Testing Similarity through Impartially Trimmed Wasserstein distances †

Pedro César Álvarez-Esteban

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

Eustasio del Barrio

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

Juan Antonio Cuesta-Albertos

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain

Carlos Matrán

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

Summary. This paper introduces an analysis of similarity of distributions based on the L_2-Wasserstein distance between trimmed distributions. Our main innovation is the use of the impartial trimming methodology, already considered in robust statistics, which we adapt to this setup. Instead of removing data at the tails as in Munk and Czado (1998) we develop a data-driven trimming method aimed at maximizing similarity between distributions. Dissimilarity is then measured in terms of the distance between the optimally trimmed distributions. We provide illustrative examples showing the improvements over previous approaches and give the relevant asymptotic results to justify the use of this methodology in applications.

Keywords: Trimmed distributions, similarity, Wasserstein distance, asymptotics, impartial trimming.

1. Introduction.

An intrinsic consequence of randomness is variability. Samples obtained from a random experiment will generally differ. This justifies the central role that similarity plays in Statistics. Since we cannot expect even two ideal samples coming from the same experiment to be the same, we should be able to detect departures from this idealized equality which cannot be reasonably justified only by randomness. Often the researcher is not really concerned about exact coincidence, but rather wants to guarantee that the parent distributions do not differ too much. For example, in bioequivalence studies, when testing two different drugs (or treatments), it is enough to check that both drugs have the same therapeutic effect on patients. To achieve this it is sufficient to have two similar distributions in the target variable. Moreover, a robust appeal to similarity leads to consider that samples initially taken to be similar, should also be considered similar after being slightly contaminated.

Munk and Czado (1998) introduced in this context a nonparametric procedure to assess similarity between distributions based on a trimmed version of the L_p-Wasserstein (or

†Research partially supported by the Spanish Ministerio de Ciencia y Tecnología and FEDER, grant BFM2005-04430-C02-01 and 02 and by the Consejería de Educación y Cultura de la Junta de Castilla y León, grant PAPIJCL VA102/06

Mallows) distance. If $\mathcal{F}_p$ is the set of probabilities on $\mathbb{R}$ with finite p-th moment, the L_p-Wasserstein distance between P and Q in $\mathcal{P}_p$ is defined as the lowest L_p-distance between random variables (r.v.'s), defined on any probability space, with distributions P and Q. Besides its intrinsic interest in connection to mass transportation problems, a main fact which makes this distance useful in statistics on the line is that it can be explicitely expressed in terms of quantile functions. If F and G are the distribution functions of P and Q and F^{-1} and G^{-1} are the respective quantile functions, then the L_p-Wasserstein distance between P and Q is given by (see, e.g., Bickel and Freedman (1981))

$$\mathcal{W}_p(P,Q) = \left[\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt\right]^{1/p} \tag{1}$$

(recall that F^{-1} is defined on $(0,1)$ by $F^{-1}(t) = \inf\{s : F(s) \geq t\}$ and satisfies that its distribution function is F when considered as a r.v. defined on the unit interval). From this it is obvious that for the probability measures based on two samples (resp. one sample and a theoretical distribution) $\mathcal{W}_p$ coincides with the L_p distance to the diagonal in a Q-Q plot (resp. probability plot). In the goodness-of-fit setting the large sample behavior of this distance, for $p = 2$, was analyzed in del Barrio, Cuesta-Albertos, Matrán, and Rodríguez-Rodríguez (1999) (see also del Barrio et al. (2000), Csörgö (2002), de Wet (2002) and del Barrio et al. (2005)), while for $p = 1$ the analysis was carried in del Barrio, Giné, and Matrán (1999).

Munk and Czado (1998) considered a trimmed version of the Wasserstein distance for assessment of similarity between the distribution functions F and G as

$$\Gamma_{\alpha,p}(F,G) := (1-2\alpha)^{-1}\left(\int_\alpha^{1-\alpha} |F^{-1}(t) - G^{-1}(t)|^p dt\right)^{1/p}. \tag{2}$$

This work has been continued in Czado and Munk (1998) and in Freitag et al. (2007). An interesting fact to be noted here is that the right hand side of the above expression equals $\mathcal{W}_p(P_\alpha, Q_\alpha)$, where P_α is the probability measure with distribution function

$$F_\alpha(t) = \frac{1}{1-2\alpha}\left(F(t) - \alpha\right), \quad F^{-1}(\alpha) \leq t < F^{-1}(1-\alpha) \tag{3}$$

and similarly for Q_α. In other words, the trimmed Wasserstein distance considered by Munk and Czado is nothing but the distance between the *trimmed distributions* P_α and Q_α (see our Definition 2.1 below). When comparing samples $\{x_1, ..., x_n\}$, $\{y_1, ..., y_m\}$, it corresponds to the distance between the sample distributions associated to the symmetrically trimmed samples (i.e. the samples obtained after removing the $[n\alpha]$ highest and the $[n\alpha]$ lowest values of both samples). This way of trimming is widely used and gives a doubtless protection against contamination by outliers. Though, the arbitrariness in the choice of the trimming zones has been largely reported as a serious drawback of procedures based on this method. In our setting the questionable fact would be why should two distributions largely different at their tails be considered similar but they should be considered as non-similar if they differ in their central parts?

To avoid this type of difficulty several estimators based on different trimming methods and general trimming techniques have been introduced in different statistical setups. Least Trimmed Squares, Minimum Volume Ellipsoids or Minimum Covariance Determinant estimators (see Rousseeuw (1985), Rousseeuw and Leroy (1987), or Maronna et al.

(2006) for references), as well as the impartial trimming methodology (see, e.g., Gordaliza (1991), Cuesta et al. (1997), García-Escudero et al. (2006) and Croux and Laine (2003) and Maronna (2005)), are based on the idea that the trimming zone should be determined by the data themselves. In our present setup we allow the trimming procedure to be chosen from the data by discarding from the sample points whith high contributions to the dissimilarity of the distributions.

To get a first idea about the differences between trimming procedures let us recall Example 1 in Munk and Czado (1998). It corresponds to a multiclinical study on cholesterol and fibrinogen levels in two sets of patients (of sizes 116 and 141) in two clinical centers. For the fibrinogen data, our impartial trimming proposal for $\alpha = 0.05$ essentially coincides with the symmetrical trimming. However, Figure 1 displays the effects of our trimming proposal for the cholesterol data, showing a significant trimming also in the middle part of the histograms corresponding to both centers. This even improves the level of similarity shown in Munk and Czado (1998), strengthening their assessment of similarity on these data.

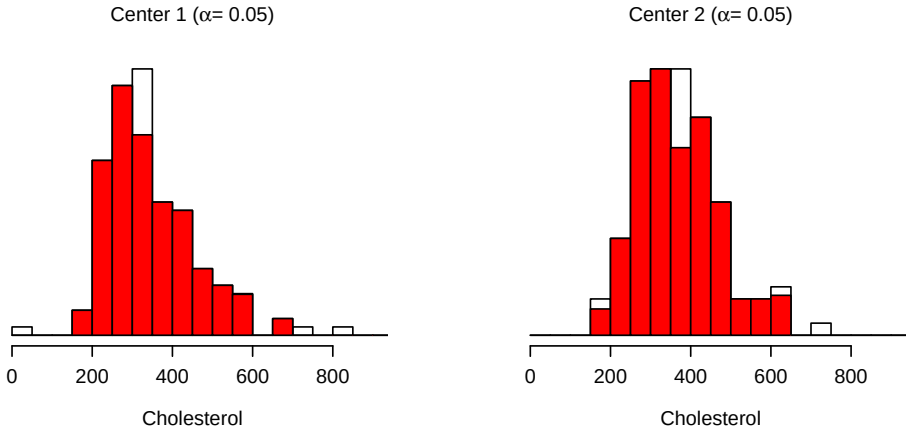

Fig. 1. Histograms corresponding to the trimmed data of cholesterol levels in two clinical centers. The white part in the bars shows the proportion of trimming in such zone giving maximum similarity to the remaining data of both centers.

The trimming method in this example is a natural extension of the impartial trimming methodology (Gordaliza (1991), Cuesta et al. (1997)) to this framework and will be introduced and analyzed in Section 2. However, we want to emphasize that in this paper we use impartial trimming not only as a way to robustify a statistical procedure but also as a method to discard a part of the data to achieve the best possible fit between two given samples or between a sample and a theoretical distribution, thus searching for the maximum similarity between them. In Section 3 we will compare our methodology with that of Munk and Czado on a real data set, showing the flexibility that the impartial trimming introduces in the similarity framework. The proofs of our results on the asymptotics of the involved statistics will be deferred to a final Appendix.

2. Measuring dissimilarities through impartial trimming.

Let $X_1, \ldots, X_n$ be i.i.d. observations with unknown common distribution F. In some instances the statistician wants to assess whether the data can be essentially assumed to follow a particular pattern, say G, except for minor distortions. In other instances a second i.i.d. sample, $Y_1, \ldots, Y_m$ with unknown common distribution G is available and we are interested in checking whether the two samples can be assumed to come from essentially equal random generators. This can be formally stated as a test of whether, at some fixed trimming level, α, both distributions, F and G, are close. To fix ideas, as in the main part of Munk and Czado (1998), we will measure closeness by the L_2-Wasserstein distance. Thus, we could consider

$$\inf_A \frac{1}{1-\alpha} \left(\int_A \left(F^{-1}(t) - G^{-1}(t) \right)^2 dt \right)^{1/2}, \tag{4}$$

where the set A varies on the Borel sets in $(0,1)$ with Lebesgue measure equal to $1-\alpha$ as the target parameter of our inferences. It is convenient, however, to introduce here a slightly more general concept: the trimming of a distribution. Trimmed probabilities can be defined in general probability spaces, although for practical purposes we will restrict ourselves to probabilities on the real line.

DEFINITION 2.1. *Let P be a probability measure on $\mathbb{R}$ and let $0 \leq \alpha < 1$. We say that a probability measure P^*, on $\mathbb{R}$, is an α-trimming of P if P^* is absolutely continuous with respect to P ($P^* \ll P$) and $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$.*

We will denote the set of α-trimmings of P by $\mathcal{T}^\alpha(P)$, namely, if $\mathcal{P}$ denotes the set of probability measures on $\mathbb{R}$, then

$$\mathcal{T}^\alpha(P) = \left\{ P^* \in \mathcal{P} : \quad P^* \ll P, \quad \frac{dP^*}{dP} \leq \frac{1}{1-\alpha} \quad P\text{-a.s.} \right\}. \tag{5}$$

The limit case in which $\alpha = 1$, $\mathcal{T}^1(P)$, is just the set of probability measures absolutely continuous with respect to P.

Obviously, trimmings in the Munk and Czado sense are included in this definition because $P_\alpha \ll P$ and

$$\frac{dP_\alpha}{dP} = \frac{1}{1-\alpha} I_{[F^{-1}(\alpha/2), F^{-1}(1-\alpha/2)]}.$$

Note that in the Munk and Czado trimmed distance (2) the effective trimming level is 2α and, accordingly, this effective trimming size was 0.1 in Figure 1.

An equivalent characterization is that $P^* \in \mathcal{T}^\alpha(P)$ if and only if $P^* \ll P$ and $\frac{dP^*}{dP} = \frac{1}{1-\alpha} f$ with $0 \leq f \leq 1$. If f takes only the values 0 and 1 then it is the indicator of a set, say A, such that $P(A) = 1-\alpha$ and trimming corresponds to considering the probability measure $P(\cdot|A)$. Definition (5) allows to reduce the weight of some regions of the measurable space without completely removing them from the feasible set.

The following proposition collects some elementary facts about trimmings.

PROPOSITION 2.1. *For any probability measure, P, on $\mathbb{R}$,*

(a) $\mathcal{T}^{\alpha_1}(P) \subset \mathcal{T}^{\alpha_2}(P)$ *if* $\alpha_1 \leq \alpha_2$.

(b) $\mathcal{T}^0(P) = \{P\}$.

(c) $\mathcal{T}^\alpha(P)$ *is a convex set.*

In the following proposition we employ the set $\mathcal{C}_\alpha$, the class of absolutely continuous functions $h : [0,1] \to [0,1]$ such that, $h(0) = 0$, $h(1) = 1$, with derivative h' such that $0 \le h' \le \frac{1}{1-\alpha}$. Compactness of this set in the $\|\cdot\|_\infty$ topology (see Lemma 3.2 in the Appendix) will be a key fact in some proofs later.

PROPOSITION 2.2. *For any real probability measure, P,*

(a) $\mathcal{T}^\alpha(P) = \{P^* \in \mathcal{P} : P^*(-\infty, t] = h\left(P(-\infty, t]\right), \quad h \in \mathcal{C}_\alpha\}$

(b) $\mathcal{T}^\alpha(U[0,1]) = \{P^* \in \mathcal{P} : P^*(-\infty, t] = h(t), 0 \le t \le 1, \quad h \in \mathcal{C}_\alpha\}$.

PROOF. Let $\mathcal{A} = \{P^* \in \mathcal{P} : P^*(-\infty, t] = h\left(P(-\infty, t]\right), \quad h \in \mathcal{C}_\alpha\}$. Given $P^* \in \mathcal{A}$, absolute continuity of h entails

$$P^*(s,t] = h\left(P(-\infty,t]\right) - h\left(P(-\infty,s]\right) = \int_{P(-\infty,s]}^{P(-\infty,t]} h'(x)\,dx \le \frac{1}{1-\alpha} P(s,t].$$

Hence, $P^* \ll P$ and $\frac{dP^*}{dP} \le \frac{1}{1-\alpha}$. Thus, $P^* \in \mathcal{T}^\alpha(P)$.

Conversely, given $P^* \in \mathcal{T}^\alpha(P)$, if F is the distribution function of P and we define $h(t) = \int_0^t \frac{dP^*}{dP}(F^{-1}(s))ds$, it is immediate that $h \in \mathcal{C}_\alpha$ and,

$$P^*(-\infty,t] = \int_{-\infty}^{t} \frac{dP^*}{dP}(s)dF(s) = \int_0^{F(t)} \frac{dP^*}{dP}(F^{-1}(s))ds = h\left(P(-\infty,t]\right).$$

Therefore $P^* \in \mathcal{A}$, and first part is proved. Part (b) is immediate from (a). □

Statement (b) in Proposition 2.2 says that the class $\mathcal{C}_\alpha$ is the class of all the distribution functions of α-trimmings of the $U[0,1]$ distribution. Then, (a) gives a characterization of the α-trimmings of every distribution in terms of the α-trimmings of the $U[0,1]$ distribution. It will be useful to write P_h for the probability measure with distribution function $h(P(-\infty,t])$. The set of α-trimmings of P can then be written $\mathcal{T}^\alpha(P) = \{P_h : h \in \mathcal{C}_\alpha\}$.

For a Borel set $A \subset (0,1)$ with Lebesgue measure $1-\alpha$ we can consider the function $h \in \mathcal{C}_\alpha$ defined by $h' = \frac{1}{1-\alpha} I_A$. It is clear then that our problem (4) will produce an upper bound for

$$\tau_\alpha(F,G) := \inf_{h \in \mathcal{C}_\alpha} \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt = \inf_{h \in \mathcal{C}_\alpha} \mathcal{W}_2^2(P_h, Q_h). \tag{6}$$

On the other hand the infimum in (6) is easily seen to be attained at function h_0 below (see (8)), associated to a set with Lebesgue measure $1-\alpha$, as it was considered in (4). This minimizer, h_0, is an *impartial α-trimming* for P and Q, and $h_0(F(x))$ and $h_0(G(x))$ are the distribution functions of the impartially α-trimmed probabilities.

We will use $(\tau_\alpha(F,G))^{1/2}$ as a measure of dissimilarity, and introduce a nonparametric test of similarity in an analogous way to that introduced by Munk and Czado. As usual in bioequivalence studies, the interest of the statisticians when analyzing similarity of distributions relies on asserting the equivalence of the involved probability distributions. In hypothesis testing this is achieved by taking equivalence or similarity as the alternative hypothesis, while dissimilarity is the null hypothesis. In agreement with this point of view,

Munk and Czado (1998) considered the testing problem with the null hypothesis being that the trimmed distance (2) exceeds some Δ-value, a threshold to be analyzed by the experimenters and the statisticians in an ad hoc way. Graphics on p-values for different Δ-values (see Figure 4 in Section 3) play a key role in this analysis, and the fact that $(\tau_\alpha(F,G))^{1/2}$ is measured in the same scale as the variable of interest favors this goal. As a final consideration regarding the use of $(\tau_\alpha(F,G))^{1/2}$ to detect similarity/dissimilarity, we must remark the fact just proved that it is the Wasserstein distance between trimmed versions of the original distributions. This allows to handle the very nice properties of this distance (see, e.g., Bickel and Freedman (1981)) in a friendly way in connection with our problem.

We will base our test of $H_0 : \tau_\alpha(F,G) > \Delta_0^2$ against $H_a : \tau_\alpha(F,G) \leq \Delta_0^2$ on the empirical counterparts of $\tau_\alpha(F,G)$, namely, $T_{n,\alpha} := \tau_\alpha(F_n,G)$, where F_n denotes the empirical d.f. based on the data, in the one sample problem and $T_{n,m,\alpha} := \tau_\alpha(F_n,G_m)$ in the two sample case. Our next results show that, under some mild assumptions on F and G, $T_{n,\alpha}$ and $T_{n,m,\alpha}$ are asymptotically normal, a fact that will be used later to approximate the critical values of H_0 against H_a. Before stating our results we introduce some notation.

We can consider the map $t :\rightarrow |F^{-1}(t) - G^{-1}(t)|$ as a random variable defined on $(0,1)$ endowed with the Lebesgue measure, ℓ. Let us denote by

$$L_{F,G}(x) := \ell\{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| \leq x\}, x \geq 0$$

its distribution function and write $L_{F,G}^{-1}$ for the corresponding (left continuous) quantile inverse. If $L_{F,G}$ is continuous at $L_{F,G}^{-1}(1-\alpha)$ then $L_{F,G}(L_{F,G}^{-1}(1-\alpha)) = 1-\alpha$, and

$$\inf_{h \in \mathcal{C}_\alpha} \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t) dt = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h_0'(t) dt, \tag{7}$$

where

$$h_0'(t) = \frac{1}{1-\alpha} I_{(|F^{-1}(t)-G^{-1}(t)| \leq L_{F,G}^{-1}(1-\alpha))} \tag{8}$$

in this case h_0 is in fact the *unique* mimimizer of the criterion functional. When $L_{F,G}$ is not continuous at $L_{F,G}^{-1}(1-\alpha)$), then, from the definition of the quantile function, we have

$$\begin{aligned} &\ell\{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| < L_{F,G}^{-1}(1-\alpha))\} \\ &\qquad \leq \quad 1-\alpha \leq \ell\{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| \leq L_{F,G}^{-1}(1-\alpha))\}. \end{aligned}$$

Thus we can also assure the existence of a set A_0 (although not necessarily unique) such that $\ell(A_0) = 1-\alpha$ and

$$\begin{aligned} &\{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| < L_{F,G}^{-1}(1-\alpha))\} \\ &\qquad \subset \quad A_0 \subset \{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| \leq L_{F,G}^{-1}(1-\alpha))\}. \end{aligned}$$

Obviously, if for any such A we consider the function I_A, then $\frac{1}{1-\alpha} I_A \in \mathcal{C}_\alpha$ and the infimum in (6) is attained at $\frac{1}{1-\alpha} I_A$. Therefore the problems (6) and (4) are equivalent.

Although the involved statistics are considerably more complex in our setting, the following results show an analogous behavior to that obtained for the symmetrically trimmed distributions in Munk and Czado (1998), leading to a similar analysis in the applications.

THEOREM 2.3. *Assume that F, $G \in \bar{\mathcal{P}}_4 = \cup_{s>4}\mathcal{P}_s$, $L_{F,G}$ is continuous at $L_{F,G}^{-1}(1-\alpha)$ and F has a continuously differentiable density $F' = f$ such that*

$$\sup_{x\in\mathbb{R}} \left| \frac{F(x)(1-F(x))f'(x)}{f^2(x)} \right| < \infty. \tag{9}$$

Then $\sqrt{n}(T_{n,\alpha} - \tau_\alpha(F,G))$ is asymptotically centered normal with variance

$$\sigma_\alpha^2(F,G) = 4\left(\int_0^1 l^2(t)dt - \left(\int_0^1 l(t)dt \right)^2 \right), \tag{10}$$

where

$$l(t) = \int_{F^{-1}(1/2)}^{F^{-1}(t)} (x - G^{-1}(F(x)))h_0'(F(x))dx,$$

and h_0 is given by (8).

This asymptotic variance can be consistently estimated by

$$s_{n,\alpha}^2(G) = \frac{4}{(1-\alpha)^2}\frac{1}{n}\sum_{i,j=1}^{n-1}(i\wedge j - \tfrac{ij}{n})a_{n,i}a_{n,j},$$

where $a_{n,i} = (X_{(i+1)} - X_{(i)})((X_{(i+1)} + X_{(i)})/2 - G^{-1}(i/n))I_{(|X_{(i)} - G^{-1}\left(\frac{i}{n}\right)| \le \ell_{F_n,G}^{-1}(1-\alpha))}$.

THEOREM 2.4. *Under the assumptions on Theorem 2.3, if G satisfies also (9) and $\frac{n}{n+m} \to \lambda \in (0,1)$ then $\sqrt{\frac{nm}{n+m}}(T_{n,m,\alpha} - \tau_\alpha(F,G))$ is asymptotically centered normal with variance $(1-\lambda)\sigma_\alpha^2(F,G) + \lambda\sigma_\alpha^2(G,F)$. This asymptotic variance can be consistently estimated by $s_{n,m,\alpha}^2 = \frac{m}{n+m}s_{n,\alpha}^2(G_m) + \frac{n}{n+m}s_{m,\alpha}^2(F_n)$.*

The proof of both statements is very similar. We will give in the Appendix only the proof of Theorem 2.3.

If $\tau_\alpha(F,G) = 0$ then Theorem 2.3 reduces to $\sqrt{n}T_{n,\alpha} \to 0$ in probability (observe that $\tau_\alpha(F,G) = 0$ implies $(x - G^{-1}(F(x)))^2 h_0'(F(x)) = 0$ for almost every x and, therefore, $\sigma_\alpha^2(F,G) = 0$). Although this would generally suffice for the applications (like those considered in Section 3), the following theorem gives the exact rate and the limiting distribution in that case. We will use the notation

$$\mathcal{C}_\alpha(F,G) = \left\{ h \in \mathcal{C}_\alpha : \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt = 0 \right\}.$$

Observe that $\mathcal{C}_\alpha(F,F) = \mathcal{C}_\alpha$, but for $F \neq G$ we have that $\mathcal{C}_\alpha(F,G)$ is a proper subset of $\mathcal{C}_\alpha$. Note also that $\mathcal{C}_\alpha(F,G) \neq \emptyset$ if and only if $\tau_\alpha(F,G) = 0$. In fact, the size of $\mathcal{C}_\alpha(F,G)$ depends on the Lebesgue measure of the set $\{t \in (0,1) : F^{-1}(t) \neq G^{-1}(t)\}$. $\tau_\alpha(F,G) = 0$ if and only if the Lebesgue measure of this last set is less than or equal to α; if it equals α then $\mathcal{C}_\alpha(F,G)$ consists of only one function, h, corresponding to $h'(t) = \frac{1}{1-\alpha}I_{(F^{-1}(t)=G^{-1}(t))}$. Lemma 3.3 in the Appendix proves that $\mathcal{C}_\alpha(F,G)$ is also a compact set for the $\|\cdot\|_\infty$ topology.

Now we are ready for the last result in this section, that establishes the asymptotic behavior of $nT_{n,\alpha}$ when F and G are equivalent at trimming level α.

Table 1. Two-sample p-values for classical tests.

Test	**p-value**	
	GPA by gender	GPA by major
Shapiro-Wilks (Sample 1)	0.0176	0.0360
Shapiro-Wilks (Sample 2)	0.0217	0.0001
Kolmogorov-Smirnov	0.0028	0.0040
Wilcoxon-Mann-Whitney	0.0004	0.0175

THEOREM 2.5. *If $\tau_\alpha(F,G)=0$, F satisfies (9) and*

$$\int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))}dt<\infty, \tag{11}$$

then

$$nT_{n,\alpha} \underset{w}{\to} \min_{h\in\mathcal{C}_\alpha(F,G)} \int_0^1 \frac{B(t)^2}{f^2(F^{-1}(t))}h'(t)\,dt,$$

where $\{B(t)\}_{0<t<1}$ is a Brownian bridge.

REMARK 2.6. Arguing as in the proof of Lemma 15 it can be seen that a.s.

$$h\mapsto \int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))}h'(t)dt$$

is $\|\cdot\|_\infty$-continuous as a function of h. Hence, it attains its minimum value on the compact set $\mathcal{C}_\alpha(F,G)$. This justifies the expression for the limiting distribution in Theorem 2.5.

3. Example.

Our analysis will be based on the variable GPA (College Grade Point Average) collected from a group of 234 students. This variable takes values from 0 to 4. The students are classified by the variables Gender and Major (1 = Computer Science, 2 = Engineering, 3 = Other Sciences). We are interested in studying the distributional similarity of the GPA obtained by males ($n=117$) and females ($m=117$), and between students with a major in computer sciences ($n=78$) and students with a major in engineering ($m=78$). Figure 2 shows the histogram for each sample.

Comparisons of these samples using classical procedures produce the results displayed in Table 1. The classical t-test is not appropriate as the Shapiro-Wilks tests reject the normality of the four samples. Then, the use of nonparametric methods like the Kolmogorov-Smirnov test (KS) or the Wilcoxon-Mann-Whitney test (WMW) is more appropriate to asses the null hypothesis that both samples come from the same distribution. The p-values of these tests reject clearly that the GPA of males and females is the same. The p-values of the KS and WMW tests (0.0040 and 0.0175, respectively) would lead us to the same conclusion in the other comparison (GPA by major).

We introduce the possibility of impartially trimming both samples as described in the previous section. Varying the trimmed proportion, α, we obtain the optimal trimming functions displayed in Figure 3. In this figure, and for each comparison, we plot the value of $|F_n^{-1}(t)-G_m^{-1}(t)|$ and the cutting values $L^{-1}_{F_n,G_m}(1-\alpha)$ for $\alpha=0.05$, 0.1 and 0.2. The first plot shows that the optimal trimming involves the lower tail, but not exactly from

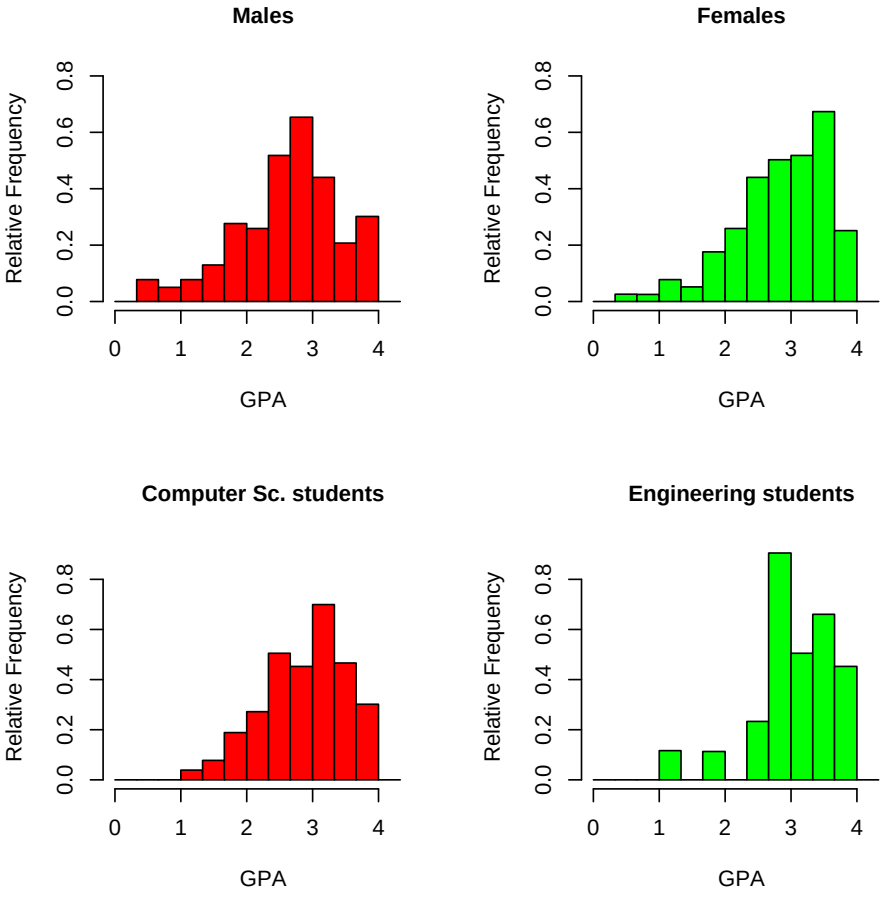

Fig. 2. Histograms for variable GPA

the lower end point. When the trimming level grows ($\alpha = 0.1$ and 0.2) the trimmed zone is not an interval and it includes points around percentiles 20%, 40%, 60% and 70%. For the second comparison, it is shown that the points that should be trimmed to make more similar both samples are between percentiles 10% and 30%. This example illustrates how the dissimilarity between two samples is not always found symmetrically in the tails of the distribution. Particularly, in the case of the first comparison the less similar zone is close to the lower tail, but not to the upper tail, where in fact, there are the more similar values.

3.1. p-values curve

To gain some insight into the assessment of the similarity or dissimilarity of the underlying distributions we can use the same p-values curves used in Munk and Czado (1998).

In order to test the null hypothesis $H_0 : \tau_\alpha(F, G) > \Delta_0^2$ against $H_a : \tau_\alpha(F, G) \leq \Delta_0^2$ in

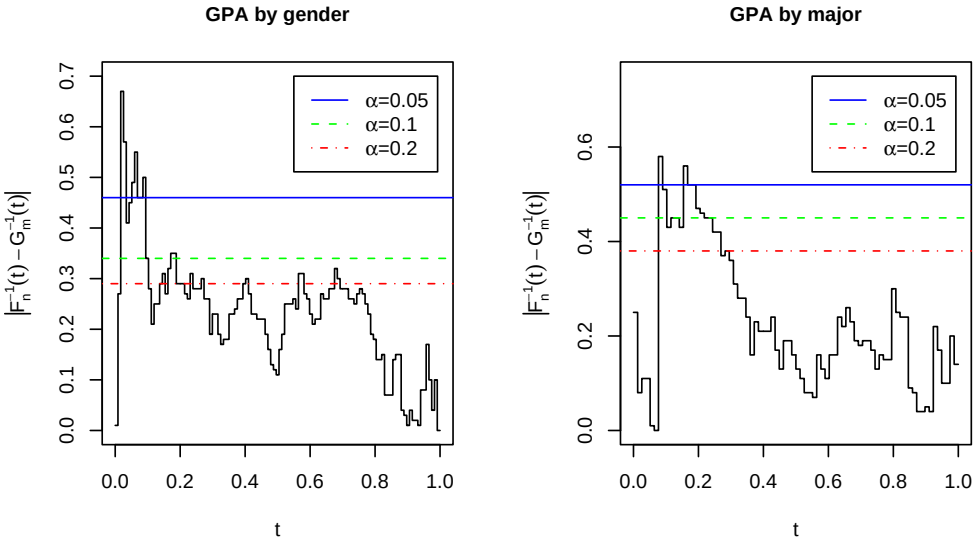

Fig. 3. Trimming functions

the two-sample comparison case, we will use the statistic

$$Z_{n,m,\alpha} = \sqrt{\frac{nm}{n+m}} \frac{(T_{n,m,\alpha} - \Delta_0^2)}{s_{n,m,\alpha}}.$$

The asymptotic p-value curve, $P(\Delta_0)$, is defined as follows,

$$P(\Delta_0) := \sup_{\{(F,G):(F,G)\in H_0\}} \lim_{n,m\to\infty} P_{F,G}\left(Z_{n,m,\alpha} \leq z_0\right) = \Phi\left(\sqrt{\frac{nm}{n+m}} \frac{T_{n,m,\alpha} - \Delta_0^2}{s_{n,m,\alpha}}\right),$$

where z_0 is the observed value of the test statistic $Z_{n,m,\alpha}$ for two given samples (note that the supremum is attained when the distance between both distributions is exactly Δ_0). These asymptotic p-value curves can be used in two ways. On one hand, given a fixed value of Δ_0 which controls the degree of dissimilarity, it is possible to find the p-value associated to the corresponding null hypothesis and then, to decide whether the distributions are similar or not. On the other hand, given a fixed test level (p-value), we can find the value of Δ_0 such that for every $\Delta \geq \Delta_0$ we should reject the hypothesis $H_0 : \tau_\alpha(F, G) \geq \Delta^2$. In other words, we can get a sound idea of the degree of dissimilarity between the distributions. To handle the values of Δ_0 the experimenter should take into account how to interpret the Wasserstein distance recalling that in the case that F and G belong to the same location family, their Wasserstein distance equals to the absolute difference of their locations.

Figure 4 displays the p-value curves using impartial trimming and symmetrical trimming for both comparisons for different trimming levels ($\alpha = 0.05$, 0.1 and 0.2). For each plot, a horizontal line has been drawn to mark a reference level for the test (0.05). The GPA points of males and females show similarity up to Δ_0 ranging from 0.32 to 0.36 (depending on the trimming size) when impartial trimmings are used. These values represent between

$100 \times 0.32/2.815 = 11.4\%$ and 12.8% of the average of the medians of the samples. However, when using symmetrical trimmings the horizontal line cuts the p-value curves for Δ_0 ranging from 0.56 to 0.59. This means between 20% and 21% of the average of the medians. A similar analysis in the comparison of the GPA value by major lead us to values of Δ_0 ranging from 0.29 to 0.36, which represent between a 9.6% and a 11.9% of the average of the medians when using impartial trimming. Instead, when using symmetrical trimming these percentages ranges from 16.6% to 19.5%.

This figure illustrates that when the dissimilarities are not in the tails of the distribution the assessment obtained using the Munk and Czado methodologies can be improved by the impartial trimmings. The computational analysis have been done with `R` statistical software

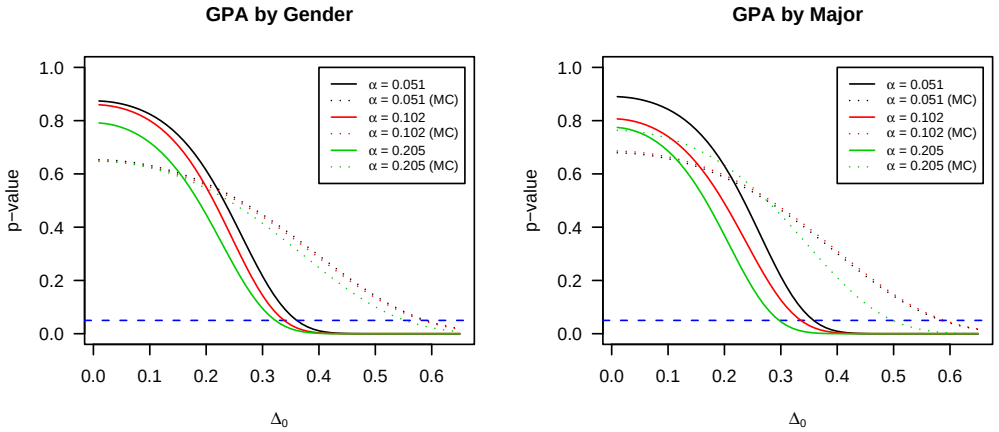

Fig. 4. p-values curves using impartial trimming and Munk & Czado (MC) methodologies.

(see R Development Core Team (2006)). The `R` programs and functions used to analyze the examples considered in this work are available at `http://www.eio.uva.es/~pedroc/R/`.

Appendix.

In this Appendix we write $\rho_n(t) = \sqrt{n} f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$ for the weighted quantile process, where f is the density function of F.

Proof (of Theorem 2.3). We can (and do) work in a sufficiently rich probability space in which there exist versions of $\{X_n\}_n$ and Brownian bridges B_n satisfying

$$n^{1/2-\nu} \sup_{\frac{1}{n} \le t \le 1-\frac{1}{n}} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} = \begin{cases} O_P(\log n), & \text{if } \nu = 0 \\ O_P(1), & \text{if } 0 < \nu \le 1/2 \end{cases}. \tag{12}$$

The existence of such a probability space is a consequence of (9), see, for instance, Theorem 6.2.1 in Csörgö and Horváth (1993).

Now we set $M_n(h) = \sqrt{n} \int_0^1 (F_n^{-1}(t) - G^{-1}(t))^2 h'(t) dt$ and

$$N_n(h) = 2 \int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{B_n(t)}{f(F^{-1}(t))} (G^{-1}(t) - F^{-1}(t)) h'(t) dt + \sqrt{n} \int_{\frac{1}{n}}^{1-\frac{1}{n}} (G^{-1}(t) - F^{-1}(t))^2 h'(t) dt.$$

Observe that

$$\begin{aligned}\sup_{h\in\mathcal{C}_\alpha}|M_n(h)-N_n(h)| \le \sqrt{n}\int_0^{\frac{1}{n}}(F_n^{-1}(t)-G^{-1}(t))^2dt+\sqrt{n}\int_{1-\frac{1}{n}}^1(F_n^{-1}(t)-G^{-1}(t))^2dt\\ +\frac{1}{\sqrt{n}}\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{|\rho_n(t)-B_n(t)|^2}{f^2(F^{-1}(t))}dt+\frac{1}{\sqrt{n}}\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{B_n(t)^2}{f^2(F^{-1}(t))}dt\\ +2\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{|\rho_n(t)-B_n(t)|}{f(F^{-1}(t))}|G^{-1}(t)-F^{-1}(t)|dt\\ =: A_{n,1}+A_{n,2}+A_{n,3}+A_{n,4}+A_{n,5}.\end{aligned}$$

The fact that $F,G\in\bar{\mathcal{P}}_4$ and Lemma 3.1 below imply $A_{n,1}\to 0$ and $A_{n,2}\to 0$ in probability. From (12) we get

$$A_{n,3}\le O_P(1)\frac{1}{\sqrt{n}}\int_{1/n}^{1-1/n}\frac{t(1-t)}{f^2(F^{-1}(t))}dt$$

and the last integral tends to 0 by Lemma 3.1. Hence $A_{n,3}\to 0$ in probability. Similarly, $A_{n,4}\to 0$ in probability. Finally, (12) yields

$$A_{n,5}\le O_P(1)n^{\nu-1/2}\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{(t(1-t))^\nu}{f(F^{-1}(t))}|G^{-1}(t)-F^{-1}(t)|dt$$

for some $\nu\in(0,1/2)$. Lemma 3.1 shows that $\int_0^1\frac{(t(1-t))^{1/2}}{f(F^{-1}(t))}|G^{-1}(t)-F^{-1}(t)|dt<\infty$. This and the dominated convergence theorem imply that the right-hand side of the last display tends to 0 in probability. Collecting the above estimates we see that $\sup_{h\in\mathcal{C}_\alpha}|M_n(h)-N_n(h)|\to 0$ in probability and, consequently, $\sqrt{n}(T_{n,\alpha}-S_{n,\alpha})\to 0$ in probability, where $\sqrt{n}S_{n,\alpha}=\inf_{h\in\mathcal{C}_\alpha}N_n(h)$. Thus, the proof will be complete if we show that $\sqrt{n}(\tilde{S}_{n,\alpha}-\tau_\alpha(F,G))$ is asymptotically $N(0,\sigma_\alpha^2(F,G))$, where

$$\sqrt{n}\tilde{S}_{n,\alpha}=\inf_{h\in\mathcal{C}_\alpha}\left[2\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}h'(t)dt+\sqrt{n}\int_0^1(G^{-1}(t)-F^{-1}(t))^2h'(t)dt\right]. \tag{13}$$

Let us denote

$$h_n=\mathrm{argmin}_{h\in\mathcal{C}_\alpha}\int_0^1(F^{-1}(t)-G^{-1}(t))^2h'(t)dt+\frac{2}{\sqrt{n}}\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}h'(t)dt.$$

Clearly $h_n'(t)\to h_0'(t)$ for almost every t. Furthermore, optimality of h_n shows $B_n\le 0$, where,

$$B_n:=\sqrt{n}\tilde{S}_{n,\alpha}-\left(2\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}h_0'(t)dt+\sqrt{n}\int_0^1(G^{-1}(t)-F^{-1}(t))^2h_0'(t)dt\right),$$

but, on the other hand,

$$\begin{aligned}B_n &= \sqrt{n}\left(\int_0^1(F^{-1}(t)-G^{-1}(t))^2h_n'(t)dt-\int_0^1(F^{-1}(t)-G^{-1}(t))^2h_0'(t)dt\right)\\ &\quad +2\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}(h_n'(t)-h_0'(t))dt=: B_{n,1}+B_{n,2}\end{aligned}$$

and $B_{n,1} \geq 0$ by optimality of h_0, while $B_{n,2} = o_P(1)$ by the dominated convergence theorem. Therefore, $B_n \to 0$ in probability, which shows that

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(F,G)) \to_w 2\int_0^1 B(t)\frac{G^{-1}(t) - F^{-1}(t)}{f(F^{-1}(t))}h_0'(t)dt. \tag{14}$$

Integrating by parts we obtain

$$\int_0^1 B(t)\frac{G^{-1}(t) - F^{-1}(t)}{f(F^{-1}(t))}h_0'(t)dt = -\int_0^1 l(t)dB(t)$$

and this proves the asymptotic normality and the expression (10) for the variance. The claim about the variance estimator follows by noting, after some algebra, that

$$s_{n,\alpha}^2 = 4\left(\int_0^1 l_n^2(t)dt - \left(\int_0^1 l_n(t)dt\right)^2\right),$$

where $l_n(t) = \int_{F_n^{-1}(1/2)}^{F_n^{-1}(t)}(x - G^{-1}(F_n(x)))h_n'(F_n(x))dx$ and $h_n(t) = \operatorname{argmin}_{h\in\mathcal{C}_\alpha}\int(F_n^{-1} - G^{-1})h'$. It can be shown that, with probability 1, $l_n(t) \to l(t)$ for almost every $t \in (0,1)$. A standard uniform integrability argument completes the proof. □

LEMMA 3.1. *If $F, G \in \bar{\mathcal{P}}_4$ then*

(i) $\sqrt{n}\int_0^{1/n}(F^{-1}(t))^2dt \to 0$; $\sqrt{n}\int_{1-1/n}^1(F^{-1}(t))^2dt \to 0$.

(ii) $\sqrt{n}\int_0^{1/n}(F_n^{-1}(t))^2dt \to 0$; $\sqrt{n}\int_{1-1/n}^1(F_n^{-1}(t))^2dt \to 0$.

(iii) $\int_0^1 \frac{\sqrt{t(1-t)}}{g(G^{-1}(t))}|F^{-1}(t) - G^{-1}(t)|dt < \infty$.

Further, if G satisfies (9)*, then*

(iv) $\frac{1}{\sqrt{n}}\int_{1/n}^{1-1/n}\frac{t(1-t)}{g^2(G^{-1}(t))}dt \to 0$.

PROOF. *(i)* For the first integral Schwarz's inequality gives

$$\sqrt{n}\int_0^{\frac{1}{n}}(F^{-1}(t))^2dt \leq \sqrt{n}\left(\int_0^{\frac{1}{n}}(F^{-1}(t))^4dt\right)^{1/2}\left(\int_0^{\frac{1}{n}}1dt\right)^{1/2} = \left(\int_0^{\frac{1}{n}}(F^{-1}(t))^4dt\right)^{1/2} \to 0;$$

the second convergence is completely similar.

(ii) We consider now the second expression. We can assume w.l.o.g. that G is concentrated on the positive real line. We have to show that $n^{-1/4}\max_{1\leq i\leq n}X_i \to 0$ in probability, or, equivalently, that $G(\varepsilon n^{1/4})^n \to 1$ for all $\varepsilon > 0$. Taking logarithms we see, using that G has finite fourth moment,

$$n\log(G(\varepsilon n^{1/4})) \simeq n(1 - G(\varepsilon n^{1/4})) \to 0.$$

(iii) We assume again that G is concentrated on the positive real line. A change of variable shows that it suffices to prove that $\int_0^\infty \sqrt{1-G(y)}|F^{-1}(G(y))|dy$ is finite. Fix $r > 4$ such

that F, G have finite r-th moment. Then $\lim_{t\to 1}(1-t)|F^{-1}(t)|^r = 0$. Hence, for large y we have $|F^{-1}(G(y))| \leq (1-G(y))^{-1/r}$ and, consequently, $\sqrt{1-G(y)}|F^{-1}(G(y))| \leq (1-G(y))^{(r-2)/2r}$. Denote by μ_r the r-th moment of G. Then, by Markov's inequality $(1-G(y))^{(r-2)/2r} \leq \mu_r^r y^{-(r-2)/2}$. This proves *(iii)* since $(r-2)/2 > 1$.

(iv) We assume for simplicity that G has support $(0,\infty)$. With the change of variable $y = G^{-1}(t)$ we can reduce the proof to showing that $\sqrt{1-G(x)}\int_0^x \frac{(1-G(y))}{g(y)}dy \to 0$ as $x \to \infty$. Observe now that $\frac{(1-G(y))}{g(y)}$ has derivative $-1 - \frac{(1-G(y))g'(y)}{g(y)}$ which, by (9), is uniformly bounded. Hence,

$$\limsup_{x\to\infty} \sqrt{1-G(x)} \int_0^x \frac{(1-G(y))}{g^2(y)}dy \to 0 \leq K \limsup_{x\to\infty} \sqrt{1-G(x)}x^2 = 0,$$

since G has finite fourth moment. □

LEMMA 3.2. *The set $\mathcal{C}_\alpha$ of all absolutely continuous functions $h : [0,1] \to [0,1]$ such that, $h(0) = 0$, $h(1) = 1$, with derivative h' such that $0 \leq h' \leq \frac{1}{1-\alpha}$ is compact for the $\|\cdot\|_\infty$ topology.*

PROOF. The set $\mathcal{C}_\alpha$ is uniformly bounded at 0 ($h(0) = 0$ for every $h \in \mathcal{C}_\alpha$) and uniformly equicontinuous ($|h(y) - h(x)| \leq \frac{1}{1-\alpha}|y-x|$ for every $h \in \mathcal{C}_\alpha$). Hence, by the Arzelá-Ascoli Theorem, $\mathcal{C}_\alpha$ is relatively compact for $\|\cdot\|_\infty$ and it suffices to show that $\mathcal{C}_\alpha$ is closed. Let us assume then that $\{h_n\}_n$ are such that $h_n \in \mathcal{C}_\alpha$ and $\|h_n - h\|_\infty \to 0$. Then

$$0 \leq h(y) - h(x) = \lim_{n\to\infty} h_n(y) - h_n(x) \leq \frac{1}{1-\alpha}(y-x), \quad \text{if } 0 \leq x \leq y \leq 1.$$

This implies that h is absolutely continuous and $0 \leq h' \leq \frac{1}{1-\alpha}$ almost everywhere. Therefore $h \in \mathcal{C}_\alpha$, which completes the proof. □

LEMMA 3.3. *If $F, G \in \mathcal{P}_2$ then the set*

$$\mathcal{C}_\alpha(F,G) = \left\{ h \in \mathcal{C}_\alpha : \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt = 0 \right\}$$

is compact for the $\|\cdot\|_\infty$ topology.

PROOF. It suffices to show that $\mathcal{C}_\alpha(F,G)$ is closed, since $\mathcal{C}_\alpha(F,G) \subset \mathcal{C}_\alpha$ and $\mathcal{C}_\alpha$ is compact by Lemma 3.2. This can be reduced to showing that

$$\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h_n'(t)dt \to \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt$$

whenever $h_n \in \mathcal{C}_\alpha$ and $\|h_n - h\|_\infty \to 0$ or, equivalently, that

$$\int_0^1 (F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 dy \to \int_0^1 (F^{-1}(h^{-1}(y)) - G^{-1}(h^{-1}(y))^2 dy. \tag{15}$$

By continuity of F^{-1} and G^{-1} (except, perhaps, at a countable set) we have $(F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 \to (F^{-1}(h^{-1}(y)) - G^{-1}(h^{-1}(y))^2$ at almost every $t \in (0,1)$. To prove (15)

it only remains to show uniform integrability of $(F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2$. But this follows from the next inequality.

$$\begin{aligned}\sup_n \int_{\{(F^{-1}(h_n^{-1}(y))-G^{-1}(h_n^{-1}(y))^2>x\}} & (F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 dy \\ &= \sup_n \int_{\{(F^{-1}(t)-G^{-1}(t))^2>x\}} (F^{-1}(t) - G^{-1}(t))^2 h_n'(t) dt \\ &\leq \frac{1}{1-\alpha} \int_{\{(F^{-1}(t)-G^{-1}(t))^2>x\}} (F^{-1}(t) - G^{-1}(t))^2 dt.\end{aligned}$$

□

Proof (of Theorem 2.5). We define $D_n(h) := n\int_0^1 (F_n^{-1}(t) - G^{-1}(t))^2 h'(t)\, dt$ and $D(h) := \int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))} h'(t) dt$ for $h \in \mathcal{C}_\alpha$. Note that

$$\begin{aligned}D_n(h) &= \int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} h'(t)\, dt + n\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)\, dt \\ &\quad +2\sqrt{n}\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))}(F^{-1}(t) - G^{-1}(t)) h'(t) dt.\end{aligned}$$

Observe also that $nT_{n,\alpha} = D_n(h_n)$ for some $h_n \in \mathcal{C}_\alpha$ If $h \in \mathcal{C}_\alpha(F,G)$ then the second and third summands in the right hand side vanish and $D_n(h) = \int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} h'(t)\, dt$. By (9) and (11) we have weak convergence of $\rho_n(\cdot)/f(F^{-1}(\cdot))$ to $B(\cdot)/f(F^{-1}(\cdot))$ as random elements in $L_2(0,1)$, see, e.g., Theorem 4.6 in del Barrio et al. (2005). By Skorohod's representation Theorem (see, e.g., Theorem 11.7.1 in Dudley (1989)) there are versions of $\rho_n(\cdot)/f(F^{-1}(\cdot))$ and $B(\cdot)/f(F^{-1}(\cdot))$ (for which we keep the same notation) such that $\|\rho_n(\cdot)/f(F^{-1}(\cdot)) - B(\cdot)/f(F^{-1}(\cdot))\|_2 \to 0$ a.s. Now for this versions we have

$$\sup_{h\in\mathcal{C}_\alpha(F,G)} |D_n(h) - D(h)| \leq \frac{1}{1-\alpha}\int_0^1 \left|\frac{\rho_n^2(t)}{f^2(F^{-1}(t))} - \frac{B^2(t)}{f^2(F^{-1}(t))}\right| dt \to 0 \quad \text{a.s.},$$

while for $h_0 \in \mathcal{C}_\alpha - \mathcal{C}_\alpha(F,G)$ we have a.s. that $D_n(h) \to \infty$ uniformly in a sufficiently small neighbourhood of h_0. Furthermore, if $h_n \to h \in \mathcal{C}_\alpha(F,G)$ then we can extract a subsequence such that $n\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h_n'(t) dt \to 0$. The result follows from the next technical Lemma. □

Lemma 3.4. *Let (X,d) be a compact metric space, $A \subset X$ compact and $\{f_n\}$, f real valued, continuous functions on X such that*

(i) $\sup_{x\in A} |f_n(x) - f(x)| \to 0$, *as* $n \to \infty$,

(ii) for $x \in X - A$ *there exists* $\varepsilon_x > 0$ *such that* $\inf_{d(y,x)<\varepsilon_x} f_n(y) \to \infty$, *as* $n \to \infty$,

(iii) if $x_n \to x \in A$ *there exists a subsequence,* $\{x_m\}$, *such that* $f_m(x_m) \to f(x)$.

Then

$$\min_{x\in X} f_n(x) \to \min_{x\in A} f(x).$$

PROOF. Choose x_n such that $f_n(x_n) = \min_{x\in X} f_n(x)$. Then there exists a converging subsequence $x_m \to x_0 \in X$. If $x_0 \in X - A$ then we fix $\varepsilon > 0$ such that $\inf_{d(y,x_0)<\varepsilon} f_n(y) \to \infty$. Since $x_m \to x_0$ and $\min_{x\in A} f_n(x) \to \min_{x\in A} f(x)$ we have that $f_m(x_m) > 2\min_{x\in A} f_m(x)$ for sufficiently large m, which contradicts the choice of x_m. Hence, $x_0 \in A$. Now, taking a further subsequence (that we keep denoting x_m we have that $f_m(x_m) \to f(x_0)$. Thus,

$$\min_{x\in A} f(x) \le f(x_0) = \lim_m f_m(x_m) = \lim_m \min_{x\in X} f_m(x) \le \lim_m \min_{x\in A} f_m(x) = \min_{x\in A} f(x)$$

and all inequalities above are, in fact, equalities. This completes the proof. □

Acknowledgments.

The data sets corresponding to the Multiclinical study on the Fibrinogen and Cholesterol levels were kindly provided by Axel Munk and Claudia Czado.

Data used in Section 3 can be found as the *majors.dat* file in the examples datasets of many statistics packages. We obtained them from the textbook by Moore and McCabe Moore and McCabe (2003).

References

Bickel, P. and D. Freedman (1981). Some asymptotic theory for the bootstrap. *Ann. Statist. 9*, 1196–1217.

Croux, C. and B. Laine (2003). Optimal subspace estimation based on trimmed square loss. Unpublished manuscript.

Csörgö, M. and L. Horváth (1993). *Weighted Approximations in Probability and Statistics*. New York: Wiley.

Csörgö, S. (2002). Weighted correlation tests for scale families. *Test 11*, 219–248.

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

Czado, C. and A. Munk (1998). Assessing the similarity of distributions—finite sample performance of the empirical Mallows distance. *J. Statist. Comput. Simulation 60*, 319–346.

de Wet, T. (2002). Goodness-of-fit tests for location and scale families based on a weighted l_2 weighted distance measure. *Test 11*, 89–107.

del Barrio, E., J. Cuesta-Albertos, and C. Matrán (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test 9*, 1–96.

del Barrio, E., J. Cuesta-Albertos, C. Matrán, and J. Rodríguez-Rodríguez (1999). Tests of goodness of fit based on the l2-wasserstein distance. *Ann. Statist. 27*, 1230–1239.

del Barrio, E., E. Giné, and C. Matrán (1999). Central limit theorem for the wasserstein distance between the empirical and the true distribution. *Ann. Probab. 27*, 1009–1071.

del Barrio, E., E. Giné, and F. Utzet (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.

Dudley, R. M. (1989). *Real Analysis and Probability*. Wadsworth & Brook/Cole.

Freitag, G., C. Czado, and A. Munk (2007). A nonparametric test for similarity of marginals - with applications to the assessment of population bioequivalence. *J. Statist. Plann. Inf. 137*, 697–711.

García-Escudero, L., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2006). A general trimming approach to robust cluster analysis. Submitted manuscript.

Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J. Approx. Theor. 64*, 162–180.

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

Maronna, R., D. Martin, and V. Yohai (2006). *Robust Statistics: Theory and Methods*. New York: Wiley.

Moore, D. S. and G. P. McCabe (2003). *Introduction to the Practice of Statistics, fourth edition*. W.H. Freeman and Company.

Munk, A. and C. Czado (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. Roy. Statist. Soc. Ser. B 60*, 223–241.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Werz (Eds.), in *Mathematical Statistics and Applications, Volume B*. Reidel, Dordrecht.

Rousseeuw, P. and A. Leroy (1987). *Robust Regression and Outlier Detection*. New York: Wiley.