Assessing when a sample is mostly normal

Pedro C. Alvarez-Esteban [a,*] Eustasio del Barrio [a]
Juan A. Cuesta-Albertos [b] Carlos Matrán [a]

[a] *Dept. de Estadística e Investigación Operativa, Universidad de Valladolid. Prado de la Magdalena s.n., 47005 Valladolid. Spain.*

[b] *Dept. Matemáticas, Estadística y Computación, Universidad de Cantabria. Avda. los Castros s.n. 39005 Santander, Spain.*

Abstract

The use of trimming procedures constitutes a natural approach to robustifying statistical methods. This is the case of goodness-of-fit tests based on a distance, which can be modified by choosing trimmed versions of the distributions minimizing that distance. In this paper we consider the L_2-Wasserstein distance and introduce the trimming methodology for assessing when a data sample can be considered mostly normal. The method can be extended to other location and scale models, introducing a robust approach to model validation, and allows an additional descriptive analysis by determining the subset of the data with the best improved fit to the model. This is a consequence of our use of data-driven trimming methods instead of more classical symmetric trimming procedures.

Key words: Model Assessment, Asymptotics, Impartial Trimming, Wasserstein distance, Similarity.

1 Introduction.

Trimming methods are a main tool in the design of robust statistical procedures. For univariate data a classical way of trimming is based on deleting the same proportion of observations in each tail of the distribution. This approach

* *Email address:* pedroc@eio.uva.es (Pedro C. Alvarez-Esteban).

[1] Research partially supported by the Spanish Ministerio de Educación y Ciencia, grant MTM2008-06067-C02-01, and 02 and by the Consejería de Educación y Cultura de la Junta de Castilla y León, GR150.

has some drawbacks. First, the implicit assumption that the possible contamination is only due to outliers. Second, the lack of "a priori" directions to trim in the multivariate setting. Several alternatives to the symmetric trimming have been proposed in the statistical literature. Among the proposed alternatives to overcome these difficulties we focus on those minimizing some distance criterium, leading to the "impartial" trimming introduced by Rousseeuw (1985) and in greater generality in Gordaliza (1991). This impartial trimming methodology is based on the idea that the trimming zone should be determined by the data themselves and has been successfully applied to different statistical problems including location estimation, (Rousseeuw, 1985; Gordaliza, 1991), regression problems (Rousseeuw, 1985), cluster analysis (Cuesta-Albertos et al., 1997; García-Escudero et al., 2003, 2008), and principal component analysis (Maronna, 2005).

This approach looks very appropriate for the goodness-of-fit framework, where the procedures are often based on minimizing distances. However, only some timid attempts have been reported in this sense so far. In fact, to our best knowledge, the only related approach is that of Munk and Czado (1998), where a symmetric trimming is introduced to robustify an analysis of similarity based on the Wasserstein distance. In our setting, the questionable fact about this approach, would be why should two distributions largely different at their tails be considered similar but they should be considered as non-similar if they are slightly different in their central parts?

This observation led to a new proposal in Alvarez-Esteban et al. (2008a), where similarity of distributions is assessed on the basis of the comparison of their trimmed versions. The approach was based on considering that two distributions are similar at level α whenever suitable chosen α-trimmed versions of such distributions coincide. This key idea is naturally related to Robustness, and can be combined with the use of a distance between probabilities to measure their degree of dissimilarity. The L_2-Wasserstein distance was the choice in Alvarez-Esteban et al. (2008a) to introduce a nonparametric test of similarity that can be considered as a robust version of a goodness-of-fit test to a completely specified distribution or, rather, a way to assess whether the *core* of the distribution underlying the data fits a fixed distribution.

In this work we show how these ideas can be used to assess whether the *core* of the distribution underlying the data can be assumed to follow a given location-scale model. For the sake of simplicity and its relevance, we consider the normal model, but it will become apparent that the methodology can be extended to cover other patterns. More precisely, we measure the minimal distance between trimmed versions of the empirical distribution and trimmed normal distributions and provide the necessary distributional theory to make it usable for inferences about its population counterpart. Our procedure involves the computation of a best trimming and it can be considered 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 to normality of the remaining data. Thus, this kind of robustification provides an added value as a descriptive tool for the analysis of the data.

On the real line, the L_2-Wasserstein distance between two probability measures can be obtained as the L_2 distance between their quantile functions, so it has an easy interpretation in terms of probability plots. In the particular case of testing for normality its use leads to a version of the omnibus Shapiro-Wilks test (see e.g. del Barrio et al., 1999, 2000, 2005). The L_2-Wasserstein distance is also well behaved with respect to trimmings (see Alvarez-Esteban et al. (2008a)) and will also be our choice here to introduce a robust approach to model validation.

This paper is organized as follows. In Section 2 we give the necessary background on trimmed distributions and Wasserstein distance and use it to introduce an estimator for the *trimmed distance to normality*. We show how to use it to assess whether a sufficiently large fraction of the distribution underlying the data can be assumed to be normal. We include in this section some asymptotic results that justify our approach. We describe the algorithm involved in the computation of our estimators and discuss further implementation details. In Section 3 we provide empirical evidence of the performance of our proposal. This will be made through real and simulated examples giving support to the procedure. Finally, an Appendix is devoted to the proof of the results.

2 Trimmed distributions in testing for normality.

2.1 Trimmed distance to normality.

Trimmed probabilities can be defined in general spaces, but for the application presented in this paper we will restrict to the real line. Let P be a probability on $\mathbb{R}$ and $0 \leq \alpha < 1$, we say that a probability P^* is an α-trimming of P if P^* is absolutely continuous with respect to P and $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$. We will denote by $\mathcal{T}_\alpha(P)$ the set of α-trimmings of P,

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

An equivalent characterization, useful to gain some insight about the meaning of an α-trimming is that $P^* \in \mathcal{T}_\alpha(P)$ if $P^* \ll P$ and there exists a function f such that $\frac{dP^*}{dP} = \frac{1}{1-\alpha} f$ where $0 \leq f \leq 1$ P-a.s. Here, $f(x)$ gives the fraction of density not trimmed at a point x in the support of P. If $f(x) = 0$, the point x is completely removed, while if $f(x) = 1$ there is no trimming at x. For those

points in the support of P where $0 < f < 1$ their weight after trimming is decreased. Note that this is a natural generalization of the common practice of trimming observations, which amounts to replacing the empirical distribution by a new version with new weights on the data: 0 for the points removed and $1/(n(1-\alpha))$ for the points kept in the sample (if we remove $k = n\alpha$ observations).

Interesting properties of α-trimmings can be found in Alvarez-Esteban et al. (2008a,b). We mention here one which is essential for the proposal in this paper. General α-trimmings can be parametrized in terms of the α-trimmings of the uniform distribution on $(0,1)$. More precisely, if $\mathcal{C}_\alpha$ is 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 $0 \leq h' \leq \frac{1}{1-\alpha}$ and we write P_h for the probability with distribution function $h(P(-\infty, t])$, then, for any real probability measure, P, we have,

$$\mathcal{T}_\alpha(P) = \{P_h : h \in \mathcal{C}_\alpha\}.$$

We note that $\mathcal{C}_\alpha$ is the set of distribution functions of α-trimmings of the uniform distributions on $(0,1)$. As a consequence, if P has distribution function F and quantile function F^{-1}, the set of α-trimmings of P equals the set of probability measures with quantile functions of type $F^{-1}(h^{-1}(t))$, $t \in (0,1)$ with $h \in \mathcal{C}_\alpha$.

Let $\mathcal{F}_2$ be the set of univariate distributions with finite second moments. Take $P, Q \in \mathcal{F}_2$ with quantile functions F^{-1}, G^{-1}, respectively. The L_2-Wasserstein distance between these two distributions is defined as

$$\mathcal{W}_2(P,Q) := \inf\left\{\left(E(X-Y)^2\right)^{1/2} : \mathcal{L}(X) = P,\, \mathcal{L}(Y) = Q\right\},$$

where X and Y are random variables defined on some arbitrary probability space. The fact that $\mathcal{W}_2$ metrizes weak convergence of probability measures plus convergence of moments of order two (see Bickel and Freedman (1981)), makes this distance specially convenient for statistical purposes. Moreover, on the real line, it equals the L_2-distance between the quantile functions, namely,

$$\mathcal{W}_2(P,Q) = \left[\int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt\right]^{1/2}.$$

It could be the case, when trying to assess normality of a data sample, $X_1, \ldots, X_n$, that some significant deviation is found but, in fact, this deviation is caused only by some small fraction of the data. It seems natural to remove or downplay the importance of the disturbing range of observations and measure the distance between the trimmed versions of the empirical measure and the normal distributions, with a trimming pattern chosen in order to

optimize fit. If we measure distance by $\mathcal{W}_2$ this amounts to considering

$$T_{n,\alpha} := \inf_{h \in \mathcal{C}_\alpha, Q \in \mathcal{N}} \mathcal{W}_2^2((P_n)_h, Q_h), \tag{1}$$

where P_n denotes the empirical distribution and $\mathcal{N}$ stands for the family of normal distributions on the line. We assume that $X_1, \ldots, X_n$ are i.i.d. observations with common distribution P. The population version of (1),

$$\tau_\alpha(P, \mathcal{N}) := \inf_{h \in \mathcal{C}_\alpha, Q \in \mathcal{N}} \mathcal{W}_2^2(P_h, Q_h), \tag{2}$$

measures how far from normality is the *core* of the underlying distribution P. We refer to $\tau_\alpha(P, \mathcal{N})$ as the (squared) *trimmed distance to normality*. If $\tau_\alpha(P, \mathcal{N}) = 0$ then there is some normal distribution Q which is equal to P after removing a fraction of mass, of size at most α, on P and Q. A small value of $\tau_\alpha(P, \mathcal{N})$ indicates that most of the distribution underlying the data is not far from normality, which might be enough for the validity of some inferences. Assessment of this small deviation from normality means, in more formal terms, fixing a threshold Δ_0^2 and testing

$$H_0 : \tau_\alpha(P, \mathcal{N}) \geq \Delta_0^2 \quad \text{vs.} \quad \tau_\alpha(P, \mathcal{N}) < \Delta_0^2. \tag{3}$$

Given the sample $X_1, \ldots, X_n$ we compute $T_{n,\alpha}$, defined in (1), and reject H_0 for small values of it.

Note that the choice of the null and the alternative hypotheses is in agreement to the fact that, as in other goodness-of-fit problems, the consequences of assuming (approximate) normality when it is not true are worse than those of the other possible error. Thus, rejecting H_0 can be done at a controlled error rate. We refer to Munk and Czado (1998) for further discussion on this issue. The difficulty posed by the arbitrary choice of the threshold Δ_0^2 can be dealt with by the consideration of the p-value curve, as in Munk and Czado (1998) or Alvarez-Esteban et al. (2008a). We turn to this point in Subsection 2.3.

It can be easily checked that $\tau_\alpha(P, \mathcal{N})$ is location invariant, but not scale invariant. In order to avoid this dependence on scale, and as usual in assessing fit to location-scale models (see, e.g., del Barrio et al. (1999)), we consider a location and scale invariant modification of $\tau_\alpha(P, \mathcal{N})$. It is convenient to think of $F^{-1}(h^{-1}(y)) =: (F^{-1} \circ h^{-1})(y)$ as a random variable defined on $(0, 1)$ and similarly for other expressions of this type. We note then that if $Q = N(\mu, \sigma^2)$, Φ denotes the standard normal distribution function and $h \in \mathcal{C}_\alpha$, we have $\mathcal{W}_2^2(P_h, Q_h) = E(F^{-1} \circ h^{-1} - \mu - \sigma \Phi^{-1} \circ h^{-1})^2$. Thus, for a fixed h,

$$v(h) := \min_{Q \in \mathcal{N}} \mathcal{W}_2^2(P_h, Q_h) = \mathrm{Var}(F^{-1} \circ h^{-1}) - \frac{\mathrm{Cov}^2(F^{-1} \circ h^{-1}, \Phi^{-1} \circ h^{-1})}{\mathrm{Var}(\Phi^{-1} \circ h^{-1})}, \tag{4}$$

the min being attained at $Q = N(\mu(h), \sigma^2(h))$, where

$$\sigma(h) = \frac{\text{Cov}(F^{-1} \circ h^{-1}, \Phi^{-1} \circ h^{-1})}{\text{Var}(\Phi^{-1} \circ h^{-1})}, \; \mu(h) = E(F^{-1} \circ h^{-1} - \sigma(h)\Phi^{-1} \circ h^{-1}). \quad (5)$$

With this notation we have $\tau_\alpha(P, \mathcal{N}) = \inf_{h \in \mathcal{C}_\alpha} v(h)$. It can be shown that this inf is attained (see Lemma A.1 in the Appendix). To simplify our exposition we make the following technical assumption:

$$v(h) \text{ admits a unique minimizer, } h_0. \quad (6)$$

Now, under (6), we define

$$\tilde{\tau}_\alpha(P, \mathcal{N}) := \frac{\tau_\alpha(P, \mathcal{N})}{r_\alpha(P)},$$

with $r_\alpha(P) = \text{Var}(F^{-1} \circ h_0^{-1})$. Note that $\tilde{\tau}_\alpha(P, \mathcal{N}) = 1 - \text{Corr}^2(F^{-1} \circ h_0^{-1}, \Phi^{-1} \circ h_0^{-1})$. Hence, $\tilde{\tau}_\alpha(P, \mathcal{N})$ is location scale invariant and satisfies $0 \leq \tilde{\tau}_\alpha(P, \mathcal{N}) \leq 1$. We refer to it as the *standardized trimmed distance to normality*. We can see in Figure 1 how $\tau_\alpha(P, \mathcal{N})$ and $\tilde{\tau}_\alpha(P, \mathcal{N})$ change with α for two choices of P: the mixture $0.9N(0, 1) + 0.1 * N(4, 1/4)$ (left) and the mixture $0.5N(0, 1) + 0.5 * N(3, 1/4)$. We can appreciate how after trimming a bit more than the level of 'contamination' we achieve almost perfect fit to normality.

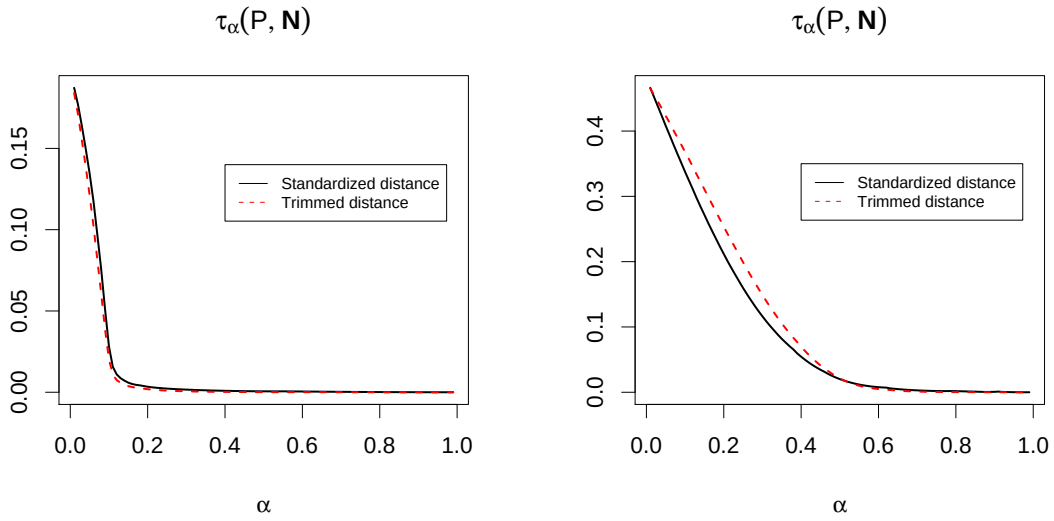

Figure 1: Trimmed distance to normality.

If we admit that our assessment about the normality of the core of the distribution should not depend on the scale of measurement of the data then we should replace the testing problem (3) by

$$H_0 : \tilde{\tau}_\alpha(P, \mathcal{N}) \geq \Delta_0^2 \quad \text{vs.} \quad \tilde{\tau}_\alpha(P, \mathcal{N}) < \Delta_0^2. \quad (7)$$

The threshold is now to be chosen in $(0,1)$ but, again, this arbitrary choice can be avoided with the use of the p-value curve.

2.2 Asymptotic theory.

In order to make $T_{n,\alpha}$ usable in pratice for testing (3) (or (7)) we include this Subsection with a result giving its asymptotic normality as well as providing a consistent estimator of the corresponding asymptotic variance. The computations involve the use of empirical versions of $\mu(h), \sigma(h)$, defined in (5), evaluated at a empirical version of h_0. To be precise we define

$$v_n(h) := \min_{Q \in \mathcal{N}} \mathcal{W}_2^2((P_n)_h, Q_h),\ h \in \mathcal{C}_\alpha. \tag{8}$$

Now, $T_{n,\alpha} = \inf_{h \in \mathcal{C}_\alpha} v_n(h)$ and, as for $v(h)$, we have that the inf is attained (Lemma A.1). We denote $h_n := \operatorname{argmin}_{h \in \mathcal{C}_\alpha} v_n(h)$ and

$$\sigma_n = \frac{\int_0^1 F_n^{-1}\Phi^{-1}h_n' - \int_0^1 F_n^{-1}h_n' \int_0^1 \Phi^{-1}h_n'}{\int_0^1 (\Phi^{-1})^2 h_n' - (\int_0^1 \Phi^{-1})h_n')^2}, \quad \mu_n = \int_0^1 (F_n^{-1} - \sigma_n \Phi^{-1})h_n'. \tag{9}$$

We refer to Subsection 2.3 below for details on the practical computation of $T_{n,\alpha}$, h_n and related estimators. Now we can state the main result in this Section.

Theorem 2.1 *If P satisfies (6), has absolute moments of order $4+\delta$, for some $\delta > 0$, and a distribution function F with continuously differentiable density $F' = f$ such that*

$$\sup_{x \in \mathbb{R}} \left| \frac{F(x)(1-F(x))f'(x)}{f^2(x)} \right| < \infty, \tag{10}$$

then

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P, \mathcal{N})) \underset{w}{\to} N(0, \sigma_\alpha^2(P, \mathcal{N}))$$

where

$$\sigma_\alpha^2(P, \mathcal{N}) = 4\left(\int_0^1 l^2(t)dt - \left(\int_0^1 l(t)dt \right)^2 \right),$$

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

$\mu(h_0)$, $\sigma(h_0)$ are as in (5) and h_0 is the minimizer defined in (6).

If $S_{n,\alpha}^2 := 4\left(\int_0^1 l_n^2(t)dt - \left(\int_0^1 l_n(t)dt \right)^2 \right)$, *where*

$$l_n(t) = \int_{F_n^{-1}(1/2)}^{F_n^{-1}(t)} (x - \mu_n - \sigma_n \Phi^{-1}(F_n(x)))h_n'(F_n(x))dx$$

and μ_n, σ_n are given in (9), then $S^2_{n,\alpha} \to \sigma^2_\alpha(P, \mathcal{N})$ in probability.

The proof of Theorem 2.1 can be found in the Appendix.

2.3 Practical issues, p-value curves, algorithms.

As we noted before, the testing problem (3) involves the choice of a threshold, Δ_0^2. Rather than choosing it in an arbitrary way we consider, as in Munk and Czado (1998) or Alvarez-Esteban et al. (2008a), the *p-value curves*. These curves are built using the asymptotic p-value computed from the test statistic $Z_{n,\alpha} := (T_{n,\alpha} - \Delta_0^2)/S_{n,\alpha}$. Note that from Theorem 2.1 we have $Z_{n,\alpha} \to N(0,1)$ in distribution if $\Delta_0^2 = \tau_\alpha(P, \mathcal{N})$ (hence, $Z_{n,\alpha} \to +\infty$ for $\Delta < \tau_\alpha(P, \mathcal{N})$ and $Z_{n,\alpha} \to -\infty$ for $\Delta > \tau_\alpha(P, \mathcal{N})$).

For each threshold value Δ_0 we compute

$$p(\Delta_0) := \sup_{F \in H_0} \lim_{n\to\infty} P_F\left(Z_{n,\alpha} \le z_0\right) = \Phi\left(\sqrt{n}\tfrac{t_{n,\alpha} - \Delta_0^2}{s_{n,\alpha}}\right),$$

where $z_0 = \sqrt{n}\frac{t_{n,\alpha} - \Delta_0^2}{s_{n,\alpha}}$ is the observed value of $Z_{n,\alpha}$. Then, we plot $p(\Delta_0)$ versus Δ_0. These p-value curves can be used in two ways. On one hand, fixing Δ_0, which controls the degree of dissimilarity, we can find the level of significance at which F cannot be considered essentially normal (at trimming level α). On the other hand, for a fixed test level (p-value), we can find the value of Δ_0 such that for every $\Delta \ge \Delta_0$ we should reject the hypothesis $H_0 : \tau_\alpha(P, \mathcal{N}) \ge \Delta^2$.

In practice we will be interested in testing (7) rather than (3). We can rewrite (7) as

$$H_0 : \tau_\alpha(P, \mathcal{N}) \ge \Delta_0^2 r_\alpha(P) \quad \text{vs.} \quad \tau_\alpha(P, \mathcal{N}) < \Delta_0^2 r_\alpha(P),$$

a family of testing problems that could be analysed using the p-value curve $p(\Delta_0 r_\alpha^{1/2}(P))$. Since $r_\alpha(P)$ is unknown we replace it by the consistent estimator $R_{n,\alpha} = \int_0^1 (F_n^{-1})^2 h_n' - \left(\int_0^1 F_n^{-1} h_n'\right)^2$ and obtain the estimated p-value curve for (7):

$$\tilde{p}(\Delta_0) := p(\Delta_0 R_{n,\alpha}^{1/2}), \quad 0 < \Delta_0 < 1.$$

The values of Δ_0 should be interpreted taking into account that $\tilde{\tau}_\alpha(P, \mathcal{N})$ takes values in $[0, 1]$. $\tilde{\tau}_\alpha(P, \mathcal{N}) = 0$ means perfect fit to normality after trimming, while large values of $\tilde{\tau}_\alpha(P, \mathcal{N})$ (close to 1) mean severe nonnormality even after trimming.

We turn now to computational details. First, to compute the value of $T_{n,\alpha}$ we observe that

$$T_{n,\alpha} = \min_{h \in \mathcal{C}_\alpha, \mu \in \mathbb{R}, \sigma \ge 0} \int_0^1 (F_n^{-1} - \mu - \sigma\Phi^{-1})^2 h' = \min_{\mu \in \mathbb{R}, \sigma \ge 0} V_n(\mu, \sigma),$$

where

$$V_n(\mu,\sigma) = \min_{h\in\mathcal{C}_\alpha} \int_0^1 (F_n^{-1} - \mu - \sigma\Phi^{-1})^2 h'.$$

If $\sigma > 0$ then $V_n(\mu,\sigma) = \int_0^1 (F_n^{-1} - \mu - \sigma\Phi^{-1})^2 h'_{n,\mu,\sigma}$, where

$$h'_{n,\mu,\sigma} = \frac{1}{1-\alpha} I_{|F_n^{-1}-\mu-\sigma\Phi^{-1}|\le k_{n,\mu,\sigma}}$$

and $k_{n,\mu,\sigma}$ is the (unique) k such that the set $\{t \in (0,1) : |F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t)| \le k\}$ has Lebesgue measure $1-\alpha$. We use this to compute numerically $V_n(\mu,\sigma)$ as follows

(1) Compute the values of $|F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t)|$ in a (fine) grid of $[0,1]$.
(2) Approximate $k_{n,\mu,\sigma}$ as the $(1-\alpha)$-quantile of these values.
(3) Approximate $V_n(\mu,\sigma)$ as the average of $(F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t))^2 h'_{n,\mu,\sigma}(t)$ over the grid.

Now, minimization of $V_n(\mu,\sigma)$ yields $T_{n,\alpha}$. We carry out this step through a simple search-in-a-grid of (μ,σ), although this could be replaced by an optimization procedure based in gradient methods avalaible in R (see e.g. *nlm*, *optim*), using the sample values as initial values.

If μ_n and σ_n are the minimizers of $V_n(\mu,\sigma)$ obtained with the above algorithm, then take

$$h_n = h_{n,\mu_n,\sigma_n}.$$

Finally, aproximate

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

computing numerically the integrals, where

$$l_n(t) := \int_{F_n^{-1}(1/n)}^{F_n^{-1}(t)} \left(x - \mu_n - \sigma_n\Phi^{-1}(F_n(x))\right) h'_n(F_n(x))dx$$

is evaluated numerically by averaging in the grid in $(0,1)$ as above. Similarly we compute $R_{n,\alpha}$.

All these procedures have been implemented in an R program avalaible at `http:// www.eio.uva.es/ ~pedroc/R`. These computations have been coded in a vectorized way and the result is that for moderate sizes of n (100-500) the time required in a PC is just a few seconds, while for big sizes (5000) is a couple of minutes, depending on the grid for (μ,σ). The grid in $[0,1]$ for the computation of $V_n(\mu,\sigma)$ was obtained splitting $[0,1]$ into 10^5 intervals of equal length.

We end this Section with a remark on the computational convenience of the normalization chosen here for the standardized trimmed distance to normality,

$\tilde{\tau}_\alpha(P, \mathcal{N})$. Other alternatives could be chosen, the most natural being perhaps $\tau^*_\alpha(P, \mathcal{N}) = \min_{h \in \mathcal{C}_\alpha} w(h)$ with

$$w(h) = \frac{v(h)}{\mathrm{Var}(F^{-1} \circ h^{-1})} = 1 - \frac{\mathrm{Cov}^2(F^{-1} \circ h^{-1}, \Phi^{-1} \circ h^{-1})}{\mathrm{Var}(F^{-1} \circ h^{-1})\mathrm{Var}(\Phi^{-1} \circ h^{-1})},$$

(recall the notation in (4)). Now $w(h)$ is location-scale free for every $h \in \mathcal{C}_\alpha$ and so is $\tau^*_\alpha(P, \mathcal{N})$. Using the method of proof of Theorem 2.1 we could prove also asymptotic normality of an empirical version of $\tau^*_\alpha(P, \mathcal{N})$. Its use in practice, however, is rather troublesome. If we try to mimick the algorithm that we used to evaluate $T_{n,\alpha}$ we should be able to compute

$$W_n(\mu, \sigma) = \min_{h \in \mathcal{C}_\alpha} \frac{\mathcal{W}_2^2((P_n)_h, N(\mu, \sigma^2)_h)}{\mathrm{Var}(P_h)}$$

for fixed μ and σ. Unfortunalety there is no easy expression for the optimal h in this minimisation problem. We could rewrite it as an optimal control problem and use appropriate numerical methods but, yet, the computational burden required for a single evaluation of $W_n(\mu, \sigma)$ discourages its use.

3 Examples and Simulations

3.1 Example 1, real data.

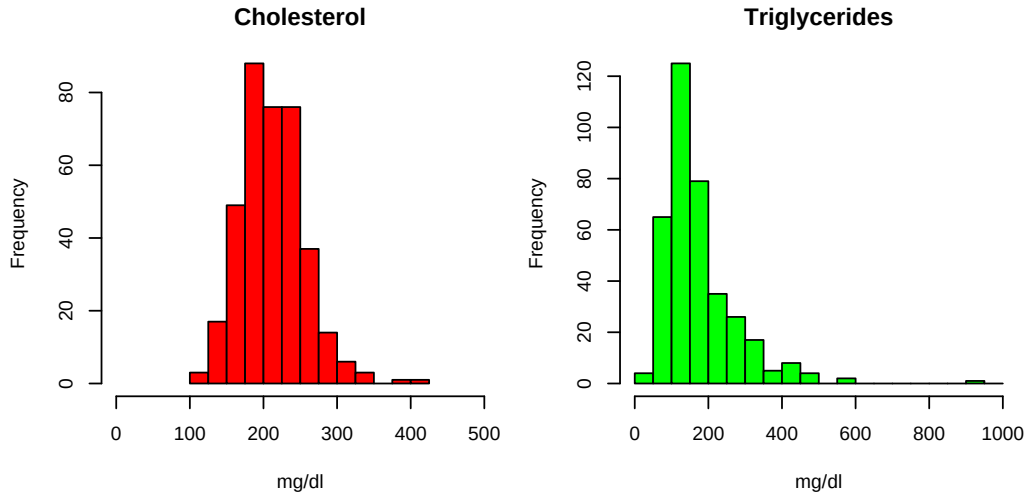

Figure 2: Histogram for variables Cholesterol and Triglycerides.

We use the variables concentration of plasma cholesterol and plasma triglycerides (mg/dl) collected from $n = 371$ patients (see Hand et al., 1994) to illustrate the application of our procedure to investigate whether these sam-

ples can be considered normal at some reasonable trimming level. Figure 2 shows the histograms for both samples. The cholesterol sample shows a slight positive skewness with two possible outliers. The triglycerides sample exhibits a clear positive skewness and a clear outlier at the right tail. Using classical procedures such as the Shapiro-Wilks test would reject the hypothesis of normality, even if we remove the mentioned outliers ($p = .0306$ and $p < .0000$, respectively).

	Cholesterol			Triglycerides		
α	μ_n	σ_n	$\tilde{\tau}_\alpha(P_n, \mathcal{N})$	μ_n	σ_n	$\tilde{\tau}_\alpha(P_n, \mathcal{N})$
0	213.31	42.07	0.026	173.94	88.90	0.193
0.05	212.08	39.64	0.005	165.92	71.27	0.112
0.10	211.75	39.88	0.004	161.16	63.87	0.089
0.20	211.02	41.68	0.002	153.15	53.04	0.048

Table 1: Distances, means and standard deviations of the best α-trimmed normal distribution.

We use instead our data-driven trimming method to obtain the best α-trimmed Gaussian approximation to each sample. Table 1 shows the means and standard deviations of the best normal approximation as well as the standardized trimmed distances to normality, $\tilde{\tau}_\alpha(P_n, \mathcal{N})$, for both samples and different trimming sizes.

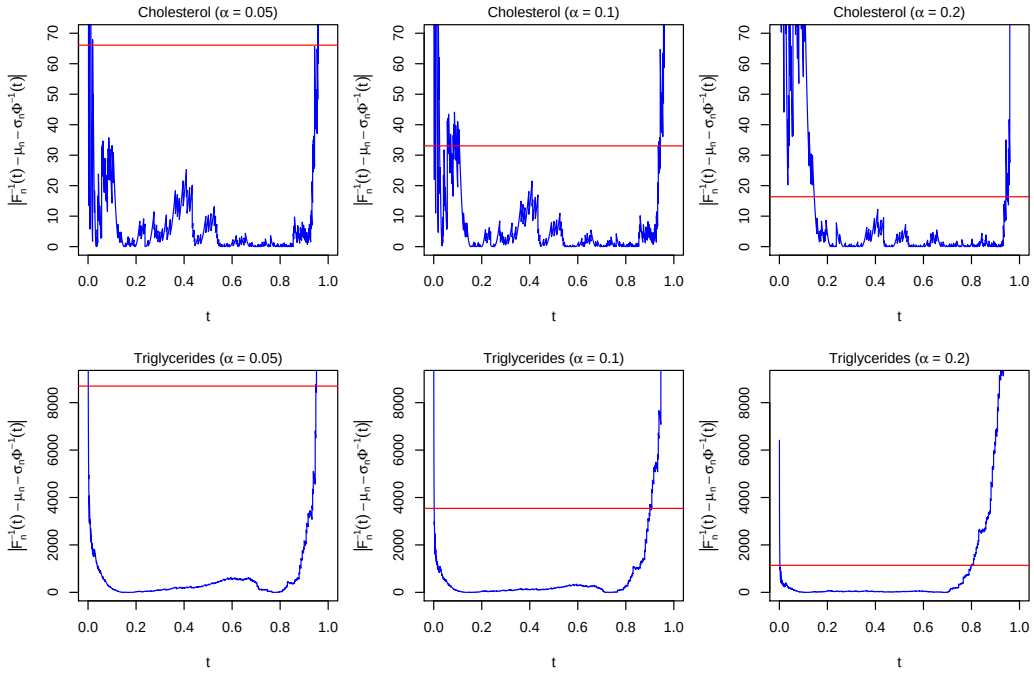

Figure 3: Optimal trimming functions for both Cholesterol and Triglycerides

samples, for different trimming sizes ($\alpha = 0.05$, 0.1 and 0.2).

Figure 3 shows the optimal trimming functions for Cholesterol and Triglycerides samples for different values of α (0.05, 0.1 and 0.2). In each graph we plot the value of $J_n(t) := |F_n^{-1}(t) - \mu_n - \sigma_n \Phi^{-1}(t)|$ and the cutting values k_{n,μ_n,σ_n}, where μ_n and σ_n are the mean and the standard deviation of the closest normal distribution (see Table 1) estimated using the algorithm described in Subsection 2.3. These plots show that trimming should be made mostly at the tails of the distributions to make them as normal as possible. This optimal trimming, though, is not symmetric, and not always removing all the observations at the tails. The plot corresponding to the Cholesterol sample and $\alpha = 0.2$ suggests that if we increase the trimming size then some observations in the center of the distribution would be trimmed.

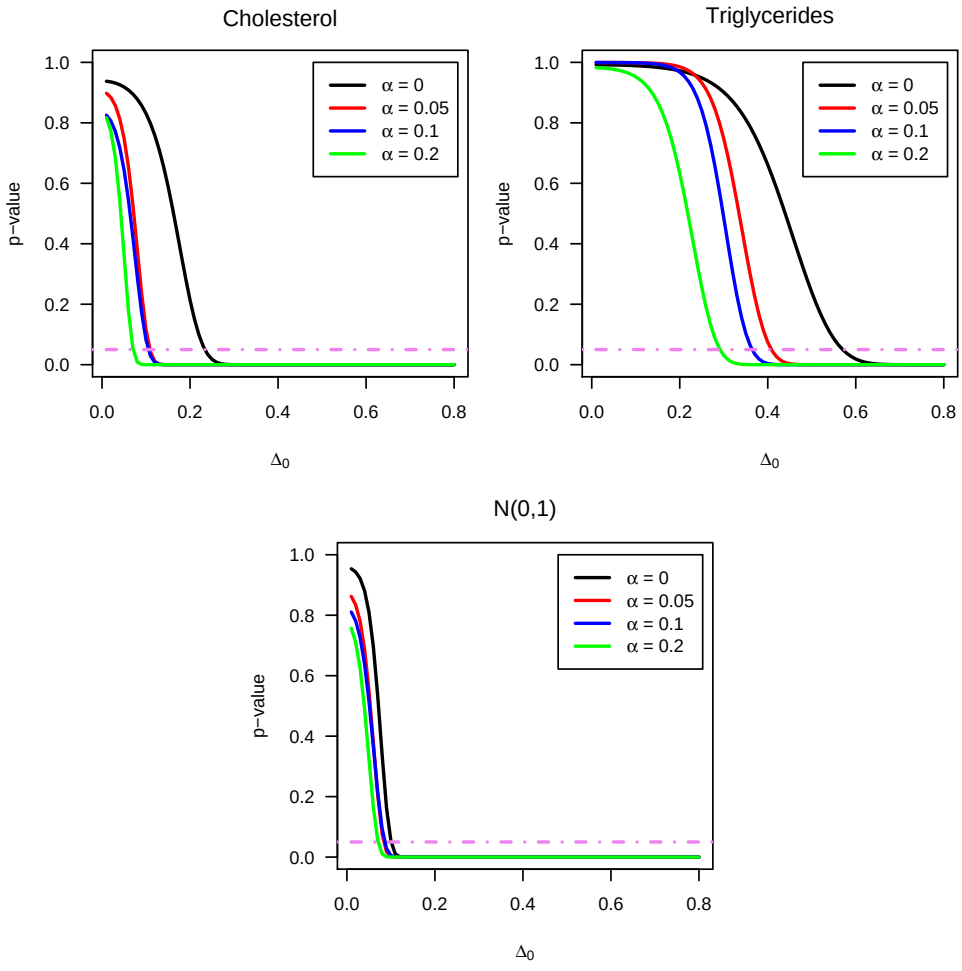

Figure 4: P-value curves for Cholesterol, Tryglicerides and the reference case. The dotted line is a reference line ($p = 0.05$).

In order to assess the degree of normality of the samples we use the p-value curves, $\tilde{p}(\Delta)$, introduced in Subsection 2.3. Figure 4 shows these curves for

Cholesterol and Triglycerides samples for different trimming sizes and the "No trimming" case. Note that although the asymptotic distribution has not been explicitly given for this case in this paper, it can be easily derived following the same arguments as for Theorem 2.1 and coincides with the limit case $\alpha = 0$ in this theorem. The third graph corresponds to a random sample of the same size ($n = 371$) drawn from the standard normal distribution. This graph has been included as a reference to simplify the assessment of the normality of the previous samples.

Although in both cases there is a significant improvement after trimming the initial 5% (fixing $p = 0.05$, from $\Delta_0 = 0.24$ when $\alpha = 0$ to $\Delta_0 = 0.11$ when $\alpha = 0.05$ for Cholesterol sample; and from $\Delta_0 = 0.57$ to $\Delta_0 = 0.41$ for Triglycerides sample), both samples exhibit a different behaviour. While in the first sample the trimmed distance to normality reaches similar values to those of a normal distribution (see the third graph), it does not in the second sample. In this last sample the trimmed distance to normality does not reach the same levels even if the trimming size is $\alpha = 0.2$ or $\alpha = 0.3$ (not shown in Figure 4). Thus, the Cholesterol sample can be considered normal after little trimming ($\alpha = 0.05$, then, mostly normal), however, the Triglycerides sample can not be considered normal at reasonable levels of trimming.

3.2 Example 2, simulated data

To better illustrate the use of the p-value curves to assess essential normality we have generated 100 random observations from six different models (two different normal models, two normal models with a small contamination that after trimming, are very close to the normality, a chi-square model and an exponential model). Figure 5 shows the associated p-value curves. Graph (a) corresponds to the N(0,1) model. The behaviour is clear, the standardized distance before trimming is very close to 0, and there is a small decrease after the initial 5% of trimming -probably due to some smoothing effect in the randomness-. Finally, a stabilization of the standardized trimmed distance is observed when α is increased. In other words, there is no improvement in the degree of normality increasing the size of trimming. Graph (b) shows the p-value curves for the N(10,4) model, quite similar to those of the previous model, illustrating that the standardization introduced in (2) performs as expected. Otherwise we would have noticed a scale effect that would affect the scale in the horizontal axis.

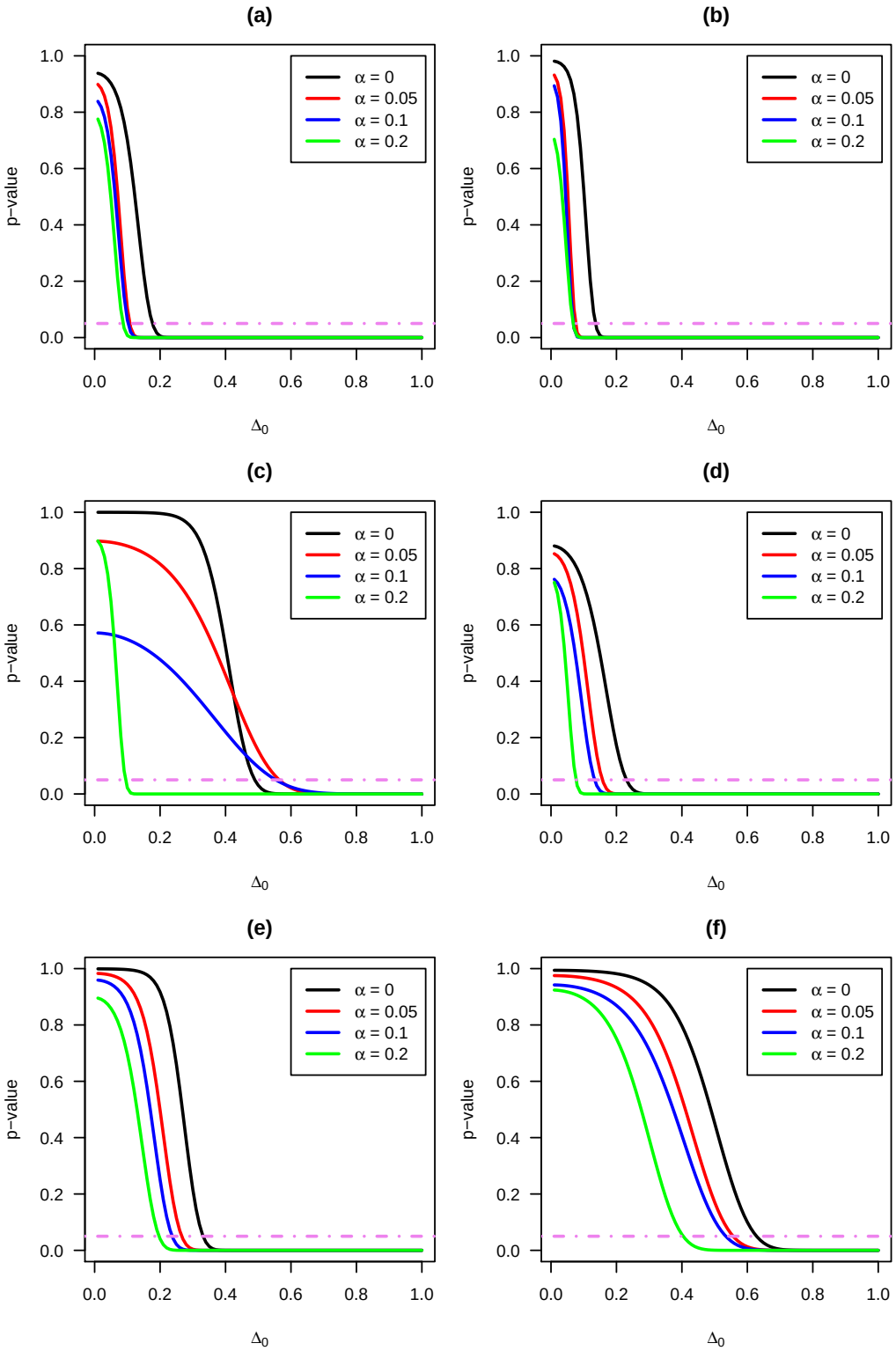

Figure 5: $p-$value curves for simulated data: (a) N(0,1); (b) N(10,4); (c) 0.9*N(0,1) + 0.1*N(-5,1); (d) 0.9*N(0,1) + 0.1*N(-3,1); (e) χ^2_4; and (f) exp(1). The dotted line is a reference line ($p = 0.05$).

Graphs (c) and (d) correspond to the mixtures $0.9 * N(0,1) + 0.1 * N(-5,1)$ and $0.9 * N(0,1) + 0.1 * N(-3,1)$, respectively. The behaviour is different. In the first case the contamination is clearly detected, fixing $p = 0.05$ the significant standardized distance is approximately $\Delta_0 = 0.5$ before trimming, and this value decreases to $\Delta_0 = 0.07$ when $\alpha = 0.2$, similar to the values observed in the normal models. Then, this sample can be considered normal at level $\alpha = 0.2$. The crossings observed in the p-value curves when $\alpha = 0.05$ or 0.1 are related to the variability in the estimation of the asymptotic variance $\sigma^2(P, \mathcal{N})$. In the second case, graph (d), the contamination is timidly detected as the significant standardized distance when $p = 0.05$ is slightly greater than those in the normal models ($\Delta_0 = 0.24$ in (d), whereas $\Delta_0 = 0.17$ in (a) or $\Delta_0 = 0.14$ in (b)). This distance decreases to values similar to those of the normal models when $\alpha = 0.05$ or 0.1.

The remaining graphs in Figure 5, (e) and (f), correspond to cases where normality is not reached at reasonable levels of trimming, a χ^2_4 model and an exponential model, respectively. In both cases the significant standardized distance after trimming is clearly far from that of the normal model ($\Delta_0 = 0.34$ and $\Delta_0 = 0.64$ vs $\Delta_0 = 0.17$ in (a)). There is also a clear improvement in this distance when the trimming size increases. However, in both cases this distance does not reach similar values to those of the normal model, even if $\alpha = 0.2$. In the chi-square case the difference with respect to the normal model is lower than in the exponential case where this standardized distance is quite far from that of the normal model ($\Delta_0 = 0.41$ in (f) vs $\Delta_0 = 0.09$ in (a)). Thus, these samples can not be considered normal at any reasonable level of trimming.

3.3 Simulation study

We finish this section with a short simulation study of the power of the proposed test to assess mostly normality for finite samples. We consider two different population models: $P_1 = 0.9 * N(0,1) + 0.1 * N(-3,1)$ and $P_2 = \chi^2_2$, the first one mostly normal and the second one farther away from normality.

We want to test the null hypothesis $H_0^i : \tilde{\tau}_\alpha(P_i, \mathcal{N}) \geq \Delta_0^2$ vs $H_a^i : \tilde{\tau}_\alpha(P_i, \mathcal{N}) < \Delta_0^2$ for different values of Δ_0 and two trimming sizes ($\alpha = 0.05$ and $\alpha = 0.1$). To do that, for each situation we obtain 10000 replicas of the statistic $\tilde{p}(\Delta_0)$ for several values of n, rejecting H_0 when $\tilde{p}(\Delta_0) < 0.05$. Tables 2 and 3 contain the observed rejection frequencies for P_1 and P_2 respectively. $\tilde{\tau}_\alpha$ represents the theoretical standardized trimmed distance and has been estimated in all cases from ten samples of 100000 observations. In both cases the simulation study shows that even for moderate sample sizes the performance of the test is quite good. We observe that the rejection frequency is low when the threshold value is smaller than the true distance and high otherwise. When the threshold

value is near the (estimated) distance then the simulated power is close to the nominal power.

	$\alpha = 0.05$, $\tilde{\tau}_\alpha(P_1, \mathcal{N}) \simeq 0.0225$					$\alpha = 0.1$, $\tilde{\tau}_\alpha(P_1, \mathcal{N}) \simeq 0.0079$				
$\Delta_0^2 =$	0.001	0.01	0.0225	0.05	0.1	0.001	0.005	0.0079	0.025	0.05
n	Frequency					Frequency				
100	0	0.0012	0.0478	0.3410	0.8298	0	0.0003	0.0058	0.2654	0.7048
200	0	0.0012	0.0484	0.5038	0.9732	0	0.0010	0.0206	0.4404	0.9182
500	0	0	0.0355	0.8114	1	0	0.0016	0.0266	0.8068	0.9988
1000	0	0	0.0392	0.9780	1	0	0.0010	0.0304	0.9748	1
5000	0	0	0.0445	1	1	0	0	0.0326	1	1
10000	0	0	0.0520	1	1	0	0	0.0460	1	1

Table 2: Observed rejection frequencies for $P_1 = 0.9 * N(0,1) + 0.1 * N(-3,1)$.

	$\alpha = 0.05$, $\tilde{\tau}_\alpha(P_2, \mathcal{N}) \simeq 0.1272$					$\alpha = 0.1$, $\tilde{\tau}_\alpha(P_2, \mathcal{N}) \simeq 0.1022$				
$\Delta_0^2 =$	0.05	0.1	0.1272	0.15	0.25	0.01	0.05	0.1022	0.15	0.25
n	Frequency					Frequency				
100	0.0004	0.0274	0.0724	0.1296	0.5232	0	0.0014	0.0770	0.2724	0.7260
200	0	0.0122	0.0534	0.1594	0.8198	0	0.0008	0.0898	0.4378	0.9352
500	0	0.0028	0.0590	0.2528	0.9954	0	0	0.1042	0.6770	0.9996
1000	0	0	0.0570	0.3718	1	0	0	0.0886	0.8828	1
5000	0	0	0.0418	0.9196	1	0	0	0.0578	1	1
10000	0	0	0.0378	0.9953	1	0	0	0.0486	1	1

Table 3: Observed rejection frequencies for $P_2 = \chi^2_2$.

A Appendix

To prove Theorem 2.1 the following technical lemma is needed,

Lemma A.1 *Let $v(h)$, $\mu(h)$ and $\sigma(h)$ be defined as in (4)-(5). Then, assuming F has finite second moment*

(a) $v(h)$, $\mu(h)$ and $\sigma(h)$ are bounded and continuous in $h \in \mathcal{C}_\alpha$ with respect to the uniform norm.
(b) $v(h)$ attains its minimum in $\mathcal{C}_\alpha$.

Proof. We show that given two square integrable quantile functions F^{-1}, G^{-1} the functional $a(h) = \int_0^1 F^{-1}(h^{-1}(t))G^{-1}(h^{-1}(t))dt$ is continuous in $\mathcal{C}_\alpha$ for the uniform norm. To check this, take $\{h_n\}_n, h_0 \in \mathcal{C}_\alpha$ such that $\|h_n - h_0\| \to 0$. This implies $h_n(F(x)) \to h_0(F(x))$ for every x. The associated quantiles satisfy $F^{-1}(h_n^{-1}(t)) \to F^{-1}(h_0^{-1}(t))$ at almost every $t \in (0,1)$ and similarly for G. To

conclude that $a(h_n) \to a(h_0)$ it suffices to show that $F^{-1} \circ h^{-1} G^{-1} \circ h^{-1}$ is uniformly integrable. But this follows from the fact that

$$\begin{aligned}
&\sup_{h \in \mathcal{C}_\alpha} \int_0^1 |F^{-1}(h^{-1}(t)) G^{-1}(h^{-1}(t))| I(|F^{-1}(h^{-1}(t)) G^{-1}(h^{-1}(t))| > K) dt \\
&\qquad = \sup_{h \in \mathcal{C}_\alpha} \int_0^1 |F^{-1}(y) G^{-1}(y)| I(|F^{-1}(y) G^{-1}(y)| > K) h'(y) dy \\
&\qquad \leq \frac{1}{1-\alpha} \int_0^1 |F^{-1}(y) G^{-1}(y)| I(|F^{-1}(y) G^{-1}(y)| > K) dy \to 0
\end{aligned}$$

as $K \to \infty$. This proves continuity of $\mathrm{Var}(F^{-1} \circ h^{-1})$, $\mathrm{Var}(\Phi^{-1} \circ h^{-1})$ and $\mathrm{Cov}(F^{-1} \circ h^{-1}, \Phi^{-1} \circ h^{-1})$. Since $\mathcal{C}_\alpha$ is compact for the uniform topology (see Alvarez-Esteban et al. (2008a)) we have that $\mathrm{Var}(\Phi^{-1} \circ h^{-1})$ attains its minimum value: $\min_{h \in \mathcal{C}_\alpha} \mathrm{Var}(\Phi^{-1} \circ h^{-1}) = \mathrm{Var}(\Phi^{-1} \circ h_{opt}^{-1})$. But this shows that $\mathrm{Var}(\Phi^{-1} \circ h^{-1})$ is bounded away from 0 in $\mathcal{C}_\alpha$ (a distribution with a density cannot have zero variance). This implies that $v(h), \mu(h)$ and $\sigma(h)$ are continuous. All the remaining claims follow from compactness of $\mathcal{C}_\alpha$. ■

To complete the proof of Theorem 2.1 we note that, similarly as in (4), we can define

$$v_n(h) := \min_{Q \in \mathcal{N}} \mathcal{W}_2^2((P_n)_h, Q_h) = \mathrm{Var}(F_n^{-1} \circ h^{-1}) - \frac{\mathrm{Cov}^2(F_n^{-1} \circ h^{-1}, \Phi^{-1} \circ h^{-1})}{\mathrm{Var}(\Phi^{-1} \circ h^{-1})}$$

and then we have

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P, \mathcal{N})) = \sqrt{n} \left(\min_{h \in \mathcal{C}\alpha} v_n(h) - \min_{h \in \mathcal{C}\alpha} v(h) \right). \tag{A.1}$$

We will obtain the conclusion in Theorem 2.1 from the study of the process $M_n(h) = \sqrt{n}(v_n(h) - v(h))$, $h \in \mathcal{C}_\alpha$. We note that, writing $\rho_n(t) = \sqrt{n}(F_n^{-1}(t) - F^{-1}(t)) f(F^{-1}(t))$ (the quantile process) we can see after some routine computations that

$$\begin{aligned}
M_n(h) &= 2 \int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))} (F^{-1}(t) - \mu(h) - \sigma(h) \Phi^{-1}(t)) h'(t) dt \qquad \text{(A.2)} \\
&+ \frac{1}{\sqrt{n}} \left[\int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} h'(t) dt - \left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))} h'(t) dt \right)^2 \right. \\
&\left. + \frac{1}{\sigma_\Phi^2(h)} \left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))} \Phi^{-1}(t) h'(t) dt - \left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))} h'(t) dt \right) \mu_\Phi(h) \right)^2 \right],
\end{aligned}$$

where $\mu_\Phi(h) = \int_0^1 \Phi^{-1} \circ h^{-1}$ and $\sigma^2_\Phi(h) = \int_0^1 (\Phi^{-1} \circ h^{-1})^2 - \mu^2_\Phi(h)$. We consider a Brownian bridge, $\{B(t)\}_{t\in(0,1)}$, and define

$$M(h) := 2\int_0^1 \frac{B(t)}{f(F^{-1}(t))}(F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))h'(t)dt.$$

Observe that $\{M(h)\}_{h\in\mathcal{C}_\alpha}$ is a centered Gaussian process with covariance function

$$K(h_1, h_2) = 4\int_0^1 l_1(t)l_2(t)dt - 4\int_0^1 l_1(t)dt \int_0^1 l_2(t)dt,$$

where

$$l_i(t) = \int_{F^{-1}(1/2)}^{F^{-1}(t)} (x - \mu(h_i) - \sigma(h_i)\Phi^{-1}(F(x)))h_i'(F(x))dx, \quad i = 1, 2.$$

This follows from noting that, integrating by parts, $M(h_i) = -2\int_0^1 l_i(t)dB(t)$. The key result in this Appendix is the following.

Proposition A.2 *Under the assumptions of Theorem 2.1 M is a tight Borel measurable map and M_n converges weakly to M in $\ell^\infty(\mathcal{C}_\alpha)$.*

Proof. We assume w.l.o.g. that there exist Brownian bridges B_n satisfying

$$n^{1/2-\nu} \sup_{\frac{1}{n}\le t\le 1-\frac{1}{n}} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} = \begin{cases} O_P(\log n), & \text{if } \nu = 0 \\ O_P(1), & \text{if } 0 < \nu \le 1/2 \end{cases} \tag{A.3}$$

(this is guaranteed by (10), see Theorem 6.2.1 in Csörgö and Horváth (1993)). Now we define.

$$N_n(h) := 2\int_0^1 \frac{B_n(t)}{f(F^{-1}(t))}(F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))h'(t)dt$$

We claim that $\|M_n - N_n\|_{\mathcal{C}_\alpha} := \sup_{h\in\mathcal{C}_\alpha} |M_n(h) - N_n(h)| \to 0$ in probability. To check this we write

$$\sup_{h\in\mathcal{C}_\alpha} \frac{1}{\sqrt{n}} \int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} h'(t)dt \le \frac{1}{1-\alpha}\frac{1}{\sqrt{n}} \int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} dt \to_{\text{Pr.}} 0, \tag{A.4}$$

where the last convergence follows from the moment assumption on F (see Lemma A.1 and the proof of Theorem 2 in Alvarez-Esteban et al. (2008a)). Thus we can (uniformly) neglect the second term in expression (A.2) for M_n. A similar bound and the fact that $\min_{h\in\mathcal{C}_\alpha} \sigma^2_\Phi(h) > 0$ is enough to control the third term. Hence, it suffices to show that

$$\sup_{h\in\mathcal{C}_\alpha} \int_0^1 \frac{|\rho_n(t) - B_n(t)|}{f(F^{-1}(t))} |F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))|dt \to_{\text{Pr.}} 0.$$

This can be done using boundedness of $\mu(h)$, $\sigma(h)$ and arguing as in the proof of Theorem 2 in Alvarez-Esteban et al. (2008a).

Since N_n and M are equally distributed, to complete the proof we have to show that M is tight or, equivalently, that it is uniformly equicontinuous in probability for some metric d for which $\mathcal{C}_\alpha$ is totally bounded (see Theorems 1.5.7 and 1.10.2 in van der Vaart and Wellner (1996)). We take d to be the uniform norm in $\mathcal{C}_\alpha$ (then $\mathcal{C}_\alpha$ is compact) and note that we have to prove that for any given $\varepsilon, \eta > 0$ there exists $\delta > 0$ such that

$$P\left(\sup_{\|h_1 - h_2\|_\infty < \delta} |M(h_1) - M(h_2)| > \varepsilon\right) < \eta.$$

From Markov's inequality and compactness we see that it is enough to show that the map $h \mapsto E|M(h)|$ is $\|\cdot\|_\infty$-continuous and this can done arguing as in the proof of Lemma A.1 ■

Proposition A.2 has the following simple, but important consequences

Corollary A.3 *Under the assumptions of Theorem 2.1*

$$\sup_{h \in \mathcal{C}_\alpha} |v_n(h) - v(h)| \to_{Pr.} 0.$$

As a consequence $\|h_n - h_0\| \to_{Pr.} 0$.

Proof. Proposition A.2 implies that $\sqrt{n}\sup_h |v_n(h) - v(h)| = \sup_h |M_n(h)|$ converges weakly to $\sup_{h \in \mathcal{C}_\alpha} |M(h)|$. This proves the first claim. To complete the proof, observe that compactness allows to extract convergent subsequences from h_n: $h_{n'} \to h_a$ and also to ensure that $v_{n'}(h) \to v(h)$ uniformly. Since $v_{n'}(h_{n'}) \leq v'_n(h)$ we see, taking limits, that h_a must be a minimizer of v. Hence, by the uniqueness asumption (6) we have $h_a = h_0$. This completes the proof. ■

Proof of Theorem 2.1. From (A.1) we see that $\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P, \mathcal{N})) = \sqrt{n}(v_n(h_n) - v(h_0)) = M_n(h_0) - \sqrt{n}(v_n(h_0) - v_n(h_n))$. Optimality implies $v(h_n) - v(h_0) \geq 0$ and $v_n(h_0) - v_n(h_n) \geq 0$. On the other hand

$$\sqrt{n}(v(h_n) - v(h_0)) + \sqrt{n}(v_n(h_0) - v_n(h_n)) = M_n(h_0) - M_n(h_n) \to_{\mathrm{Pr.}} 0,$$

the last convergence implied by Proposition A.2, Corollary A.3 and equicontinuity. From this we get $\sqrt{n}(v_n(h_0) - v_n(h_n)) \to_{\mathrm{Pr.}} 0$, hence $\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P, \mathcal{N}))$ converges in distribution to $M(h_0)$, proving the first part of Theorem 2.1. The second claim can be proved with the aid of Corollary A.3 and continuity arguments as in Lemma A.1. We skip details. ■

References

Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2008a). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, **103**, 697-704.

Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2008b). Similarity of probability measures through trimming. Submitted.

del Barrio, E.; Cuesta-Albertos, J.A.; Matrán, C. and Rodríguez Rodríguez, J. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.* **27**: 1230-1239.

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

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.

Bickel, P. J., and Freedman, D. A. (1981). Some Asymptotic Theory for the Bootstrap, *Ann. Statist.*, **9**, 1196–1217.

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

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

García Escudero, L.A.; Gordaliza, A. and Matrán, C. (2003). Trimming tools in exploratory data analysis. *J. Comput. and Graph. Stat.* **12**, 434–449.

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.

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

Hand, D.J.; Daly, F.; Lunn, A.D.; McConway. K.J. and Ostrowski, E. *A Handbook of Small Data Sets*. Chapman & Hall. London.

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

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 (2008). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org*. Viena, Austria.

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

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer. New York.