use the parametrization in (7). We have $P_{n,\alpha_n} = (P_n)_{h_n}$, $Q_{n,\alpha_n} = (Q_n)_{l_n}$, for some $h_n, l_n \in \mathcal{C}_{\alpha_n}$. Writing F_n^{-1}, G_n^{-1}, F^{-1} and G^{-1} for the quantile functions of P_n, Q_n, P and Q we have $\mathcal{W}^2(P_{n,\alpha_n}, Q_{n,\alpha_n}) = \|F_n^{-1} \circ h_n^{-1} - G_n^{-1} \circ l_n^{-1}\|_2$, with $\|\cdot\|_2$ denoting the usual norm in $L_2(0,1)$, namely, $\|b\|_2^2 = \int_0^1 b^2$. Now

$$\|(F_n^{-1} \circ h_n^{-1} - G_n^{-1} \circ l_n^{-1}) - (F^{-1} \circ h_n^{-1} - G^{-1} \circ l_n^{-1})\|_2$$
$$\leq \|F_n^{-1} \circ h_n^{-1} - F^{-1} \circ h_n^{-1}\|_2 + \|G_n^{-1} \circ l_n^{-1} - G^{-1} \circ l_n^{-1}\|_2$$
$$\leq \frac{1}{\sqrt{1-\alpha_n}}(\|F_n^{-1} - F^{-1}\|_2 + \|G_n^{-1} - G^{-1}\|_2),$$

where we have used that $\int_0^1 (F^{-1}(h^{-1}(t)) - G^{-1}(h^{-1}(t))^2\,\mathrm{d}t = \int_0^1 (F^{-1}(x) - G^{-1}(x)^2 h'(x)\,\mathrm{d}x$. The assumptions on P and Q ensure that, as in (9), $\|F_n^{-1} - F^{-1}\|_2 + \|G_n^{-1} - G^{-1}\|_2 = \mathrm{O}_P(n^{-1/2})$. On the other hand, on E_n,

$$\|F_n^{-1} \circ h_n^{-1} - G_n^{-1} \circ l_n^{-1}\|_2 = \mathcal{W}_2(P_{n,\alpha_n}, Q_{n,\alpha_n}) \leq \mathcal{W}_2(\tilde{P}_{n'_1}, \tilde{Q}_{n'_2}) = \mathrm{O}_P(n^{-1/2}).$$

Combining these two facts we see that $\mathcal{W}_2(P_{h_n}, Q_{h_n})I_{E_n} = \|F^{-1} \circ h_n^{-1} - G^{-1} \circ l_n^{-1}\|_2 I_{E_n} = \mathrm{O}_P(n^{-1/2})$. Using (12) we see that $\mathcal{W}_2(P_{h_n}, P_0) = \mathrm{O}(n^{-\rho/2})$. Since $\mathcal{W}_2(P_{h_n}, P_{n,\alpha_n}) = \mathrm{O}_P(n^{-1/2})$, we conclude that $\mathcal{W}_2(P_{n,\alpha_n}, P_0)I_{E_n} = \mathrm{O}(n^{-\rho/2})$. Convexity and a similar argument for Q_{n,α_n} yield the result. □

Proof of Example 1. The fact that $d_{\mathrm{TV}}(P, Q) = \alpha$ follows from noting (with some abuse of notation) that for $\tilde{F}^{-1} \in \mathcal{R}_\alpha(P)$ and $\tilde{G}^{-1} \in \mathcal{R}_\alpha(Q)$

$$\tilde{F}^{-1}(t) \leq F^{-1}\big(\alpha + (1-\alpha)t\big) \leq \tilde{G}^{-1}(t).$$

Hence, the probability P_0 with quantile $F_0^{-1}(t) = F^{-1}(\alpha + (1-\alpha)t)$ is the unique element in $\mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$. Next we observe that, for $\tilde{F}^{-1} \in \mathcal{R}_{\alpha_n}(P)$,

$$F^{-1}(t) \leq F^{-1}\big(\alpha_n + (1-\alpha_n)t\big)$$
$$\leq F_0^{-1}(t) + \big(F^{-1}\big(\alpha_n + (1-\alpha_n)t\big) - F^{-1}\big(\alpha + (1-\alpha_n)t\big)\big).$$

Similarly, if $\tilde{G}^{-1} \in \mathcal{R}_{\alpha_n}(Q)$, $\tilde{G}^{-1}(t) \geq F_0^{-1}(t) - (F^{-1}(\alpha_n + (1-\alpha_n)t) - F^{-1}(\alpha + (1-\alpha_n)t))$ and, combining both inequalities, we get $|F_0^{-1}(t) - \tilde{F}^{-1}(t)| \leq |\tilde{F}^{-1}(t) - \tilde{G}^{-1}(t)| + |F^{-1}(\alpha_n + (1-\alpha_n)t) - F^{-1}(\alpha + (1-\alpha_n)t)|$ and the bound follows from the triangle inequality. □

Proof of Example 2. We write F_0 for the distribution function of P_0, hence, $F_0^{-1}(y) = \mu/2 + F^{-1}((1-\alpha)y)$ for $y \in (0, 1/2]$ and $F_0^{-1}(y) = -\mu/2 + F^{-1}(\alpha + (1-\alpha)y)$ for $y \in [1/2, 1)$. Similarly, we write $\tilde{F}_n$ and $\tilde{G}_n$ for the distribution functions of $\tilde{P}_n$ and $\tilde{Q}_n$, respectively. Necessarily, $\tilde{P}_n(0,\infty) \leq \frac{1}{1-\alpha_n}(1 - F(\frac{\mu}{2})) = \frac{1}{2}(1 + \frac{K}{(1-\alpha_n)\sqrt{n}})$. We write $\beta_n = \frac{1}{2} - \tilde{P}_n(0,\infty)$. It follows from the fact that $\mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) \to 0$ that $\mathcal{W}_2(\tilde{P}_n, P_0) \to 0$ and, therefore, that $\beta_n \to 0$. We give next a lower bound for $\mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n)$, assuming that $\beta_n > 0$. If this is the case

$$\tilde{F}_n^{-1}(t) \leq -\frac{\mu}{2} + F^{-1}\biggl(\alpha + (1-\alpha_n)(t-\beta_n) + \frac{K}{2\sqrt{n}}\biggr), \qquad t \in \biggl(0, \frac{1}{2} + \beta_n\biggr). \tag{14}$$

On the other hand $\tilde{G}_n^{-1}((1-\alpha_n)t) \geq \mu/2 + F^{-1}((1-\alpha_n)t)$. Standard computations show that there is a unique $a = a(\beta_n) > 0$ such that $F(a - \frac{\mu}{2}) - F(a + \frac{\mu}{2}) + \alpha = (1-\alpha)\beta_n$ and that

$$-\frac{\mu}{2} + F^{-1}\bigl(\alpha + (1-\alpha)(t-\beta)\bigr) \leq \mu/2 + F^{-1}\bigl((1-\alpha)t\bigr)$$

for $t \in (\frac{1}{1-\alpha}F(-a - \frac{\mu}{2}), \frac{1}{2})$. From this we get that

$$\mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) \geq \sqrt{g_1(\beta_n)} - s_{n,1} - s_{n,2}, \tag{15}$$

where $g_1(\beta) = \int_{F(-a-\mu/2)/(1-\alpha)}^{1/2} (\mu + F^{-1}((1-\alpha)t) - F^{-1}(\alpha + (1-\alpha)(t-\beta)))^2\,\mathrm{d}t$, $s_{n,1}^2 = \int_{F(-a-\mu/2)/(1-\alpha)}^{1/2} (F^{-1}((1-\alpha)t) - F^{-1}((1-\alpha_n)t))^2\,\mathrm{d}t$, $s_{n,2}^2 = \int_{F(-a-\mu/2)/(1-\alpha)}^{1/2} (F^{-1}(\alpha + (1-\alpha)(t-\beta_n)) - F^{-1}(\alpha + (1-\alpha_n)(t-\beta_n) + \frac{K}{2\sqrt{n}}))^2\,\mathrm{d}t$. A routine use of Taylor expansions yields $\lim_{\beta\to 0+} \frac{g_1(\beta)}{\beta^{5/2}} = (1-\alpha)^{3/2}\frac{\sqrt{|f'(\mu/2)|}}{f^2(\mu/2)} > 0$, $s_{n,1}^2 = \mathrm{O}(\sqrt{\beta_n}n^{-1})$ and $s_{n,2}^2 = \mathrm{O}(\sqrt{\beta_n}n^{-1})$. From this and (15) we obtain

$$\beta_n = \mathrm{O}(n^{-2/5}), \tag{16}$$

with a similar bound being satisfied by $\gamma_n = \frac{1}{2} - \tilde{Q}_n(-\infty, 0)$.

We turn now to the upper bound for $\mathcal{W}_2(\tilde{P}_n, P_0)$. From the triangle inequality we get

$$\begin{aligned}\mathcal{W}_2(\tilde{P}_n, P_0) &\leq \biggl(\int_0^{1/2} (\tilde{F}_n^{-1} - F_0^{-1})^2\biggr)^{1/2} + \biggl(\int_{1/2}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2\biggr)^{1/2}\\ &\leq \mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) + \biggl(\int_0^{1/2} (\tilde{G}_n^{-1} - F_0^{-1})^2\biggr)^{1/2} + \biggl(\int_{1/2}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2\biggr)^{1/2}.\end{aligned}$$

We consider next $\int_{1/2}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2$. Since $\tilde{P}_n \in \mathcal{R}_{\alpha_n}(P)$ we have

$$\tilde{F}_n^{-1}(t) \leq -\frac{\mu}{2} + F^{-1}\bigl(\alpha_n + (1-\alpha_n)t\bigr), \qquad t \in (0,1). \tag{17}$$

Keeping the above notation for β_n, assume first that $\beta_n \le 0$. Then

$$\tilde{F}_n^{-1}(t) \ge -\frac{\mu}{2} + F^{-1}\left(\alpha + (1-\alpha_n)t + \frac{K}{2\sqrt{n}}\right), \qquad t \in \left(\frac{1}{2}, 1\right) \tag{18}$$

(this follows upon noting that $\tilde{F}_n^{-1}(\frac{1}{2}+) \ge 0$ and $\tilde{F}_n^{-1}(t) = F^{-1}(h^{-1}(t))$, h^{-1} growing with slope at least $1-\alpha_n$). For $t \in (\frac{1}{2}, 1)$, (17) and (18) still hold if we replace $\tilde{F}_n^{-1}$ by F_0^{-1}. Hence, in this case $\int_{1/2}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2 \le \int_{1/2}^1 (F^{-1}(\alpha_n + (1-\alpha_n)t) - F^{-1}(\alpha_n + (1-\alpha_n)t - \frac{K}{2\sqrt{n}}))^2 \, \mathrm{d}t =: s_{n,3}^2$.

If $\beta_n > 0$, then, arguing as above, we have

$$\tilde{F}_n^{-1}(t) \ge -\frac{\mu}{2} + F^{-1}\left(\alpha + (1-\alpha_n)(t-\beta_n) + \frac{K}{2\sqrt{n}}\right), \qquad t \in \left(\frac{1}{2} + \beta_n, 1\right), \tag{19}$$

while (14) holds in $(0, \frac{1}{2} + \beta_n)$. Now we use the bound

$$\left(\int_{1/2}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} \le \left(\int_{1/2}^{1/2+\beta_n} (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} + \left(\int_{1/2+\beta_n}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2}$$

and proceed as follows. For $t \in (\frac{1}{2} + \beta_n, 1)$ (17) and (19) hold again after replacing $\tilde{F}_n^{-1}$ by F_0^{-1}. This and the triangle inequality yield

$$\begin{aligned} &\left(\int_{1/2+\beta_n}^1 (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} \\ &\quad \le \left(\int_{1/2+\beta_n}^1 \bigl(F^{-1}(\alpha + (1-\alpha)t) - F^{-1}(\alpha + (1-\alpha)(t-\beta_n))\bigr)^2 \, \mathrm{d}t\right)^{1/2} \\ &\qquad + 2\left(\int_{1/2}^1 \left(F^{-1}(\alpha_n + (1-\alpha_n)t) - F^{-1}\left(\alpha_n + (1-\alpha_n)t - \frac{K}{2\sqrt{n}}\right)\right)^2 \mathrm{d}t\right)^{1/2} \\ &\quad = \sqrt{g_2(\beta_n)} + 2s_{n,3}. \end{aligned} \tag{20}$$

For the interval $(\frac{1}{2}, \frac{1}{2} + \beta_n)$ we write $\underline{G}^{-1}(t) = \frac{\mu}{2} + F^{-1}((1-\alpha_n)t)$ (the minimal quantile function in $\mathcal{R}_{\alpha_n}(Q)$). Then $(\int_{1/2}^{1/2+\beta_n} (\tilde{F}_n^{-1} - F_0^{-1})^2)^{1/2} \le (\int_{1/2}^{1/2+\beta_n} (\tilde{F}_n^{-1} - \underline{G}^{-1})^2)^{1/2} + (\int_{1/2}^{1/2+\beta_n} (\underline{G}^{-1} - F_0^{-1})^2)^{1/2}$. We observe now that $\tilde{G}^{-1}(t) \ge \underline{G}_n^{-1}(t)$ and also that, for $t \in (\frac{1}{2}, \frac{1}{2} + \beta_n)$, $-\frac{\mu}{2} + F^{-1}(\alpha + (1-\alpha)(t-\beta_n)) \le 0 \le \frac{\mu}{2} + F^{-1}((1-\alpha)t)$. Combining these facts with (14) we obtain

$$\begin{aligned} |\tilde{F}_n^{-1}(t) - \underline{G}^{-1}(t)| \le {} & |\tilde{F}_n^{-1}(t) - \tilde{G}_n^{-1}(t)| \\ & + \bigl|F^{-1}\bigl((1-\alpha_n)t\bigr) - F^{-1}\bigl((1-\alpha)t\bigr)\bigr| \\ & + \left|F^{-1}\left(\alpha + (1-\alpha_n)(t-\beta_n) + \frac{K}{2\sqrt{n}}\right) - F^{-1}\bigl(\alpha + (1-\alpha)(t-\beta_n)\bigr)\right|. \end{aligned}$$

As a consequence,

$$\begin{aligned}
&\left(\int_{1/2}^{1/2+\beta_n} (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} \\
&\quad\le \mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) + \left(\int_{1/2}^{1/2+\beta_n} \bigl(\mu + F^{-1}\bigl((1-\alpha)t\bigr) - F^{-1}\bigl(\alpha + (1-\alpha)t\bigr)\bigr)^2 \,\mathrm{d}t\right)^{1/2} \\
&\qquad + 2\left(\int_{1/2}^{1/2+\beta_n} \bigl(F^{-1}\bigl((1-\alpha_n)t\bigr) - F^{-1}\bigl((1-\alpha)t\bigr)\bigr)^2 \,\mathrm{d}t\right)^{1/2} \\
&\qquad + \left(\int_{1/2}^{1/2+\beta_n} \left(F^{-1}\left(\alpha + (1-\alpha_n)(t-\beta_n) + \frac{K}{2\sqrt{n}}\right)\right.\right. \\
&\qquad\qquad\qquad \left.\left. - F^{-1}\bigl(\alpha + (1-\alpha)(t-\beta_n)\bigr)\right)^2 \mathrm{d}t\right)^{1/2} \\
&\quad = \mathcal{W}_2(\tilde{P}_n, \tilde{Q}_n) + \sqrt{g_3(\beta_n)} + 2s_{n,4} + s_{n,5},
\end{aligned}$$

where $g_3(\beta) = \int_{1/2}^{1/2+\beta} (\mu + F^{-1}((1-\alpha)t) - F^{-1}(\alpha + (1-\alpha)t))^2 \,\mathrm{d}t$. Again a Taylor expansion shows that $g_3(\beta_n) = \mathrm{O}(\beta_n^3) = \mathrm{o}(n^{-1})$. Similarly, we get $s_{n,j} = \mathrm{o}(n^{-1})$, $j = 4, 5$, and, as a consequence

$$\left(\int_{1/2}^{1/2+\beta_n} (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} = \mathrm{O}(n^{-1/2}). \tag{21}$$

Collecting the estimates in (20) and (21), we obtain

$$\left(\int_{1/2}^{1} (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} \le \sqrt{g_2(\beta_n)} + 2s_{n,3} + \mathrm{O}(n^{-1/2}). \tag{22}$$

We note next that F^{-1} has a bounded derivative and, as a consequence, $s_{n,3}^2 = \mathrm{O}(n^{-1})$. Similarly, we find that $g_2(\beta_n) = \mathrm{O}(\beta_n^2)$. Summarizing,

$$\left(\int_{1/2}^{1} (\tilde{F}_n^{-1} - F_0^{-1})^2\right)^{1/2} = \mathrm{O}(n^{-2/5}).$$

A similar analysis works for $\int_0^{1/2} (\tilde{G}_n^{-1} - F_0^{-1})^2$ and completes the proof. □

Proof of Proposition 3. We take $(X_{1,1}, Y_{1,1})$ to be an optimal coupling for P and Q with respect to the $\|x - y\|^p$-cost and $(X_{1,i}, Y_{1,i})$, $2 \le i \le n$, and $(X_{2,j}, Y_{2,j})$, $1 \le j \le m$, independent copies of $(X_{1,1}, Y_{1,1})$ (hence $E\|X_{i,j} - Y_{i,j}\|^p = \mathcal{W}_p^p(P, Q)$). Then $S_{n,m} = \min_\pi (a(\pi))^{1/p}$ and $T_{n,m} = \min_\pi (b(\pi))^{1/p}$, where

$$a(\pi) = \sum_{1\le i\le n, 1\le j\le m} \pi_{i,j} \|X_{1,i} - X_{2,j}\|^p,$$

$b(\pi)$ is defined similarly by replacing $X_{i,j}$ by $Y_{i,j}$ and π takes values in the set of $n \times m$ matrices with non-negative entries $\pi_{i,j}$ such that $\sum_{1\le j\le m} \pi_{i,j} = \frac{1}{n}$ and $\sum_{1\le i\le n} \pi_{i,j} = \frac{1}{m}$.

We observe next that, by the triangle inequality,

$$\begin{aligned} |a(\pi)^{1/p} - b(\pi)^{1/p}| &\le \Biggl(\sum_{1\le i\le n, 1\le j\le m} \pi_{i,j} \|(X_{1,i} - X_{2,j}) - (Y_{1,i} - Y_{2,j})\|^p \Biggr)^{1/p} \\ &\le \biggl(\frac{1}{n} \sum_{1\le i\le n} \|X_{1,i} - Y_{1,i}\|^p \biggr)^{1/p} + \biggl(\frac{1}{m} \sum_{1\le j\le m} \|X_{2,j} - Y_{2,j}\|^p \biggr)^{1/p}. \end{aligned}$$

As a consequence, we have that $|S_{n,m} - T_{n,m}|$ is upper bounded by the right-hand side of the above display and, from the elementary inequality $(a+b)^p \le 2^{p-1}a^p + 2^{p-1}b^p$ for non-negative a, b, we get

$$\begin{aligned} E(S_{n,m} - T_{n,m})^p &\le 2^{p-1} E\|X_{1,1} - Y_{1,1}\|^p + 2^{p-1} E\|X_{2,1} - Y_{2,1}\|^p \\ &= 2^p \mathcal{W}_p^p(P, Q). \end{aligned}$$

This completes the proof. □

Acknowledgements

Research partially supported by the Spanish Ministerio de Ciencia e Innovación, Grant MTM2008-06067-C02-01, and 02 and by the Consejería de Educación y Cultura de la Junta de Castilla y León, GR150.

The authors would like to thank two anonymous referees for their careful reading of the manuscript, their suggestions and the pointers to relevant references that helped us to greatly improve our original version.

References

[1] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2008). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.* **103** 697–704. MR2435470

[2] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2011). Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. H. Poincaré Probab. Statist.* **47** 358–375. MR2814414

[3] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2010). Assessing when a sample is mostly normal. *Comput. Statist. Data Anal.* **54** 2914–2925.

[4] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

[5] Buja, A. (1986). On the Huber–Strassen theorem. *Probab. Theory Relat. Fields* **73** 149–152. MR0849070

[6] Caffarelli, L.A. and McCann, R.J. (2010). Free boundaries in optimal transport and Monge–Ampère obstacle problems. *Ann. of Math.* (2) **171** 673–730. MR2630054

[7] Cascos, I. and López-Díaz, M. (2008). Consistency of the α-trimming of a probability. Applications to central regions. *Bernoulli* **14** 580–592. MR2544103

[8] Csörgő, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Chichester: Wiley. MR1215046

[9] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: An attempt to robustify quantizers. *Ann*. *Statist*. **25** 553–576. MR1439314

[10] del Barrio, E., Giné, E. and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11** 131–189. MR2121458

[11] Figalli, A. (2010). The optimal partial transport problem. *Arch*. *Ration*. *Mech*. *Anal*. **195** 533–560. MR2592287

[12] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *Ann*. *Statist*. **36** 1324–1345. MR2418659

[13] Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J*. *Approx*. *Theory* **64** 162–180. MR1091467

[14] Gower, J.C. (2006). Measures of similarity, dissimilarity, and distance. In *Encyclopedia of Statistical Sciences*, 2nd ed. (S. Kotz, C.B. Read, N. Balakrishnan and B. Vidakovic, eds.) **12** 7730–7738. New York: Wiley.

[15] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann*. *Math*. *Statist*. **35** 73–101. MR0161415

[16] Huber, P.J. (1965). A robust version of the probability ratio test. *Ann*. *Math*. *Statist*. **36** 1753–1758. MR0185747

[17] Huber, P.J. and Strassen, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann*. *Statist*. **1** 251–263. MR0356306

[18] Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics* **47** 264–273. MR2164700

[19] Martínez-Camblor, P., De Uña-Álvarez, J. and Corral, N. (2008). k-sample test based on the common area of kernel density estimators. *J*. *Statist*. *Plann*. *Inference* **138** 4006–4020. MR2455983

[20] Massart, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math*. **1896**. Berlin: Springer. MR2319879

[21] Raghavachari, M. (1973). Limiting distributions of Kolmogorov–Smirnov type statistics under the alternative. *Ann*. *Statist*. **1** 67–73. MR0346976

[22] Rieder, H. (1977). Least favorable pairs for special capacities. *Ann*. *Statist*. **5** 909–921. MR0468005

[23] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, *Vol*. *B* (*Bad Tatzmannsdorf*, 1983) (W. Grossmann, G. Pflug, I. Vincze and W. Werz, eds.) 283–297. Dordrecht: Reidel. MR0851060

Received June 2010 and revised December 2010